frontiers | Research Topics

Evidence-informed reasoning of pre- and in-service teachers

Edited by
Ingo Kollar, Martin Greisel and Robin Stark

Published in
Frontiers in Education

ISSN 1664-8714
ISBN 978-2-8325-2377-3
DOI 10.3389/978-2-8325-2377-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Evidence-informed reasoning of pre- and in-service teachers

Topic editors
Ingo Kollar — University of Augsburg, Germany
Martin Greisel — University of Augsburg, Germany
Robin Stark — Saarland University, Germany

Topic Coordinator
Theresa Krause-Wichmann — Saarland University, Germany

Citation
Kollar, I., Greisel, M., Stark, R., eds. (2023). *Evidence-informed reasoning of pre- and in-service teachers*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2377-3

Table of contents

TYPE Editorial
PUBLISHED 25 April 2023
DOI 10.3389/feduc.2023.1188022

OPEN ACCESS

EDITED AND REVIEWED BY
Ramona Maile Cutri,
Brigham Young University, United States

*CORRESPONDENCE
Ingo Kollar
✉ ingo.kollar@uni-a.de

RECEIVED 16 March 2023
ACCEPTED 12 April 2023
PUBLISHED 25 April 2023

CITATION
Kollar I, Greisel M, Krause-Wichmann T and Stark R (2023) Editorial: Evidence-informed reasoning of pre- and in-service teachers. *Front. Educ.* 8:1188022.
doi: 10.3389/feduc.2023.1188022

Editorial: Evidence-informed reasoning of pre- and in-service teachers

Ingo Kollar[1]*, Martin Greisel[1], Theresa Krause-Wichmann[2] and Robin Stark[2]

[1]Chair for Educational Psychology, University of Augsburg, Augsburg, Germany, [2]Chair for Personality Development and Education, Saarland University, Saarbrücken, Germany

KEYWORDS
evidence-informed reasoning, pre-service teachers, in-service teachers, teacher professionalism, educational theories, scaffolding

Editorial on the Research Topic
Evidence-informed reasoning of pre- and in-service teachers

Evidence-informed reasoning as an important requirement for pre- and in-service teachers

How can I help my students acquire the skill of dividing fractions? How can I increase my students' learning motivation? What is the reason for Fiona's learning difficulties? These are just a couple of problems that teachers face in their classes on a daily basis. To competently cope with such problems, teachers should be able to retrieve, use, and apply evidence from Educational Science, Educational Psychology, and subject-matter didactics, and, in that way, engage in "evidence-informed reasoning" (e.g., Greisel et al., 2023).

This Research Topic assembles scientific contributions that refer to four questions: (1) What does evidence-based education mean and why is it important? (2) What are barriers for pre- and in-service teachers' evidence-informed reasoning? (3) How can pre- and in-service teachers' evidence-informed reasoning be scaffolded? (4) How does in-service teachers' evidence-informed reasoning impact student performance?

The contributions within this Research Topic

What does evidence-based education mean and why is it important?

In their contribution, Dekker and Meeter discuss and evaluate the arguments with which evidence-based education and especially the view that randomized controlled trials should be regarded most important to inform educational practice are criticized. Taking the critique into account, they propose not to dismiss evidence-based education in general and randomized controlled trials in particular, but show how they should be administered and complemented to be more informative to research and practice.

What are barriers for pre- and in-service teachers' evidence-informed reasoning?

Three contributions of this Research Topic show that *unfavorable beliefs and low trust toward educational science* may act as barriers for pre- and in-service teachers' evidence informed-reasoning:

Voss presents three empirical studies that indicate that pre-service teachers tend to hold skeptical beliefs about the importance of educational science for the solution of educational problems, especially when they have little experience with educational science as a domain, and when they have a background in natural sciences. In turn, such unfavorable beliefs seem to go hand in hand with pre-service teachers' low engagement in educational science courses.

Schmidt et al. investigate to what extent in-service teachers trust knowledge claims from educational research. They find that teachers' trust in claims from educational research is higher than their trust in claims made on the basis of anecdotal evidence, and that their trust in educational science is positively related to general trust in science. Yet, the authors also show that teachers trust knowledge claims from educational research more when they confirm their prior beliefs.

Similarly, Futterleib et al. show that pre-service teachers tend to devalue findings from educational science when they do not confirm their prior beliefs. However, this only seems to apply when the evidence is strong and unambiguous. If the evidence leaves more room for interpretation, pre-service teachers might find other ways to protect their beliefs instead of devaluing science. Furthermore, pre-service teachers assess educational science as pertinent to investigate educational topics independent of whether they are confronted with belief-challenging evidence.

Two articles of this Research Topic show that pre- and in-service teachers' *suboptimal skills regarding the retrieval and argumentative use of educational evidence* may act as another barrier for pre- and in-service teachers' evidence-informed reasoning:

Zimmermann et al. investigate how pre-service teachers search for information on educational topics on the internet. They find that pre-service teachers' search strategies are often suboptimal, especially when it comes to evaluating the trustworthiness of websites and the quality of their content. Yet, even though the authors hypothesized that the employment of advanced search strategies should depend on pre-service teachers' internet-specific epistemological beliefs, the results do not support this assumption.

Bauer et al. demonstrate that pre-service teachers often have difficulties using evidence from educational science when arguing for or against diagnostic judgments. They show empirically that diagnostic argumentation consists of three facets (justification of a diagnosis with evidence, disconfirmation of differential diagnoses, and transparency regarding the processes of evidence generation) and demonstrate that pre-service teachers often perform poorly on all three of these facets when arguing for or against certain diagnostic judgments.

Finally, two articles in this Research Topic stress the problem that pre- and in-service teachers sometimes *lack appropriate scientific knowledge and hold misconceptions* on important educational evidence that guide their decision-making processes:

The study by Surma et al. indicates that novice secondary school teachers hold widespread misconceptions regarding the effectiveness of different study strategies. Additionally, they demonstrate that novice teachers are unaware of specific strategies that educational research has shown to be effective (such as summarizing or spaced practice). These findings call for interventions that help novice teachers acquire scientific and sound knowledge on effective study strategies.

Similarly, Ferguson and Bråten show that for some topics (such as the alleged existence of learning styles), misconceptions and educational myths are also prevalent among Norwegian pre-service teachers, while for other topics, they seem to argue in an evidence-informed manner. Further, Norwegian pre-service teachers seem to focus especially on teacher behavior as cause for student performance and less on student factors. In general, participants rarely refer to educational research during pedagogical decision-making.

How can pre- and in-service teachers' evidence-informed reasoning be scaffolded?

The contributions introduced so far indicate a clear need for interventions that help pre- and in-service teachers develop their beliefs and competences regarding evidence-informed reasoning further. Six articles in this Research Topic investigate how such interventions could look like:

Rochnia and Gräsel investigate how to increase the utility value pre-service teachers attribute to educational sciences when solving pedagogical problems. Pre-service teachers read a short description of either empirical results or a theoretical reflection model which both illustrated the utility of educational sciences. Then, they had to summarize it or connect it to their own lives. While utility value was found to increase in all conditions, the four interventions did not differ in their effects.

Grimminger-Seidensticker and Seyda study how attitudes and self-efficacy toward inclusive teaching among physical education pre-service teachers can be supported. Their participants either received an information-based seminar, a seminar that combined theoretical input with practical exercises, or no training. While pre-service teachers' self-efficacy did not change in any condition, the intervention that combined theory input with practical exercises showed the most positive effects on some of their attitudes toward inclusive teaching.

Engelmann et al. focus on improving pre-service teachers' abilities to critically appraise scientific literature. After all participants were introduced to a set of criteria to appraise scientific evidence, they were provided with model solutions to several pedagogical problems and either explained them to a learning partner (interactive condition) or to themselves (constructive condition). While students' skills improved significantly from pre- to post-test, no differential effects for the experimental conditions appeared.

Krause-Wichmann et al. compared the effects of different sample solutions pre-service teachers received after they had analyzed an authentic classroom case. These solutions either included example-free or example-based instruction on functional procedures, and were combined with either example-free, example-based or no instruction on dysfunctional procedures. The authors find example-based instruction, both on functional and on dysfunctional procedures, to work best in order to help pre-service teachers develop their evidence-informed reasoning scripts further.

Lohse-Bossenz et al. report on the development of a vignette-based instrument to measure pre-service teachers' abilities to apply scientific knowledge in ambivalent educational situations. Participants were well able to spot the differences in the quality levels of teacher behavior that were described in the different vignettes. Further, a second study shows that an intervention that was designed to improve participants' theoretical knowledge led to a further increase in students' performance.

Tannert et al. investigate how best to scaffold pre-service teachers' conceptual knowledge and reasoning about video cases through signaling. In one condition, participants were informed about the use of signals within the videos, whereas students in the other condition were not. Results indicate that pre-service teachers from the informed condition acquired more conceptual knowledge than their uninformed counterparts, while there were no differential effects on reasoning.

How does in-service teachers' evidence-informed reasoning impact student performance?

Groß Ophoff et al. investigate the impact of teachers' engagement with educational research on student performance. Their findings show that students of teachers from schools with a strong climate toward research use indeed perform better than students of teachers working at schools that are less research-informed. Further, they find that trust among colleagues and organizational learning has a positive impact on research use climate, which in turn acts as a mediator for student performance.

Conclusions

The articles assembled in this Research Topic provide important answers on what evidence-informed reasoning is, what barriers pre- and in-service teachers face when engaging in evidence-informed reasoning, how their evidence-informed reasoning skills can be scaffolded, and what impact teachers' evidence-informed reasoning has on student performance. That way, the contributions in this Research Topic hold great potential to inform educational practice, research, and policy-making.

Author contributions

IK provided a first draft of this article and finalized the manuscript. MG, TK-W, and RS commented extensively. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Deutsche Forschungsgemeinschaft (KO 3462/3-1; STA 596/8-1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Greisel, M., Wekerle, C., Wilkes, T., Stark, R. and Kollar, I. (2023). Pre-service teachers' evidence-informed reasoning: do attitudes, subjective norms, and self-efficacy facilitate the use of scientific theories to analyze teaching problems? *Psychol. Learn. Teach.* 22, 20–38. doi: 10.1177/14757257221113942

frontiers | Frontiers in Education

TYPE Conceptual Analysis
PUBLISHED 26 August 2022
DOI 10.3389/feduc.2022.941410

Check for updates

OPEN ACCESS

EDITED BY
Ingo Kollar,
University of Augsburg, Germany

REVIEWED BY
Robin Stark,
Saarland University, Germany
Maria Zimmermann,
Humboldt University of Berlin,
Germany

*CORRESPONDENCE
Izaak Dekker
Izaak.dekker@gmail.com

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 11 May 2022
ACCEPTED 11 August 2022
PUBLISHED 26 August 2022

CITATION
Dekker I and Meeter M (2022)
Evidence-based education: Objections
and future directions.
Front. Educ. 7:941410.
doi: 10.3389/feduc.2022.941410

Evidence-based education: Objections and future directions

Izaak Dekker[1,2]* and Martijn Meeter[3]

[1]Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, Netherlands,
[2]Research Centre Urban Talent, Rotterdam University of Applied Sciences, Rotterdam, Netherlands,
[3]LEARN! Research Institute, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Over the past two decades, educational policymakers in many countries have favored evidence-based educational programs and interventions. However, evidence-based education (EBE) has met with growing resistance from educational researchers. This article analyzes the objections against EBE and its preference for randomized controlled trials (RCTs). We conclude that the objections call for adjustments but do not justify abandoning EBE. Three future directions could make education more evidence-based whilst taking the objections against EBE into account: (1) study local factors, mechanisms, and implementation fidelity in RCTs, (2) utilize and improve the available longitudinal performance data, and (3) use integrated interventions and outcome measures.

KEYWORDS

evidence-based education, evidence-based policy, randomized controlled trial, implementation science, evidence-informed education, context-centered research

Introduction

There is a global consensus about the value of good education. Educational science shows that teachers, programs, and methods can greatly influence learning gains. Policymakers are increasingly eager to prioritize investments in methods, trainings, and approaches that are proven to be most effective in line with the tenets of evidence-based education (EBE). This EBE movement coincided with enormous investments in education in the United States ("No child left behind" act in 2002 and the "Every Student Succeeds" act in 2015), the United Kingdom ("What works network" in 2013), China (Slavin et al., 2021), and recently some other European countries (e.g., National Program of Education in the Netherlands).

Yet, many educational scientists and educators seem reluctant to endorse EBE, and EBE seems to only slowly find its way into educational practice (Dagenais et al., 2012; Van Schaik et al., 2018; Joram et al., 2020). Critiques against EBE are numerous and highly cited. Scholars criticize the status of randomized controlled trials (RCTs) and generalizations based on them (e.g., Deaton and Cartwright, 2018; Morrison, 2021). Others question the cost-effectiveness of educational RCTs, or whether EBE restricts attention to those interventions that can be studied with RCTs (e.g., Cowen, 2019). A third strain of critique targeted the broader EBE paradigm and its moral implications for the teaching profession (e.g., Biesta, 2007, 2010; Wrigley, 2018). The sheer volume of criticism might deter practitioners from EBE. Many indeed opt for using a –seemingly middle-ground- position of "evidence *informed* education," although stakeholders often use the terms interchangeably (Nelson and Campbell, 2017).

Strikingly, there is limited dialogue between proponents and critics of EBE. Researchers aligned with the EBE movement have not always thoroughly dealt with criticism against EBE's preference for RCTs and the wider potential repercussions of EBE for the teaching profession. Slavin (2008, 2017, 2020) and Slavin et al. (2021) discussed a selection of the objections against RCTs and EBE but left others unanswered. On the other hand, some critics may have created a "straw man" by equating EBE with exclusive reliance on quantitative RCTs (e.g., Wrigley, 2018) and a technocratic view of the teaching profession (e.g., Biesta, 2010). The debate runs the risk of losing its intellectual use when the opposing sides divide into separate streams of scholarship. This conceptual article contributes to EBE and the educational research literature by analyzing the critiques against the EBE and its usage of RCTs, and by proposing ways forward that take these arguments into account.

The rise of evidence-based education

In a lecture on "Teaching as a research-based profession" in 1996 (published in 2000), Hargreaves compared the educational profession to the medical profession. Based on his comparison he proposed that it would improve education if, similar to medical science, practitioners could and would make more use of evidence. In an article that meant to define EBE Davies (1999) later stated that:

> educational activity is often inadequately evaluated by means of carefully designed and executed controlled quasi-experiments, surveys, before-and-after studies, high-observational studies, ethnographic studies which look at outcomes as well as processes, or conversation and discourse analytic studies that link micro structures and actions to macro level issues. Moreover, research and evaluation studies that do exist are seldom searched for systematically, retrieved and read, critically appraised for quality, validity, and relevance, and organized and graded for power of evidence (p. 109).

He went on to define the task of the EBE movement as: (1) the capacity and discipline of educators to pose answerable questions about education, know where to find evidence, assess the evidence, and determine its relevance to their educational needs, and (2) the power of educational scientists to establish sound evidence where it is lacking.

Slavin (2002) subsequently specifically addressed the need for large-scale experimental evaluations to answer questions about effectiveness. Causal relations cannot be directly seen, they have to be inferred from observations or measurements. The logic of controlled manipulation is the strongest way to support such an inference, and randomization with an adequate sample offers a method that enables a comparison between two groups that are the same except for receiving the treatment (Slavin, 2002; Duflo and Banerjee, 2017). In the minds of many, EBE became synonymous with such experiments (e.g., Newman, 2017; Cowen, 2019; Wrigley and McCusker, 2019). However, large-scale experiments are complicated and costly to execute, and although they have become more prevalent since 2000, they remain altogether a rare phenomenon in the educational field (Cook, 2007; Pontoppidan et al., 2018; Slavin, 2020). Moreover, as is clear from Davies' quote mentioned above, EBE is and should be broader than experimental studies or RCTs.

Objections to evidence-based education

The pleas of Davies (1999), Hargreaves (2000), and Slavin (2002) for more EBE stirred a rich variety of critiques from within the educational research community. Although EBE stands for both gathering evidence where it is lacking and improving the capacity of educators to make use of evidence, most criticism of EBE is targeted at its preference for RCTs. Perhaps this is due to the dominance of RCTs in the medical science which EBE emulates, or to Slavin's (2002) influential call for experimental research to determine "what works."

Cook (2002, 2007) summarized the objections to performing RCTs into (1) philosophical objections (e.g., experiments imply a descriptive theory of causation that is inferior to explanatory theories of causation), (2) practical arguments (e.g., offering a potentially beneficial intervention only to the treatment group generates inequity), (3) undesirable trade-offs (external vs. internal validity), (4) the objection that schools will not use experimental results, and (5) objections that favor other types of study designs (e.g., quasi-experiments, preferred by researchers who value design control over statistical control). Since Cook presented his "typology," several new objections and new insights regarding EBE and RCTs in education were published. Some build on arguments within the existing categories, other ontological, socio-economic, and normative objections seem to belong to altogether new categories (e.g., Biesta, 2007; Cowen, 2019).

This analysis builds on the articles from Cook but reorganizes the used categories in order to prevent conceptual overlap and make them more parsimonious. The scope of Cook's "Philosophical objections" is too wide since philosophy encompasses both epistemology and ethics. Cook places ethical arguments in the "practical" category, but the term "practical" is more easily associated with other concerns such as category 4 (schools will not use the results). Undesirable trade-offs (Cook's third category) can be of a epistemological nature, but could also be ethical, or practical. We therefore cluster all criticisms into three types. Objections are categorized as "epistemic" when they target methodological questions or assumptions and consequences at the level of philosophy of science (when do we know what causes what, for example). Socio-economic objections target the feasibility or repercussions of the EBE paradigm. Finally, normative objections are ethical by nature and object to the purpose (or lack thereof) of EBE.

Epistemic objections

Several critics have raised epistemic and methodological objections to RCTs within EBE. Deaton and Cartwright (2018) and Cartwright (2019) described how RCTs can only give us unbiased estimates when randomization does not generate a random imbalance on variables that are not measured in a baseline test and covariates or confounders are not correlated with the treatment. When the sample is a convenience sample, which is often the case, point estimates from the sample should not be generalized to the broader population or other populations (scaling up) or individuals (drilling down). Joyce and Cartwright (2020) add that external validity in education is problematic because, in their view, educational contexts have great influence on how treatments work. They suggest that educational researchers should therefore study why and how something might work in a specific context. This means studying potential support factors, derailers, and the local structures that afford necessary causal pathways in addition to average treatment effects (Joyce, 2019; Joyce and Cartwright, 2020).

These epistemic arguments point out the limitations of RCTs and urge for improved RCTs and the use of additional types of study designs. However, neither is incompatible with the EBE maxim that urges educators to use the best available evidence. In his treatise against the dominance of RCTs, Morrison grudgingly admits that "*pace* Churchill, the RCT is the worst form of design except for all the others" (2021, p. 211). In other words: there is potentially much wrong with RCTs, but even more with other designs as a method of inferring causal relationships. Contributions such as Joyce and Cartwright (2020) raise the standard for the educational sciences and EBE and urge both scholars, practitioners, and policymakers, to be more knowledgeable about the type of research that could ideally answer contextual questions. From this perspective, RCTs should be improved and be complimented by other types of research but still play a vital role.

There are more radical epistemic (and ontological) objections against EBE. Biesta (2007, 2010) argued that education is an "open and semiotic system," which he defines as "systems that do not operate through physical force but through the exchange of meaning" (Biesta, 2010, p. 496). What causes learning is influenced by many variables that cannot be controlled and depends on interpretations by learners. We can therefore not determine "causes" in a deterministic manner. Does this objection posit a real threat to EBE? All of society could be argued to be an open and semiotic system, so taken literally it would make experimentation in all of the social sciences impossible. However, the "semiotic" (interpretation-dependent) nature of education does not preclude experimentation. How educational interventions are interpreted may be subject to regularities, and these may then underlie replicable results. In lab experiments, researchers can attempt to manipulate the factors of interest and hold constant all other relevant ones. This is impossible in field experiments, and most social scientists are aware that many confounding variables could impact results (Duflo and Banerjee, 2017). The combination of lab and field experiments brings us as close as we can get to provisionally "proving" causal relationships. Replications of experimental studies, [which are estimated to constitute only 0.13% of studies of the articles in leading educational journals (Makel and Plucker, 2014)], would further consolidate the reliability of the findings. The remaining uncertainty is completely compatible with EBE's maxim of using "the best available evidence."

The interpretation-dependent nature of many educational interventions makes it valuable to study cognitive and affective factors and processes in addition to behavior. Over the past decades, several scholars therefore rightly pleaded for studying *mechanisms* as well as effects in order to understand why interventions might cause certain outcomes. This is one of the epistemological requirements of critical realism. Understanding the mechanisms that drive the effects of interventions increases the chance of successfully translating an intervention to another context. Several scholars accordingly developed theories that help us to predict and measure the interactions between interpretations and behavior. The theory of identity-based motivation, for example, is based on studies of how students interpret the role of school for their future identity (Oyserman et al., 2002, 2006; Oyserman and Destin, 2010). In lab and field experiments, Oyserman and her colleagues subsequently tested and showed how these interpretations can be altered. Because they tested every step in the mechanism and formulated how the implementation fidelity can be monitored (Oyserman, 2015; Horowitz et al., 2018), this intervention proved transferable to different contexts.

Another set of Biesta's objections targets the epistemology that EBE assumes. In his articles, Biesta proposes using Dewey's epistemology to ground educational science. Instead of using a representational model of knowledge (spectator view) we should use Dewey's transformational model which assumes that reality is constantly changing. The transformational epistemology asserts that it is only possible to determine in hindsight what worked but never what works, because of the changing nature of reality and because the experimental methods of science change or distort the very reality that they aim to measure.

Summarizing the epistemology for EBE as a "spectator view" is too simplistic and ignores the work done by philosophers of science such as Searle (e.g., 1999) and many others. EBE is usually grounded in critical or scientific realism which entails that (ontologically) the world can exist independently of the mind (or science) and that (epistemologically) theories about this world can be approximately true. Dewey's epistemology is problematic because it erroneously reduces the existence of all theoretical constructs (among which causality) to operational relations (Bulle, 2018). Reducing all theoretical constructs to operational relations means that concepts "can be grasped only in and through the activity which constitutes it" (Dewey, 1891, p. 144). Vygotsky aptly criticized Dewey's reduction of theoretical constructs to operational relations in the following manner:

"It is impossible, to assimilate the role of the work tool, which helps man subject natural forces to his will, with that of the sign, which he uses to act upon himself. The tool is externally oriented whereas the sign is internally oriented. Attempts to equate the sign with the external tool, as it is the case in John Dewey's works, lose the specificity of each type of activity, artificially reducing them into one" (Vygotsky, 1978, p. 53).

Dewey's pragmatist epistemology has, for these reasons, been cast aside in epistemology, psychology, and the natural sciences, but it is still foundational for some social-constructivist views that are present in teacher education (among which Biesta's criticism of EBE). According to Northrop (1946), pragmatism's presence in western teacher education led to an overestimation of practical work and an underestimation of theoretical mastery; undermining the obligation to master the subjects that one teaches. However, even if we, for the sake of the argument, followed this epistemology, it would still be compatible with learning from experiences and experiments (e.g., from RCTs). Inferring what will work from what worked can never be done with absolute certainty, but what has or hasn't worked in the past will often provide the best available evidence for both theorized causal and "operational" relations. Surely Biesta does not suggest ignoring evidence about what worked (toward a relevant purpose) in the past when we choose educational interventions. This would limit even the use of the professional judgment that Biesta propagates, as this is also based on previous experiences.

A final interesting epistemic objection to how RCTs are currently used in EBE was raised by Zhao (2017). He argued that educational researchers too often fail to take "side-effects" into account in their trials. If we narrowly focus on one learning outcome, we might fail to notice trade-offs. Emulating medical science, as EBE purports to do, should include using a wider range of relevant outcome measures in RCTs to monitor side effects. Zhao claims that even some of the most contested subjects in educational research might be "appeased" if we acknowledged the trade-offs of different interventions. Using direct instruction as a didactic teaching strategy leads to higher learning outcomes, but this fails to convince critics who instead value the potential "costs" to creativity or professional flexibility. Experiments that report on learning outcomes, as well as impact on creativity and curiosity, will be more constructive to the debate (Zhao, 2017). Studying potential side effects requires researchers to improve their study designs (e.g., to exploratively search for potential side effects qualitatively, track long-term effects, also measure student and teacher wellbeing, etc.) and be aware of potential trade-offs.

Socio-economic objections

Performing and replicating large-scale experimental evaluations is complicated and expensive (Morrison, 2019). Do they offer a good return on investment? Some scholars criticize EBE, and large-scale RCTs, for being ineffective in solving relevant questions to the field (e.g., Thomas, 2016). Lortie-Forgues and Inglis (2019) recently analyzed 141 of the large-scale (median n = 2,386) educational RCTs commissioned by the Education Endowment Foundation (EEF) and the National Centre for Educational Evaluation and Regional Assistance (NCEE) to assess the magnitude and precision of their findings. Unencouragingly, they found that some 40% of RCTs they analyzed produced uninformative results: results were consistent both with finding no effect at all, or with a large effect comparable to 1 year or maturation and instruction (Bloom et al., 2008). The interesting question that they raised was, "why?" They suggested three explanations: (A) the theory on which the programs are based is unreliable (B) the educational programs are ineffective because they have been poorly designed or implemented (C) the studies are underpowered because the outcome measures they use contain more "noise" than we previously assumed. Explanation C is similar to an underlying cause of the wider "replication crisis" in psychology and other sciences (Maxwell et al., 2015); replication studies with large enough sample size or better outcome measures would eventually "solve" the problem by filtering out null findings (and positive findings) that result from mere chance. In the other two cases (A and B), the field experiment is doing education as a whole a service—it is either showing that some intervention should not be used because it is based on faulty theories, or that it requires thorough attention to implementation. For this reason, it would be good if monitoring implementation fidelity became standard practice within the field. However, none of the explanations incentivize school leaders to fund a large-scale evaluation. Few school leaders feel for investing in a study that is likely to show that the efforts of their colleagues led to non-significant or small effects. This suggests a need for governments to reserve sufficient research funding to accompany educational innovation (Pontoppidan et al., 2018).

Cowen (2019) raised an interesting objection against the predominance of RCTs that evidence-based policy has caused. He observes that EBE allows policymakers to target interventions that teachers have to apply instead of policies which they are accountable for themselves. EBE favors teacher-level interventions over structural change of the educational system given that the effects of the latter are near-impossible to measure with an RCT. Letting teachers teach mathematics with certain didactics can be evaluated with an RCT, a structural overhaul of the educational system not. This "bias" does have an upside. Structural overhauls of the educational system come with great costs (both financial and mental) and peril; this in itself should be an argument to be more conservative when it comes to structural reorganizations than with classroom interventions. Moreover, Cowen (2019) points out that it could be solved if EBE would draw from the full range of available research techniques when it comes to studying potential benefits to structural changes to educational systems. This is, again, compatible with the EBE maxim to use the best available evidence.

Another way in which socio-economic objections about the costs of large-scale evaluations can be taken into account as well as possible is by properly weighing the effects that are found. Greenberg and Abenavoli (2017) and Kraft (2020) recently offered insightful suggestions on how our interpretation of experimental evidence should be improved. Many RCTs use outcome measures developed specifically to measure the expected effects (often in the form of a survey), and measure the effects, with standardized effect sizes (Cohen's *d* in particular), of targeted instead of universal interventions. Specifically designed outcome measures used shortly after the intervention inflate expectations of the effects on actual practical outcome measures such as standardized tests and long-term effects. Studying targeted interventions means using a more homogeneous sample, which by definition leads to smaller variance in the dependent variables and thus larger effect sizes (Greenberg and Abenavoli, 2017). Cohen's *d* does not take relative risks into account and therefore "overvalues" small-scale trials with low variance. The effects of universal interventions on standardized test outcomes have therefore often been undervalued compared to targeted interventions with specific outcome measures. Kraft (2020) suggests using a different interpretation of effect sizes that takes the design of the study (large-scale, heterogeneous sample, "real" outcome measures, etc.), costs per pupil, and scalability of the intervention into account. This should help us in making sense of large-scale RCT outcomes and help define what we should interpret as successful educational innovations.

Normative objections

Normative objections against EBE are targeted at the aims of EBE, the paradigm which it stands for, or the moral implications that it has. While epistemic and socio-economic arguments primarily address predominance of RCTs, normative arguments have mainly been aimed at the broader EBE paradigm. In a range of articles and books, Biesta (e.g., 2007, 2010) argued that EBE is misguided because education is not effect-driven but value-driven, it is an inherently normative profession. Learning should always be directed at some educational good. Biesta divides educational goods into three categories: qualification, socialization, and subjectification. According to Biesta EBE is misguided because it places too much emphasis on qualification and too little on subjectification, and because EBE will inherently value only those outcomes that can be measured.

There are two things to consider here. Are the goals of EBE misguided? And are there educational goods that cannot be measured? Every researcher should be transparent about outcome measures. Every society and school should likewise test transparent learning goals and outcomes with every single examination that is undertaken. Outcome measures such as reading and math achievement are prevalent because there is an overwhelming democratic consensus about their value. The more idiosyncratic and subjective goals become, being a good citizen, or being a good person even, the less democratic consensus can be found on what they are, how they can be taught, and how they should be measured. As soon as a social or personal educational good is agreed upon, researchers can study it as an academic performance measure. In elementary schools and secondary schools in most western countries, the educational goods are partly defined by democratic governments, and partly by schools that may be accountable to local districts (as, e.g., in Britain and the United States) or to parents (either through parent councils or when they compete for students with other schools). In post-tertiary education goals are largely determined by the teaching staff and representatives of a vocational field. Once a school or institution chooses a certain educational good, they will usually find ways to assess it. If a vocational school targeted at hotel management considers "hospitality" an important educational good, they will find ways to teach it, and also to assess it. If an art school wants its students to create authentic masterpieces incorporating personal subjectivity they will find a way to grade this. The problem of the educational researcher, how to measure educational goods for which there is no standardized test, is therefore shared by the teacher or curriculum designer, and a teacher's solution can also be used by the researcher. The argument of Biesta (2010) and others (e.g., Wrigley, 2018; Akkerman et al., 2021) rightly draws attention to the importance of outcome measures both in education and educational research. Their position becomes incompatible with EBE once they argue that there are educational goods about which there is public consensus, that you can teach to students, but cannot evaluate. The combination of these three premises is an argument against human ingenuity; it presupposes that teachers will not find a way to assess what they find important, and it seems an untenable position.

Discussion

Newton et al. (2020) offered a useful model for "pragmatic" EBE for practitioners. The final part of this analysis will build upon their model by suggesting three directions for furthering EBE based on the earlier discussed objections to EBE.

Context-*centered experiments*

RCTs and especially large-scale field experiments fulfill an important "deciding" role in the ecosystem of educational research. However, to realize this potential they should meet high standards of rigor (Morrison, 2021): among others, be based on theory, have sufficient power, use baseline measures, randomly assign, and use clear protocols. In addition to these regular standards, educational researchers conducting experiments should strive to meet three further standards that make experiments more useful to educational practice.

The first thing to consider is the context in which the experiment is conducted (Deaton and Cartwright, 2018). This means studying support factors, derailers, and the local structures that afford causally necessary pathways. Qualitative case studies or qualitative evaluations of these factors can be of great added value to field experiments. This allows us to not only learn if something worked in a specific context, but why it worked differently in several contexts.

Second, studying the causal step-wise process that explains how interventions work, will allow interventions to be applied more reliably and transparently. Interventions with a clear mechanism allow both researchers and teachers to look "under the hood" whenever an intervention is not producing the expected effects. "Replication with variation," studying both the outcome as well as the mechanisms, is a suitable way to do this (Locke, 2015).

Third, implementation should be an integral part of the research design (Moir, 2018). Implementation science has already been employed in clinical, health, and community settings, but is relatively new within education (Lyon et al., 2018). In a systematic review of the role of implementation fidelity in educational interventions, Rojas-Andrade and Bahamondes (2019) found that the different aspects of implementation fidelity, and particularly exposure and responsiveness, were linked to outcomes in 40% of the studies. There are many different implementation fidelity frameworks, one suitable example for the educational sciences is Horowitz et al. (2018) adaptation of the framework of Carroll et al. (2007). This framework suggests evaluating program differentiation (is the intervention different from what was done before in this context?), dosage (how much of the intervention did students receive?), adherence (did the students receive the intervention in the intended sequence?), quality of delivery (did the students experience the key points as true and easy to process?), and student responsiveness (how did the students react to the adherence and quality of delivery?).

All these standards surely do not make it easier, or less expensive, to conduct large-scale educational experiments. They should therefore preferably be used when a causal issue is important but either lacks evidence or when the evidence is contradictory (Cook, 2007). These high demands shall not always be met, but offer a standard to aspire to, in order to make educational experiments even more useful. The earlier referred to examples of the research into identity-based motivation (e.g., Horowitz et al., 2018), research into goal-setting theory (Morisano et al., 2010; Locke, 2015; Dekker, 2022), and recent rigorous experiments (e.g., Yeager et al., 2022a,b), fulfill several of these demands and show that steps toward this ideal are possible.

Play to the strengths of the educational domain

Many critics suggest that EBE is hard or even impossible because the educational domain is different from domains such as medicine or agriculture (Morrison, 2021). Some aspects of education do indeed make effectiveness studies complicated. Yet, there are also aspects that could potentially be beneficial to EBE.

Schools, colleges, and universities keep track of grades, status, and many other student and course variables. There is an abundance of longitudinal performance data already available to most schools, colleges, and universities. Grading itself is not free from bias and noise, but with the appropriate statistical methods (e.g., growth modeling, or multilevel growth modeling) predictors of performance change can be studied over time. These methods could improve our insight into long-term effects of RCTs or longitudinal studies where an experimental design is not possible or suitable for the question at hand. Although grades are important, they do not represent the only educational goods.

Additionally, most schools, colleges, and universities evaluate their lessons, curriculum, and teachers. These types of student evaluations can be targeted at anything, and could, potentially, have research value. Potentially, they rarely stand up to scholarly standards (Newton et al., 2020). They are rarely designed with the scientific rigor that the students who fill them out have to 1 day adhere to. EBE should not just be known for using or advocating experimental studies, it should be known for a more scientific approach to educational data as well.

One example of how this could be approached is the development of research into blended learning. Future studies into effective forms of blended learning can combine online user data with qualitative evaluations of onsite education and performance data, to configure the optimal blends of online and onsite education for specific courses.

Integrated interventions and outcome measures

Two critiques against currently used outcome measures could bolster EBE. Zhao (2017) proposed studying potential trade-offs of an intervention. Biesta (2007) argued that instead of asking "what works" and implying that the educational good is self-explanatory, educational researchers should ask which educational goods are at stake. This means reflecting on and taking responsibility for transparently chosen outcome measures (Akkerman et al., 2021). At the start of college, for example, students' performance and mental health are interrelated in several ways (e.g., Dekker et al., 2020). Interventions that aim to improve either learning outcomes or mental health during this phase should preferably monitor both to test whether the targeted outcome did not come at the expense of the other. Several scholars pursue to integrate these different aspects into the concepts themselves: Kuh et al. (2005), for example, proposed using the term student success to stand for a combination of academic achievement engagement, satisfaction and the acquisition of skills, etc. Schreiner (2010) similarly introduced the concept of academic thriving to stand for a combination of performance, community, and wellbeing. When possible, package interventions could target combinations of outcomes by addressing the underlying problems or motivation (e.g., Morisano et al., 2010; Schippers and Ziegler, 2019). In some cases, the potential trade-offs or side effects might be less known. In these cases, it would be wise to qualitatively explore whether students experienced any unpredicted effects from participating in the experiment.

Conclusion

In this article, we discussed the criticism against EBE and its preference for experimental studies. EBE stands for a combination of (1) the duty of educational professionals to raise answerable questions, search for evidence, assess it, and carefully apply it to practice and (2) the duty of educational researchers to provide rigorous evidence where it is lacking. Most of the criticism from the research community is directed at the implications of the second "duty" or the overarching pursuit of EBE. The arguments raised against EBE and the RCTs that often come with it call for a nuanced view on the usefulness of different types of research designs and disciplines. No argument, however, warrants ignoring the best available evidence when designing education. There are many problems to consider when interpreting outcomes from RCTs (e.g., they create only a probabilistic equivalence between the groups being contrasted, and then only at pre-test, and many of the ways used to increase internal validity can reduce external validity). Yet, in most instances, experimental studies offer the least unreliable estimators of effectiveness.

While reviewing higher education practices, Newton et al. (2020) describe how, even today, ineffective teaching practices and subjective student evaluations persist. The opposite of EBE is not RCT-free educational evidence, but practice based on no evidence at all, or a wrong application or interpretation of evidence. The recently growing evidence base from experimental studies can improve the influence of educational research on educational practice. Especially if they are conducted according to high standards of rigor. One risk that should be avoided though, is catering to a need for extremely brief answers to simplified questions: "what works?" Articles, reviews, and books that summarize research findings about what works into oversimplified promises fall short of delivering on their promises. As the philosopher Hilary Putnam supposedly put it: "a philosophy that can be put in a nutshell, belongs in one." Dumbing down and summarizing too much stimulates wrong interpretations of evidence.

Educational researchers that aspire to contribute to EBE have a responsibility to conduct rigorous research that takes both epistemic, economic and normative objections into account. Educational professionals, in turn, have a responsibility to be curious about, and carefully search and assess the available evidence.

Author contributions

ID: conceptualization and writing—original draft. MM: writing—review and editing. Both authors contributed to the article and approved the submitted version.

Acknowledgments

We wish to thank the two reviewers, Erik van Schooten, Ellen Klatter, Michaéla Schippers, Hannah Bijlsma, and Esther van Dijk for their helpful feedback.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Akkerman, S. F., Bakker, A., and Penuel, W. R. (2021). Relevance of educational research: An ontological conceptualization. *Educ. Res.* 50:9. doi: 10.3102/0013189X211028239

Biesta, G. J. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educ. Theory* 57, 1–22. doi: 10.1111/j.1741-5446.2006.00241.x

Biesta, G. J. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. *Stud. Philos. Educ.* 29, 491–503. doi: 10.1007/s11217-010-9191-x

Bloom, H. S., Hill, C. J., Black, A. R., and Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *J. Res. Educ. Effect.* 1, 289–328. doi: 10.1080/19345740802400072

Bulle, N. (2018). What is wrong with Dewey's theory of knowing. *Ergo Open Access J. Philos.* 5, 575–606. doi: 10.3998/ergo.12405314.0005.021

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., and Balain, S. (2007). A conceptual framework for implementation fidelity. *Implement. Sci.* 2:40. doi: 10.1186/1748-5908-2-40

Cartwright, N. (2019). What is meant by "rigour" in evidence-based educational policy and what's so good about it? *Educ. Res. Eval.* 25, 63–80. doi: 10.1080/13803611.2019.1617990

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educ. Eval. Policy Anal.* 24, 175–199. doi: 10.3102/01623737024003175

Cook, T. D. (2007). Randomized experiments in education: Assessing the objections to doing them. *Econ. Innov. New Technol.* 16, 331–355. doi: 10.1080/10438590600982335

Cowen, N. (2019). For whom does "what works" work? The political economy of evidence-based education. *Educ. Res. Eval.* 25, 81–98. doi: 10.1080/13803611.2019.1617991

Dagenais, C., Lysenko, L., Abrami, P. C., Bernard, R. M., Ramde, J., and Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: A review of empirical research. *Evid. Policy J. Res. Debate Pract.* 8, 285–309. doi: 10.1332/174426412X654031

Davies, P. (1999). What is evidence-based education? *Br. J. Educ. Stud.* 47, 108–121. doi: 10.1111/1467-8527.00106

Deaton, A., and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* 210, 2–21. doi: 10.1016/j.socscimed.2017.12.005

Dekker, I. (2022). *Academic thriving: Optimising student development with evidence-based higher education*, Ph.D thesis. Guildford: PURE.

Dekker, I., De Jong, E. M., Schippers, M. C., De Bruijn-Smolders, M., Alexiou, A., and Giesbers, B. (2020). Optimizing students' mental health and academic performance: AI-enhanced life crafting. *Front. Psychol.* 11:1063. doi: 10.3389/fpsyg.2020.01063

Dewey, J. (1891). "How do concepts arise from percepts?," in *The early works of John Dewey. Volume 3: 1889-1892. Early Essays And Outline Of A Critical Theory Of Ethics*, eds J. A. Boydston and G. E. Axetell (Carbondale: Southern Illinois University Press), 142–146.

Duflo, E., and Banerjee, A. (eds) (2017). *Handbook Of Field Experiments*. Amsterdam: Elsevier.

Greenberg, M. T., and Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *J. Res. Educ. Effect.* 10, 40–67. doi: 10.1080/19345747.2016.1246632

Hargreaves, D. H. (2000). "Teaching as a research-based profession: possibilities and prospects," in *Leading Professional Development In Education*, eds B. Moon, J. Butcher, and E. Bird (London: Psychology Press), 200–210.

Horowitz, E., Sorensen, N., Yoder, N., and Oyserman, D. (2018). Teachers can do it: Scalable identity-based motivation intervention in the classroom. *Contemp. Educ. Psychol.* 54, 12–28. doi: 10.1016/j.cedpsych.2018.04.004

Joram, E., Gabriele, A. J., and Walton, K. (2020). What influences teachers' "buy-in" of research? Teachers' beliefs about the applicability of educational research to their practice. *Teach. Teach. Educ.* 88:102980. doi: 10.1016/j.tate.2019.102980

Joyce, K. E. (2019). The key role of representativeness in evidence-based education. *Educ. Res. Eval.* 25, 43–62. doi: 10.1080/13803611.2019.1617989

Joyce, K. E., and Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *Am. Educ. Res. J.* 57, 1045–1082. doi: 10.3102/0002831219866687

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educ. Res.* 49, 241–253. doi: 10.3102/0013189X20912798

Kuh, G. D., Kinzie, J., Schuh, J. H., and Whitt, E. J. (2005). *Student Success In College: Creating Conditions That Matter*. Hoboken, NJ: Jossey-Bass.

Locke, E. A. (2015). Theory building, replication, and behavioral priming: Where do we need to go from here? *Perspect. Psychol. Sci.* 10, 408–414. doi: 10.1177/1745691614567231

Lortie-Forgues, H., and Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educ. Res.* 48, 158–166. doi: 10.3102/0013189X19832850

Lyon, A. R., Cook, C. R., Brown, E. C., Locke, J., Davis, C., Ehrhart, M., et al. (2018). Assessing organizational implementation context in the education sector: Confirmatory factor analysis of measures of implementation leadership, climate, and citizenship. *Implement. Sci.* 13, 1–14. doi: 10.1186/s13012-017-0705-6

Makel, M. C., and Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educ. Res.* 43, 304–316. doi: 10.3102/0013189X14545513

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am. Psychol.* 70, 487–498. doi: 10.1037/a0039400

Moir, T. (2018). Why is implementation science important for intervention design and evaluation within educational settings? *Front. Educ.* 3:61. doi: 10.3389/feduc.2018.00061

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O., and Shore, B. M. (2010). Setting, elaborating, and reflecting on personal goals improves academic performance. *J. Appl. Psychol.* 95, 255–264. doi: 10.1037/a0018478

Morrison, K. (2019). Realizing the promises of replication studies in education. *Educ. Res. Eval.* 25, 412–441. doi: 10.1080/13803611.2020.1838300

Morrison, K. (2021). *Taming Randomised Controlled Trials In Education: Exploring Key Claims, Issues And Debates*. Milton Park: Routledge.

Nelson, J., and Campbell, C. (2017). Evidence-informed practice in education: Meanings and applications. *Educ. Res.* 59, 127–135. doi: 10.1080/00131881.2017.1314115

Newman, J. (2017). Deconstructing the debate over evidence-based policy. *Crit. Policy Stud.* 11, 211–226. doi: 10.1080/19460171.2016.1224724

Newton, P. M., Da Silva, A., and Berry, S. (2020). The case for pragmatic evidence-based higher education: A useful way forward? *Front. Educ.* 5:583157. doi: 10.3389/feduc.2020.583157

Northrop, F. S. C. (1946). *The Meeting Of East And West: An Inquiry Concerning World Understanding*. New York, NY: The Macmillan Company.

Oyserman, D. (2015). *Pathways To Success Through Identity-Based Motivation*. Oxford: Oxford University Press.

Oyserman, D., Bybee, D., and Terry, K. (2006). Possible selves and academic outcomes: How and when possible selves impel action. *J. Pers. Soc. Psychol.* 91:188. doi: 10.1037/0022-3514.91.1.188

Oyserman, D., and Destin, M. (2010). Identity-based motivation: Implications for intervention. *Couns. Psychol.* 38, 1001–1043. doi: 10.1177/0011000010374775

Oyserman, D., Terry, K., and Bybee, D. (2002). A possible selves intervention to enhance school involvement. *J. Adolesc.* 25, 313–326. doi: 10.1006/jado.2002.0474

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O. J., Opheim, V., Gustafson, S., et al. (2018). Randomised controlled trials in Scandinavian educational research. *Educ. Res.* 60, 311–335. doi: 10.1080/00131881.2018.1493351

Rojas-Andrade, R., and Bahamondes, L. L. (2019). Is implementation fidelity important? A systematic review on school-based mental health programs. *Contemp. Sch. Psychol.* 23, 339–350. doi: 10.1007/s40688-018-0175-0

Schippers, M. C., and Ziegler, N. (2019). Life crafting as a way to find purpose and meaning in life. *Front. Psychol.* 10:2778. doi: 10.3389/fpsyg.2019.02778

Schreiner, L. A. (2010). The "thriving quotient": A new vision for student success. *About Campus* 15, 2–10. doi: 10.1002/abc.20016

Searle, J. R. (1999). *Mind, Language And Society: Philosophy In The Real World*. New York, NY: Basic Books.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educ. Res.* 31, 15–21. doi: 10.3102/0013189X031007015

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educ. Res.* 37, 5–14. doi: 10.3102/0013189X08314117

Slavin, R. E. (2017). Evidence-based reform in education. *J. Educ. Stud. Placed Risk* 22, 178–184. doi: 10.1080/10824669.2017.1334560

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educ. Psychol.* 55, 21–31. doi: 10.1080/00461520.2019.1611432

Slavin, R. E., Cheung, A. C. K., and Zhuang, T. (2021). How could evidence-based reform advance education? *ECNU Rev. Educ.* 4, 7–24. doi: 10.1177/2096531120976060

Thomas, G. (2016). After the gold rush: Questioning the" gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harv. Educ. Rev.* 86, 390–411. doi: 10.17763/1943-5045-86.3.390

Van Schaik, P., Volman, M., Admiraal, W., and Schenke, W. (2018). Barriers and conditions for teachers' utilisation of academic knowledge. *Int. J. Educ. Res.* 90, 50–63. doi: 10.1016/j.ijer.2018.05.003

Vygotsky, L. S. (1978). "Internalization of higher psychological functions" in *Mind in society: The development of higher psychological processes*, eds M. Cole, V. John-Steiner, S. Scribner and E. Souberman (Cambridge, MA: Harvard University Press), 52–58.

Wrigley, T. (2018). The power of 'evidence': Reliable science or a set of blunt tools? *Br. Educ. Res. J.* 44, 359–376. doi: 10.1002/berj.3338

Wrigley, T., and McCusker, S. (2019). Evidence-based teaching: A simple view of "science". *Educ. Res. Eval.* 25, 110–126. doi: 10.1080/13803611.2019.1617992

Yeager, D. S., Bryan, C. J., Gross, J. J., Murray, J. S., Krettek Cobb, D., Santos, P. H. F., et al. (2022a). A synergistic mindsets intervention protects adolescents from stress. *Nature* 607, 512–520. doi: 10.1038/s41586-022-04907-7

Yeager, D. S., Carroll, J. M., Buontempo, J., Cimpian, A., Woody, S., Crosnoe, R., et al. (2022b). Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychol. Sci.* 33, 18–32. doi: 10.1177/09567976211028984

Zhao, Y. (2017). What works may hurt: Side effects in education. *J. Educ. Chang.* 18, 1–19. doi: 10.1007/s10833-016-9294-4

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/feduc.2022.909255

Check for updates

OPEN ACCESS

EDITED BY
David Pérez-Jorge,
University of La Laguna, Spain

REVIEWED BY
Stamatios Papadakis,
University of Crete, Greece
Matias Arriagada-Venegas,
Universidad de Concepcion, Chile

*CORRESPONDENCE
Elke Grimminger-Seidensticker
elke.grimminger.seidensticker@uni-paderborn.de

SPECIALTY SECTION
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

RECEIVED 31 March 2022
ACCEPTED 22 August 2022
PUBLISHED 08 September 2022

CITATION
Grimminger-Seidensticker E and Seyda M (2022) Enhancing attitudes and self-efficacy toward inclusive teaching in physical education pre-service teachers: Results of a quasi-experimental study in physical education teacher education. *Front. Educ.* 7:909255.
doi: 10.3389/feduc.2022.909255

Enhancing attitudes and self-efficacy toward inclusive teaching in physical education pre-service teachers: Results of a quasi-experimental study in physical education teacher education

Elke Grimminger-Seidensticker[1]* and Miriam Seyda[2]

[1]Department of Exercise and Health, Faculty of Science, University of Paderborn, Paderborn, Germany, [2]TU Dortmund University, Dortmund, Germany

As many teachers feel overwhelmed by teaching inclusively, teacher education programs have to find ways to prepare them for this future challenge. Due to particular conditions, physical education (PE) teachers might be a particular target group in this context. With regard to the state of research, there is a need to identify inputs or learning situations that might improve the best physical education teachers' competencies and underlying cognitive and affective-motivational aspects, such as attitudes, self-efficacy, or stress perception, to empower physical education teachers for teaching physical education inclusively. Practical experiences seem to be a key aspect in this context. Therefore, we conducted a quantitative evaluated quasi-experimental intervention study with physical education pre-service teachers to test different forms of promoting inclusion competencies and their underlying cognitive and affective-motivational constructs. Intervention group 1 (IG 1) followed an information-based seminar, whereas intervention group 2 (IG 2) was also taught theoretical units in combination with practical lessons in the gym that were prepared and conducted by the pre-service teachers themselves. The control group (CG) did not receive any specific information or practical experiences on inclusively teaching physical education. We first hypothesized that both intervention groups (IG 1 and IG 2), in contrast to the control group, would significantly improve their attitudes toward inclusion and the self-efficacy to teach inclusively, and would decrease their perceived stress related to teaching physical education inclusively (hypothesis 1). Second, we hypothesized that participants of intervention group 2 would have a significantly stronger increase in positive attitudes toward inclusion, a stronger increase in self-efficacy, and a greater decrease in the level of perceived stress related to teaching physical education inclusively than participants of intervention group 1 (hypothesis 2). Based on ANCOVA analysis, we found significant results for some subscales of the attitudes, but no significant results for stress perception and self-efficacy. In total, the teaching strategy in intervention group 2 seemed to work best in enhancing physical education pre-service teachers' inclusion competencies.

KEYWORDS

inclusive education, physical education, pre-service teachers, intervention study, teacher education

Introduction

To date, inclusion and inclusive education have been implemented in schools in numerous countries (Tant and Watelain, 2016; Harant, 2017). Based on the United Nations Convention on the Rights of Persons with Disabilities, which was ratified in Germany in 2009, inclusive schooling of all children is mandatory. In the context of education, inclusion can be defined as the "means of increasing participation in learning by all students so that their educational needs can be met" (Qi and Ha, 2012, p. 258). In contrast to this all-encompassing understanding of inclusion, schools and educational reforms often only focus on the aspect of inclusion which denotes the integration of pupils with disabilities into regular schools. This has subsequently led to the directive that "persons with disabilities can access an inclusive, quality and free primary education and secondary education on an equal basis with others in the communities in which they live" (United Nations [UN], 2014 Article 24–Education, 2.b).

While discussing inclusion and inclusive education in schools, the special role of teachers and their competencies has to be considered. Following Blömeke et al. (2015), we understand competencies as complex ability constructs that are context-specific and measurable either by cognitive and affective-motivational aspects (dispositional view) and are finally observable in performance (behavioral view). Conceptually, competencies can be modeled as a continuum. They are learnable and can thus be improved, e.g., *via* continuing education or teacher education programs at university. However, to successfully implement inclusion at school, teachers not only need adequate (continuing) education and sufficient support (e.g., special equipment) but also positive attitudes toward inclusion (Frankel et al., 2010) as well as self-efficacy (Block et al., 2013). With respect to the success of inclusion, to a large extent it depends on the "willingness of teachers to cater to the pupils with disability in their class" (Gilor and Katz, 2017, p. 294). Studies have shown that a positive attitude toward inclusion has a positive effect on teaching and the integration of students with special educational needs in regular classes (Avramidis and Norwich, 2002; Sharma et al., 2008; De Boer et al., 2011; Killoran et al., 2014; Lüke and Grosche, 2018b). However, the school subject and the form of disability affects teachers' attitudes toward inclusion: physical education (PE) teachers are less favorable toward including children with orthopedic, visual or mental impairment in their PE classes, whereas they are ambivalent about including children with emotional and behavioral difficulties (Hutzler et al., 2019). As attitudes toward inclusion have to be focused on a specific school subject, we are focusing our paper on PE (pre-service) teachers. In this context, studies have shown that even if PE (pre-service) teachers have positive attitudes toward inclusion, they do not feel well prepared to teach inclusively (Reuker et al., 2016 for Germany; in the international context: Morley et al., 2005; O'Brien et al., 2009), and they feel hardly any self-efficacy in teaching children with emotional and behavioral difficulties or mental, visual, or hearing impairment (Leineweber and Thomas, 2017). Fejgin et al. (2005) even point out a positive relationship between the number of students with special educational needs in a PE class with the PE teachers' experienced burnout.

Thus, we need to identify inputs or learning situations that might improve the best PE teachers' competencies and underlying cognitive and affective-motivational aspects, such as attitudes, self-efficacy, or stress perception, to empower PE teachers for teaching PE inclusively. "Teaching is currently seen as an overwhelming profession and it is important to bring beginning teachers to the place where they feel they are capable and will be more emotionally equipped to take on the stressors of the classroom" or in our context, of the gym (Specht and Metsala, 2018, p. 69). Furthermore, pre-service teacher education programs may be the optimal time to address future teachers' attitudes and concerns about inclusive education, trying to change them in a positive way (Sharma et al., 2006; Specht and Metsala, 2018). In this context, Lautenbach and Heyder (2019) were able to identify 23 studies in *general* teacher education that attempted to promote pre-service teachers' attitudes and self-efficacy, using different teaching strategies (purely theoretically driven, a combination of theoretical inputs and practical experiences or purely practical experiences), and evaluated the impact of these teaching strategies in pre-and post-intervention designs. The study results were inconsistent, showing mostly positive changes in pre-service teachers' attitudes, no changes or seldom negative changes. However, the biggest limitation of these previous studies is the absence of control groups in the study design. Thus, it is difficult to link changes to the educational input (Lautenbach and Heyder, 2019).

Although, we could suggest that particular inputs in teacher education seminars could have a positive impact on attitudes toward inclusion as well as on self-efficacy, there is still a need to examine the extent of changes in different seminar designs, and with a special focus on PE as a particular school subject compared to other subjects. Teaching PE takes place under particular conditions: (1) a large space, where many students move around in different ways and places at the same time, which requires a particular kind of classroom or better gym management to guarantee corporal and emotional safety of all students (e.g., Grube et al., 2018). (2) Particular requirements regarding teachers' own physical conditions, like own sport motor competencies or volume of the voice (e.g., Brouwers et al., 2011). (3) The openness of learning situations in PE offers students more possibilities of displaying and negotiating peer relationships. In this context, heterogeneity might result in exclusion and mobbing processes if PE teachers are unable to handle, in an appropriate way, differences in sport motor competencies or in social popularity within their students (e.g., Grimminger, 2013, 2014). (4) PE is a field where the ability or disability of moving in the right way or having the societally accepted body shape is constantly displayed and that might be experienced as a humiliating context if PE teachers are not sensitive toward embarrassing situations (Kerner et al., 2018). This is mostly the case if PE teachers follow traditional sport approaches where norms of performances and the "right" body are anchored. These traditional norms might be a barrier for teaching PE inclusively (Giese and Ruin, 2018). (5) Another barrier for including students with special educational needs in PE is the fear for pupils' safety and the possible negative impact on peers (e.g., Ko and Boswell, 2013). Thus, PE seems to be a particular school subject in the context of inclusion and PE teacher education needs to be tailored to these particular aspects.

In particular, in recent years, there is an increase in the number of intervention studies to enhance PE (pre-service) teachers' attitudes, self-efficacy, or competencies in teaching PE inclusively. The studies vary in length, formats of teaching courses (only theoretically, combination of theory and practice in different contexts) as well as in the study design (implementation of a control group or not). In this context, Barber's case study (2018) is one of the rare qualitative studies. The author took up the challenge of traditional sport approaches and with it linked norms and values of (dis-)ability, and developed a "program modification in pre-service PE teacher designed to interrupt misconceptions and to construct new understandings of "dis"ability, to assist teacher education students in beginning to develop a philosophy of full inclusion" (p. 520). The intervention consisted of a one-day visit to an "Abilities Center" where able-bodied and disabled individuals interacted with each other. As part of the visit, the pre-service teachers participated in para sport sessions. The impact of this visit on the PE pre-service teachers' attitudes toward inclusion and their perception of (dis-)ability was examined with pre- and post-focus groups, videography of student reflections, and individual interviews. Barber (2018) noted relatively minor changes in the pre-service teachers' attitudes. However, changes in the perceptions of (dis-)ability could be identified. The participants became more open to understanding the potential of the variety of movement experiences for inclusion. Nevertheless, this change was not reflected in their approach to lesson planning. All the other studies are quantitative intervention studies, like Taliaferro and Harris (2014) who examined the impact of a one-day workshop on PE teachers' self-efficacy to include students with autism spectrum disorder. The changes in the intervention group were not significant when compared to the control group. Additionally, a non-significant impact on judgments about inclusion was shown by a two-day workshop with PE teachers (Haegele et al., 2018; no control group was implemented). Therefore, we could conclude that the length of an intervention might be crucial. Subsequently, Taliaferro et al. (2015) examined the effect of participation in one of two 15-week adapted physical education courses, with a nine-week experience practicum, on pre-service teachers' self-efficacy toward the inclusion of students with special educational needs. The results showed a significant increase in the pre-service teachers' self-efficacy across all disability categories. Therefore, these results implicated the importance of a certain length of intervention as well as the relevance of practical experiences. However, as the researchers did not enroll a control group in their study, the precise effect of the intervention may be unclear. Thus, Reina et al. (2019) ran a pre- and post-test quasi-experimental intervention study with PE teachers with an intervention group, who followed the so-called *Incluye-T* training program, and with a control group who did not participate in this special program. Each of the six face-to-face sessions (lasting for 3 h) combined a theoretical component with a practical component, when PE teachers were asked to modify activities, equipment, and instruction for students with special educational needs, via the use of simulations. Physical disabilities, intellectual disabilities, and visual impairment were addressed. The researchers examined the effect of this special training program on PE teachers' self-efficacy and could note significant improvements for the participants in the intervention group compared to the control group. The development was not affected by school type or gender. In their follow-up study, Reina et al. (2021) could even prove the impact of a teacher designed and implemented disability awareness program on the attitudes of students toward inclusion. After following the *Incluye-T* training program, the PE teachers developed and implemented their own disability awareness activities in their regular PE lessons. The results showed that combined activities that focus on several disabilities/impairments had the highest impact on students' disability awareness.

As pre-service PE teachers do not yet have the same practical experiences as PE teachers who participate in an in-service training, the design and content of a training program for pre-service PE teachers should be particularly tailored. Furthermore, PE in higher education is a subject for which almost half of the teaching courses are practical sport courses in which pre-service teachers develop mostly their own sport performance as well as their competencies to teach a certain sport discipline (Erhorn et al., 2020). Thus, Zach et al. (2012) concluded that extending field experiences as well as movement and sport classes that emphasize the practice of teaching methods improve PE pre-service teachers' self-efficacy. Hopkins et al. (2018) also underlined the potential of fieldwork experiences for the development of a teacher identity. Thus, Lautenbach et al. (2020) implemented a quasi-experimental intervention study with PE pre-service teachers to test the effect of a theoretically based seminar in comparison to a theoretically based seminar with practical experiences, i.e., participation in inclusive sport groups, in comparison to a control group with no particular input about inclusion. The researchers aimed the enhancement of the pre-service teachers' attitudes and self-efficacy toward inclusion, and the reduction of stress perception of teaching inclusively. Although the results underlined the advantage of the combined theory-practical seminar, participating in an inclusive sport group might not be enough with respect to PE pre-service teachers' prospective jobs that will require teaching inclusively.

Finally, we can conclude that previous intervention studies on the enhancement of (pre-service) PE teachers' attitudes, self-efficacy, and teaching competencies in an inclusive setting underline the importance of the intervention length as well as the relevance of the combination of theoretical inputs with practical experiences that should be pedagogically accompanied and reflected upon. Nevertheless, we do not know yet which kind of practical experience might be successful in pre-service PE teacher education. Reina et al. (2019) showed the impact of combining theory with teaching practices for the inclusion of students with physical disabilities, intellectual disabilities, or visual impairment for PE teachers. However, is this teaching strategy also successful for pre-service PE teachers with less teaching experience? And finally, is the simulation approach in the practical units the only way to translate theory into practice?

Following these principal questions, we conducted a quasi-experimental intervention study with two different intervention groups (IG 1 and IG 2) and a control group (CG). We wanted to examine the following research question: Which form of intervention might have a greater impact on attitudes toward inclusion, on stress perception, and on the self-efficacy to teach PE inclusively among PE pre-service teachers' Therefore, the intervention groups underwent two different teaching programs for the enhancement of important influencing factors, like attitudes toward inclusion and self-efficacy in teaching inclusively, as well as for a reduction in perceived stress regarding an inclusive physical education class. Whereas IG 1 followed an information-based seminar, including theoretical units on how to plan PE lessons in an inclusive PE context, IG 2 was also taught these theoretical units and additionally put the theoretical aspects to the test in practical lessons in the gym. The PE pre-service teachers were responsible for these practical lessons by planning and conducting the lessons with their fellow pre-service teachers. Their planning was guided by current recommendations for designing inclusive physical education classes (e.g., Tiemann, 2015; Giese and Weigelt, 2017), taking into account the type of impairment (e.g., learning disabilities, emotional and behavioral difficulties, physical impairment, chronic diseases and intellectual impairment) that should be specifically addressed in each lesson. Like in the study of Reina et al. (2019), the pre-service teachers should find pedagogical adaptions or modifications, e.g., in content, rules or material, with regard to the particular impairment. However, in contrast to Reina et al. (2019), their fellow pre-service teachers were not asked to simulate the impairment for various reasons: (1) Not all types of impairment can be simulated, or the simulation might possibly result in a display of stereotypes; e.g., how a child with emotional and behavioral difficulties might behave in PE in the pre-service teachers' view. Finally, stereotypes should not be actively initiated; instead, they should be critically reflected upon to sensitize pre-service teachers to their own stereotypes in relation to teaching behavior (Macrae et al., 1994). (2) The aim of the intervention study was not to help pre-service PE teachers experience what it is like to have a disability; that is the aim of the disability simulation approach (McGowan, 1999). The aim of the intervention study was to sensitize pre-service PE teachers to pedagogical strategies for teaching PE inclusively. (3) The effect of the disability simulation approach on attitudes toward inclusion is ambivalent; it could have also a negative impact as participants without disabilities could experience the simulation as overwhelming due to missing coping strategies (e.g., Flower et al., 2007). The lessons planned and conducted by the pre-service teachers were reflected upon afterward by all seminar participants and were also commented upon by the university lecturer. The aim of this teaching scenario was to give pre-service teachers mastery experiences, as they seem to be crucial predictors of teachers' self-efficacy regarding inclusive teaching (Wilson et al., 2020). The control group was only theoretically taught how to plan a physical education lesson without specific information about inclusive pedagogical concepts.

The study was theoretically based on the theory of planned behavior (Ajzen, 1991) and the self-efficacy theory of Bandura (1977) to link attitudes with behavior. Thus, it could be argued that attitudes toward inclusion can theoretically as well as empirically be linked to teaching behavior (Yeo et al., 2014). However, teachers' attitudes toward inclusion depend on different contextual variables, like gender, age, length of professional experience, amount and quality of teachers' acquaintance with persons with disabilities, amount and quality of teachers' academic training and practicum, the degree of perceived self-efficacy or competence, the impact of children's disability attributes, and the school environment including type or level of school (e.g., Specht and Metsala, 2018; Hutzler et al., 2019). In the theory of planned behavior (Ajzen, 1991), self-efficacy is one of the factors besides personal external factors, attitudes and subjective norms that influences the formation of a behavior intention that might lead to a concrete behavior. As Savolainen et al. (2020) have shown in a longitudinal cross-lagged study that self-efficacy influences attitudes toward inclusion, the enhancement of self-efficacy of (pre-service) teachers seems to be crucial. Specht and Metsala (2018) also strengthen that especially in pre-service teacher programs, positive experiences in inclusive educational settings are important for the development of self-efficacy. Bandura (1977) hypothesized that self-efficacy determines whether individuals act, how much effort they will expend and how long they will continue acting in the face of obstacles and failures. According to Bandura (1977), self-efficacy makes a difference in how people think, feel and act. Teachers' self-efficacy is a domain-specific aspect (in contrast to general self-efficacy) and can be defined as "the teacher's belief in her or his ability to organize and execute the courses of action required to successfully accomplish a specific task in a particular context" (Tschannen-Moran et al., 1998, p. 233). Following Bandura (1977), self-efficacy can be increased by four sources: (1) mastery experiences or personal experiences (e.g., performance exposure), (2) vicarious experiences (e.g., observing a model performing a task that is within one's own abilities), (3) social persuasion (e.g., verbal messages or social encouragement) and (4) physiological and/or affective responses (e.g., state of anxiety). In this context, Martins et al. (2015) showed that pre-service teachers with high self-efficacy highlight during a practicum, for example, planning and teaching practice as a mastery experience, lesson observation as a vicarious experience and post-lesson discussions as persuasion, whereas pre-service teachers with low self-efficacy might associate the same learning situations with negative emotions. Thus, self-efficacy can be a resource to deal with stress (Lazarus and Folkman, 1984). Stress occurs if the subjective appraisal of the relevance of a stressor and the appraisal of the individual's perceived resources to deal with this stressor lead to the conclusion that the resources are not enough to handle the stressor. Hutzler et al. (2005) have suggested that self-efficacy influences attitudes toward inclusion moderated by stress perception. Increasing self-efficacy could therefore reduce perceived stress, enhance attitudes toward inclusive classes and teaching inclusively and, finally, improve concrete teaching behavior in inclusive classes.

With regard to the previous intervention studies in PE teacher education, we wanted to examine in our quasi-experimental study which form of intervention might have a bigger impact on attitudes toward inclusion, on stress-perception and on the self-efficacy to teach inclusively among PE pre-service teachers. Based on the results of previous intervention studies, we first hypothesized that both intervention groups (IG 1 and IG 2), in contrast to the control group, would significantly improve their attitudes toward inclusion and the self-efficacy to teach inclusively, and would decrease their perceived stress related to teaching PE inclusively (hypothesis 1). Second, we hypothesized that participants of IG 2 would have a significantly stronger increase in positive attitudes toward inclusion, a stronger increase in self-efficacy, and a stronger decrease in the level of perceived stress related to teaching PE inclusively than participants of IG 1 (hypothesis 2).

Materials and methods

Participants

In total, 86 PE pre-service teachers (mean age = 25.17; *SD* = 2.24; 34 females) from two universities were enrolled in the study, with IG 1 and CG at one university (taught by two different teachers) and IG 2 at another university. The groups did not differ in gender distribution, age and practical experience with inclusion at school as well as outside school (e.g., volunteer service), meaning previous contact or experiences with persons with disabilities (only operationalized as "yes or no" and not measured with respect to intensity or quality of experience). They differed, however, in their theoretical experiences with the topic of inclusion after counting the number of seminars previously attended that dealt with the topic of inclusion. IG 2 had significantly fewer pre-service teachers without previous theoretical experience than the other groups. **Table 1** gives an overview of the study population at pre-measurement.

Participant recruitment was carried out through surveys in compulsory master's seminars at the two universities. Allocation

TABLE 1 Overview of the study population at pre-measurement.

	IG 1 (*n* = 49)	IG 2 (*n* = 17)	CG (*n* = 20)
Gender	20 females*	9 females	5 females
	27 males*	7 males	12 males
Age	mean age = 25.49 (SD = 2.23)	mean age = 24.41 (SD = 1.91)	mean age = 25.00 (SD = 2.50)

* Two persons did not indicate their gender.

to the groups was performed by pre-service teachers' selection of the respective seminar. Participation in the study was voluntary. The pre-service teachers were told that non-participation had no negative consequences for them and that they could quit the study at any time. The study protocol was approved by the ethics committee of TU Dortmund University (approval number: 2017 3).

Data collection

To test the explicit attitudes toward inclusion, we used two questionnaires to measure attitudes on a macrosystem level and on a classroom level. For the macrosystem level, we implemented the questionnaire on "Attitudes toward an Inclusive School System" (Lüke and Grosche, 2018a; 5-point Likert scale). In accordance with the definition of attitudes as constructs that consist of affective, cognitive and behavior-related components, this questionnaire includes three subscales: "emotion" (five items; example item: "It would be great if all children could be taught in an inclusive school system"; α = 0.69), "behavioral intentions" (five items; example item: "I would be willing to participate actively in the development of an inclusive school system"; α = 0.69) and "cognition" (10 items; example item: "I think an inclusive school system wouldn't be able to get all children to their best individual performances"; α = 0.83). To measure attitudes toward inclusion at the classroom level, we included the "Questionnaire on Attitudes toward Inclusion for Teachers" (Seifried and Heyl, 2016; 6-point Likert scale) with its three subscales: "promotion of academic competencies" (six items; example item: "Children with special needs will be equally supported in both an inclusive class and in a special needs class"; α = 0.85), "willingness to teach inclusively" (five items; example item: "I can imagine teaching an inclusive class next term"; α = 0.71) and "social inclusion" (four items; example item: "Children with special needs will be treated well by other children in an inclusive class"; α = 0.51; H = 0.87). The level of subjectively perceived stress when imagining teaching PE inclusively was measured using the validated Primary Appraisal Secondary Appraisal Scale (PASA) (Gaab, 2009; 6-point Likert scale). The PASA consists of four subscales assessing "challenge" (four items; example item: "The situation is a challenge for me"; α = 0.42; H = 0.90) and "threat" (four items; example item: "This situation scares me"; α = 0.73), which forms the primary appraisal scale (α = 0.44; H = 0.79). The secondary appraisal scale (α = 0.80) consists of the subscale "self-concept of own competencies" (four items; example item: "I know what I have to do in this situation"; α = 0.52; H = 0.79) and the subscale "control expectancy" (four items; example item: "I can control a lot myself of what I can do in this situation"; α = 0.74). A stress index as an indicator for stress perception can be calculated by subtracting the secondary appraisal from the primary appraisal mean scores (Gaab, 2009). A higher stress index indicates a higher subjective level of stress. Although the PASA questionnaire examines aspects of self-efficacy in the particular context of teaching PE inclusively, we measured the general teacher-related self-efficacy by the Teacher Self-Efficacy Scale (Schwarzer and Jerusalem, 1999; 4-point Likert scale). This scale consists of 10 items [example item: "I am convinced that I can develop creative ways to cope with system constraints (such as budget cuts and other administrative problems) and continue to teach well"; α = 0.77]. As Cronbach's alpha values of some scales might indicate a low consistency (α < 0.60), we calculated the Coefficient H for the maximal reliability of these scales. Coefficient H tends to provide the highest estimates of internal consistency. It is not affected by the addition of poor items because its intended use is for optimally weighted scales. In optimally weighted scales, items are differentially weighted, so an unrelated item does not affect reliability (McNeish, 2018). The calculated Coefficient H for all the scales with a low Cronbach's alpha value were good. Therefore, we can also presume the reliability of these scales.

The time between the pre- and post-questionnaire was fifteen weeks, and it was used for a specific seminar structure (90-min sessions per week). This was designed to deal specifically with the topic of inclusion in the intervention groups, while the control group had no explicit units on the topic. The two intervention groups differed in the special design of the seminar. While IG 1 dealt with the topic only on a theoretical level, IG 2 focused on a combination of practical and theoretical examinations of inclusive situations as described above.

Treatment procedures

IG 1 received a theoretical introduction to the Universal Design for Learning (CAST, 2011; Hall et al., 2012) as well as the 6 plus 1 model (Tiemann, 2016) as guidelines for the planning and analyses of inclusive PE classes. The topic of inclusive language was introduced with the example of formulating rules for a soccer play in PE class; this idea was then transferred to the general wording in PE exercises. Furthermore, methods of individual promotion were theoretically discussed.

IG 2 first got a theoretical introduction to basic terms and teaching concepts of PE with respect to inclusion (Kullmann et al., 2014; Tiemann, 2015, 2016), heterogeneity and individual promotion in physical education (Fediuk, 2008), reflective co-education (Gieß-Stüber, 2012; Kastrup and Kleindienst-Cachay, 2016) and intercultural competence (Grimminger-Seidensticker and Möhwald, 2017). Hereby, we conceptually understood inclusion as the creation of a positive and meaningful learning environment for all students, not only for those with special educational needs (Overton et al., 2017). In the second part of the master's seminar, a practical examination was the focus. The pre-service teachers were asked to plan and implement with their fellow pre-service teachers teaching sequences in which students with and without specific educational needs were taught inclusively. The teaching sequences were reflected upon after implementation. Different specific educational needs were addressed as examples: learning disabilities, emotional and behavioral difficulties, physical impairment, chronic diseases, and intellectual impairment. In the five practical units of the seminar, the respective procedure was as follows: pre-service teachers were first informed about the important characteristics of the impairment – its genesis and distribution and the resultant experience and behavior of the affected adolescents. Subsequently, they demonstrated in a sport unit with their peers a pedagogical inclusive handling of the impairment (without simulating the addressed impairment). The aim of this procedure was to give mastery experiences to the pre-service teachers who were responsible for the lesson. In the subsequent reflection phase, a discussion was held about which teaching adaptations specific to the impairment were perceived by the participating pre-service teachers and how they were experienced and evaluated by them. In addition, pre-service teachers disclosed their ideas and presented them for discussion. Finally, the demonstrated sport units were classified according to the theoretical models of inclusive physical education. The aim of this step was to initiate social persuasion processes and to give other pre-service teachers, who were not responsible for the lesson, vicarious experiences.

The control group also discussed methods for the individual promotion of children's competencies, but not in detail, and focused on the steps for planning and analyzing PE lessons in general.

Statistical analyses

In order to determine the effects of the intervention, we selected analysis of variance with repeated measures as the statistical test method (King, 2010). Before testing, we ran *t*-tests and ANOVAs to check for mean differences in the dependent variables "attitudes" (including all subscales used), "stress perception" and "self-efficacy" in the three groups (IG 1, IG 2, CG) at pre-measurement. As already mentioned, teachers' attitudes toward inclusion depend on different contextual variables, like gender, age, length of professional experience, amount and quality of teachers' acquaintance with persons with disabilities, amount and quality of teachers' academic training and practicum, the degree of perceived self-efficacy or competence, the impact of children's disability attributes and the school environment (Hutzler et al., 2019). Thus, we also checked for mean differences in the dependent variables due to gender and school form, as well as for former experiences with inclusion at the theoretical level and in practice at school and outside school. These procedures garnered information about possible covariates in the subsequent ANOVAs with two measurement points (pre- and post-) for the dependent variables "attitudes" (including all subscales used), "stress perception" and "self-efficacy" within the three groups. We also included a test for equality of variances (Levene's test) of the dependent variables. If this condition was violated, additional non-parametric procedures were carried out to test the development over time (Friedmann test) and the differences between the groups (Kruskal-Wallis test). This concerned the variables "willingness to teach inclusively," "stress perception" and "self-efficacy." Since all non-parametric analyses confirmed the results of the variance analyses, only the latter will be presented.

In all ANOVAs, Bonferroni was selected for the confidence interval adjustment to hold the error rate for multiple comparisons to $\alpha = 0.05$, and the contrast *repeated* was chosen. Effect sizes, such as partial eta squared and Cohen's *d*, were interpreted following Cohen (1988). All analyses were calculated with IBM SPSS Statistics Version 25 (RRID:SCR_016479).

Results

Descriptive statistics

Table 2 shows that at pre-measurement, significant differences in all the subscales of attitudes already existed within the three groups: "emotion" [$F(2) = 33.59, p < 0.001; \eta^2 = 0.51$]; "behavioral intentions" [$F(2) = 23.19, p < 0.001; \eta^2 = 0.42$]; "cognition" [$F(2) = 6.79, p = 0.002; \eta^2 = 0.18$]; "promotion of academic competencies" [$F(2) = 9.56, p < 0.001; \eta^2 = 0.24$]; "willingness to teach inclusively" [$F(2) = 10.13, p < 0.001; \eta^2 = 0.25$]; "social inclusion" [$F(2) = 39.38, p < 0.001; \eta^2 = 0.56$]. This was true as well as in the level of self-efficacy [$F(2) = 4.16, p = 0.020; \eta^2 = 0.12$]. Specifically, IG 2 significantly differed from IG 1 and CG in all variables except "self-efficacy," whereas IG 2 differed only from CG. Furthermore, males and females differed in the dependent variable "behavioral intentions" [$t(62) = 2.05, p = 0.044$; Cohen's $d = 0.53$]. Those with previous theoretical experiences with the topic of inclusion differed from those with no previous theoretical experiences with the dependent variables "emotion" [$t(65) = 2.27, p = 0.027$; Cohen's $d = 0.56$], "behavioral intentions" [$t(65) = 3.06, p = 0.003$; Cohen's $d = 0.76$], "promotion of academic competencies" [$t(62) = 2.55, p = 0.013$; Cohen's $d = 0.65$] and "social inclusion" [$t(64) = 2.10, p = 0.040$; Cohen's $d = 0.53$]. Those with practical experiences with inclusion at school differed from those with no practical experiences at school in the dependent variable "cognition" [$t(20) = 2.64, p = 0.016$; Cohen's $d = 1.18$]. Pre-service teachers with practical experiences with inclusion outside school differed from pre-service teachers with no practical experiences with inclusion outside school in the dependent variable "stress perception" [$t(60) = -2.34, p = 0.022$; Cohen's $d = 0.61$].

TABLE 2 Descriptive statistics of the dependent variables at pre-measurement.

Group	N	Mean "emotion" (SD)	Mean "behavioral intentions" (SD)	Mean "cognition" (SD)	Mean "promotion of academic competencies" (SD)	Mean "willingness to teach inclusively" (SD)	Mean "social inclusion" (SD)	Mean "stress perception" (SD)	Mean "self efficacy" (SD)
IG1	35	2.03 (0.37)	2.06 (0.54)	2.01 (0.40)	3.41 (0.54)	3.28 (0.42)	3.64 (0.52)	−0.07 (0.46)	3.14 (0.37)
IG 2	17	3.16 (0.61)	3.05 (0.60)	2.48 (0.72)	4.26 (0.88)	3.91 (0.74)	5.01 (0.81)	−0.0.51 (1.09)	3.25 (0.26)
CG	15	2.25 (0.52)	1.99 (0.41)	1.88 (0.37)	3.41 (0.74)	3.15 (0.49)	3.48 (0.31)	−0.07 (0.46)	2.91 (0.32)
Males	35	2.30 (0.52)	2.12 (0.71)	2.07 (0.55)	3.56 (0.74)	3.30 (0.60)	3.82 (0.68)	−0.12 (0.70)	3.13 (0.37)
Females	29	2.40 (0.72)	2.47 (0.64)	2.14 (0.52)	3.67 (0.83)	3.53 (0.60)	4.09 (0.94)	−0.43 (0.78)	3.13 (0.34)
Theory yes	36	2.53 (0.77)	2.52 (0.72)	2.10 (0.64)	3.87 (0.84)	3.45 (0.68)	4.15 (0.98)	−0.42 (0.86)	3.16 (0.37)
Theory no	33	2.17 (0.47)	2.03 (0.55)	2.11 (0.41)	3.39 (0.62)	3.33 (0.49)	3.73 (0.58)	−0.14 (0.55)	3.06 (0.33)
Practical school yes	19	2.32 (0.73)	2.04 (0.74)	2.27 (0.32)	3.49 (0.39)	3.45 (0.39)	4.03 (0.87)	0.21 (0.63)	3.06 (0.32)
Practical school no	3	2.67 (0.46)	2.13 (0.50)	1.73 (0.38)	3.50 (0.17)	3.00 (0.20)	3.67 (0.14)	−0.54 (0.81)	3.30 (0.44)
Practical outside school yes	29	2.37 (0.69)	2.34 (0.65)	2.00 (0.62)	3.66 (0.78)	3.50 (0.74)	4.05 (0.86)	−0.52 (0.80)	3.15 (0.35)
Practical outside school no	38	2.36 (0.66)	2.26 (0.72)	2.18 (0.47)	3.62 (0.79)	3.33 (0.46)	3.89 (0.84)	−0.10 (0.64)	3.09 (0.36)

IG 1, intervention group 1; IG 2, intervention group 2, CG, control group; N, sample size; SD, standard deviation. Differences in the number of participants are due to dropout or missing answers in the concerned variables.

Impact of the intervention

As there were significant mean differences in the dependent variables at pre-measurement, we decided to calculate repeated ANCOVA measures with the pre-test score as the covariate (Jamieson, 2004). Thus, we set gender and previous theoretical or practical experience with inclusion at school or outside school as covariates if there were mean differences in the dependent variables at pre-measurement within the different groups.

Concerning attitudes toward inclusion, we observed for "emotion," "behavioral intentions" and "social inclusion" neither a significant interaction nor a significant main effect from pre- to post-measurement. However, for "cognition," "promotion of academic competencies" and "willingness to teach inclusively," we identified significant effects that will be outlined in the following. For "cognition," we observed a significant time effect from pre- to post-measurement [F (1.29) = 22.03, $p < 0.001$; $\eta^2 = 0.43$] as well as a significant group-by-dependent variable interaction effect, mediated by the covariate [F (1.29) = 6.09, $p = 0.002$; $\eta^2 = 0.39$]. *Post hoc* analysis revealed that the attitude "cognition" increased significantly in all three groups [IG 1 $mean_{post} = 2.27$, $SD_{post} = 0.35$; t (19) = −3.24, $p = 0.004$; Cohen's $d = 0.78$; IG 2 $mean_{post} = 3.18$, $SD_{post} = 0.54$; t (10) = −5.94, $p < 0.001$; Cohen's $d = 0.98$; CG $mean_{post} = 2.25$, $SD_{post} = 0.41$; t (3) = −4.98, $p = 0.016$; Cohen's $d = 0.87$]. However, by controlling for the mean differences at pre-measurement, the increase in IG 2 was significantly higher than in IG 1 [t (29) = −1.95, $p = 0.047$; Cohen's $d = 0.76$]. However, there was no significant difference in the increase to CG or between IG 1 and CG. For "promotion of academic competencies," we found a significant group-by-time interaction effect [F (1.29) = 3.43, $p = 0.049$; $\eta^2 = 0.22$]. *Post hoc* analysis showed that the attitude "promotion of academic competencies" increased significantly in IG 2 from pre- to post-measurement ($mean_{post} = 4.46$, $SD_{post} = 0.77$) in comparison to IG 1 ($mean_{post} = 3.13$, $SD_{post} = 0.43$) and CG ($mean_{post} = 3.08$, $SD_{post} = 0.11$), for which no significant changes could be observed [F (2) = 8.63, $p = 0.001$; $\eta^2 = 0.37$]. For "willingness to teach inclusively," we noted a significant time effect from pre- to post-measurement [F (1.29) = 7.03, $p = 0.013$; $\eta^2 = 0.20$] as well as a significant covariate-by-dependent variable interaction effect [F (1.29) = 6.24, $p = 0.018$; $\eta^2 = 0.18$]. Post analysis revealed that only in IG 2 did the attitudes "willingness to teach inclusively" increase significantly from pre- to post-measurement [t (9) = −3.55, $p = 0.006$; Cohen's $d = 0.79$], and that the increase in IG 2 was significantly higher than in IG 1 [F (2) = 7.79, $p = 0.002$; $\eta^2 = 0.33$].

Regarding stress perception and self-efficacy, there were no significant interactions or main effects from pre- to post-measurement.

Discussion

Preparing pre-service teachers for teaching inclusively seems to be one of the most important challenges of teacher education. With the ratification of the United Nations' Convention on the Rights of Persons with Disabilities, inclusion has become not only a societal desire but a personal right. The increasing number of children with special needs who are taught in a "regular" school context has led to new challenges for PE teachers in not always ideal structural conditions. As teacher competencies seem to be a key factor in teaching successfully in inclusive settings, teacher education programs must focus on the promotion of these domain-specific competencies, which can be seen as new or additional skills, behaviors and beliefs (Darling-Hammond and Bransford, 2005). As PE differs in several aspects from other school subjects, a particular focus on (pre-service) PE teachers is needed.

In recent years, several intervention studies on (pre-service) PE teachers have been conducted to enhance their attitudes and self-efficacy toward teaching inclusively. The results underline the importance of the intervention length as well as the relevance of the combination of theoretical inputs with practical experiences that should be pedagogically accompanied and reflected upon. Similar to studies of general education, some of the previous studies with (pre-service) PE teachers were limited in their results due to methodological problems (e.g., no implementation of a control group). Furthermore, it was still unclear which kind of practical experience in combination with theoretical inputs might be (more) successful in improving attitudes, self-efficacy, and stress perception of inclusive teaching in pre-service PE teacher education. Thus, we conducted a quasi-experimental intervention study with pre- and post-measurement, consisting of fifteen 90-min teaching sessions, and comparing different intervention programs (IG 1, IG 2) in relation to a control group. We hypothesized that the two different intervention groups (IG 1, IG 2) would develop enhanced attitudes toward inclusion and self-efficacy regarding teaching inclusively as well as decreased perceived stress related to teaching PE inclusively compared to the control group. Furthermore, we hypothesized that the development in IG 2, where the PE pre-service teachers underwent a combination of theoretical input and practical application, would be greater than in IG 1, where the PE pre-service teachers followed an exclusively theoretical information-based seminar.

Based on the ANCOVA analysis, we can state significant results for some subscales of the attitudes. However, there were no significant results for stress perception and self-efficacy.

The attitude "cognition" increased significantly in all three groups from pre- to post-measurement. By controlling for the mean differences at pre-measurement, the increase in IG 2 was significantly higher than in IG 1. Thus, the teaching units of the intervention groups and of the control group seem to have led to an enhancement of the cognitive aspect of attitudes toward inclusion, meaning that an inclusive school system can be implemented without any problems and that it would enhance social justice. In particular, PE pre-service teachers of the IG 2 fostered this cognitive aspect after the intervention. Whereas the information-based approach was also successful in promoting the attitude that teaching inclusively is possible, the theoretical approach in combination with practical experience had even more impact. The attitude "promotion of academic competencies" increased significantly from pre to post in IG 2, whereas there was no significant change in IG 1 and CG. Thus, PE pre-service teachers from IG 2 experienced an increase in the belief that teachers can support equally both children with and without special needs in an inclusive context. It appears that the practical experience in which pre-service teachers of IG 2 planned PE lessons according to students' special educational needs (e.g., learning disabilities, emotional and behavioral difficulties, physical impairments, chronic diseases and intellectual impairments) increased PE pre-service teachers' positive attitudes toward dealing with different levels of proficiency, whereas the particular and non-particular information-based approach to teaching in an inclusive context had no impact.

The attitude "willingness to teach inclusively" showed a significant increase only in IG 2. This attitude can be seen as a behavioral-affective component and is content-related with respect to self-efficacy or control expectancy. PE pre-service teachers from IG 2 felt significantly more competent to teach inclusively after the intervention and were less afraid of teaching inclusively than PE pre-service teachers from IG 1 or from CG. Although, they did not teach "real" students with disabilities, the combination of theory and practical teaching experiences with fellow pre-service teachers had more impact on their attitudes than the purely information-based approach in IG 1 or the non-specific theoretical approach in the CG. This might be also a limitation of the present study, as teaching PE lessons with fellow pre-service teachers is an artificial context. In contrast to Reina et al. (2019), we decided not to work with a simulation procedure to prevent the risk of triggering stereotypes (Macrae et al., 1994), and to prevent negative effects of overwhelming and missing coping strategies in simulation procedures (Flower et al., 2007). However, we need future study courses in which PE pre-service teachers are theoretically prepared to teach an inclusive PE class and put into practice these theoretical units with an inclusive school class. This field experience should be theoretically reflected upon afterward. However, getting into contact with the target group seems not always the best and most effective way to promote teachers' competencies and their underlying cognitive-affective constructs, such as attitudes or self-efficacy, as it has been shown by Anttila et al. (2018) that putting into practice theoretically planned PE lessons with a group of asylum seekers does not enhance *per se* PE pre-service teachers' intercultural competencies and attitudes; it might even have a counteractive effect by emphasizing patriarchal feelings of superiority. Thus, we should develop teaching formats that combine theory-driven seminar units with practical experiences in teaching an inclusive PE class, which should be implemented carefully. In this context, reflection upon the experiences and especially feelings during the practical teaching situations seems to be crucial. To initiate the reflection process, video sequences from the video-recorded lessons conducted by the pre-service teachers could be used as stimulators. Therefore, it might be important for the pre-service teachers to decide for themselves which video sequences to use for the reflection (Hinternesch et al., 2021).

Regarding the study design, a mixed-methods study combining quantitative and qualitative data seems to be crucial to understand theoretically expected or unexpected changing processes. In this regard, our study is limited with respect to quantitative data, and we do not know, for example, if and what the PE pre-service teachers experienced with respect to the intervention programs as sources to enhance their self-efficacy, as Martins et al. (2015) determined. We also did not operationalize the impact of the pre-service teachers' attitudes and self-efficacy on concrete teaching behavior outside the study context. Finally, the impact of these and other context variables (e.g., school environment) is one of the most frequently lacking or ambivalent links in research on teaching (PE) inclusively (Hutzler et al., 2019), and it needs to be tackled urgently in future studies on (PE pre-service) teachers.

Furthermore, the reflections upon the teaching experiences should also focus on the role of power relations that are anchored in the school context and have an impact on how (PE) teachers perceive their role and privileges. This aspect is underlined in seminar formats using the critical pedagogy approach, which should also be carefully implemented so as not to overwhelm pre-service teachers (see, e.g., Shelley and McCuaig, 2018). Following Barber (2018), it seems also to be important to reflect upon traditional understandings of (dis-)ability, performance, body norms, and values that interfere with a traditional understanding of elite sport. In this context, it should not be neglected that most pre-service PE teachers enter PE teacher education programs with particular (positive) socialization experiences in the (elite) sport context, and they try to remain connected to the field of sport via their occupation (O'Neil and Richards, 2018). Many PE teacher candidates were successful athletes, who had experienced physical education curricula dominated by traditional teaching strategies and contents, and they have already developed strong subjective theories about traditional sport content in PE (Richards et al., 2014). Furthermore, Morgan and Hansen (2008) found out that PE teachers' teaching strategies are linked to personal school experiences in PE. These socialization effects are reinforced by a recruitment strategy for future PE teachers that focuses on traditional sport competencies and selects those who mirror and share the traditional values and norms (Richards et al., 2020). In combination with a teaching curriculum that is also traditionally orientated toward sport, pre-service PE teachers are not challenged in their experiences and attitudes. The PE teacher candidates perceive their study courses, and later PE, just as a continuation of their sport identity (Curtner-Smith, 2017). Therefore, there is a need to reflect upon the recruitment strategies for PE teacher education and to develop a procedure to attracts PE teacher candidates representing diversity in sport and movement experiences (O'Neil and Richards, 2018) as well as in social backgrounds like gender, ethnicity, race, and socioeconomic status (Flintoff and Webb, 2012). Furthermore, we need to rethink PE teacher curricula and look for teaching strategies that challenge PE teacher candidates' experiences and attitudes. This also includes the need for critical biography work on teacher educators' socialization experiences, norms, and values, as they might be seen as crucial role models for a future inclusive PE setting (Flintoff et al., 2015).

Conclusion

In conclusion, by taking into account the small sample size of our study as limitation, the combination of a theoretical approach with practical experiences in teaching PE units with peer pre-service PE teachers seems to have more impact on the development of positive attitudes toward inclusion than only theoretically based approaches. Thus, our study underlines the importance of mastery experiences in settings that might be similar like the "real" professional context but can be still considered as a kind of "test environment." The complexity might be reduced, and pre-service teachers might not feel overwhelmed due to probably (still) missing teaching competencies. Thus, the reflection upon the practical experiences and especially feelings during the practical teaching situations seems to be crucial. Finally, we need to develop teaching formats that are on the one hand adapted to PE pre-service teachers' competencies, but challenge on the other hand the professional development by systematically implementing opportunities for critical (self-)reflection.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by TU Dortmund University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

EG-S conceptualized the design, was responsible for data management and analysis (parametric), and wrote the original draft including visualizing tables. MS implemented the study design at her university, developed and described the content for intervention group, ran non-parametric analysis, and commented on the original draft. Both authors contributed to the article and approved the submitted version.

Funding

This work was part of the project Dortmund Profile for Inclusion-Oriented Learning and Teacher Training (DoProfiL). DoProfiL was part of the "Qualitätsoffensive Lehrerbildung," joint initiative of the Federal Government and the Länder, which aims to improve the quality of teacher training. The program was funded by the Federal Ministry of Education and Research (01JA1630).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Anttila, E., Siljamäki, M., and Rowe, N. (2018). Teachers as frontline agents of integration: Finnish physical education students' reflections on intercultural encounters. *Phys. Educ. Sport Peda.* 23, 609–622. doi: 10.1080/17408989.2018.1485141

Avramidis, E., and Norwich, B. (2002). Teachers' attitudes towards integration/inclusion: A review of the literature. *Eur. J. Spec. Needs Educ.* 17, 129–147. doi: 10.1080/08856250210129056

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioural change. *Psychol. Rev.* 84, 191–215. doi: 10.1037/0033-295X.84.2.191

Barber, W. (2018). Inclusive and accessible physical education: Rethinking ability and disability in pre-service teacher education. *Sport Educ. Soc.* 23, 520–532. doi: 10.1080/13573322.2016.1269004

Block, M. E., Hutzler, Y., Barak, S., and Klavina, A. (2013). Creation and validation of the self-efficacy instrument for physical education teacher education majors toward inclusion. *Adapt. Phys. Activ. Q.* 30, 184–205. doi: 10.1123/apaq.30.2.184

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (2015). Beyond dichotomies. Competence viewed as a continuum. *Z. Psychol.* 22, 3–13. doi: 10.1027/2151-2604/a000194

Brouwers, A., Tomic, W., and Boluijt, H. (2011). Job demands, job control, social support and self-efficacy beliefs as determinants of burnout among physical education teachers. *Eur. J. Psychol.* 7, 17–39. doi: 10.5964/ejop.v7i1.103

CAST (2011). *The UDL Guidelines.* Available Online at: https://udlguidelines.cast.org/more/downloads (accessed Mar 30, 2022).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Curtner-Smith, M. (2017). "Acculturation, recruitment, and the development of orientations," in *Teacher Socialization In Physical Education: New Perspectives*, eds K. A. R. Richards and K. L. Gaudreault (New York, NY: Routledge), 33–46.

Darling-Hammond, L., and Bransford, J. D. (2005). *Preparing Teachers For A Changing World. What Teachers Should Learn And Be Able To Do.* San Francisco, CA: Jossey-Bass.

De Boer, A., Pijl, S. J., and Minnaert, A. (2011). Regular primary schoolteachers' attitudes towards inclusive education: A review of the literature. *Int. J. Incl. Educ.* 15, 331–353. doi: 10.1080/13603110903030089

Erhorn, J., Langer, W., and Möller, L. (2020). Vorbereitung angehender Sportlehrkräfte auf einen inklusiven Sportunterricht' Eine kritische Bestandsaufnahme hochschuldidaktischer Lehrformate. [Preparing future physical education teachers for inclusive physical education' A critical review of university teaching formats]. *Ger. J. Exerc. Sport Res.* 50, 1–14. doi: 10.1007/s12662-020-00668-5

Fediuk, F. (2008). *Sport In Heterogenen Gruppen. Integrative Prozesse In Sportgruppen Mit Behinderten Und Benachteiligten Menschen. [Sport In Heterogeneous Groups. Integrative Processes In Sports Groups With Disabled And Disadvantaged People.* Aachen: Meyer & Meyer.

Fejgin, N., Talmor, R., and Erlich, I. (2005). Inclusion and burnout in physical education. *Eur. Phys. Educ. Rev.* 11, 29–50. doi: 10.1177/1356336X05049823

Flintoff, A., and Webb, L. (2012). 'Just open your eyes a bit more': The methodological challenges of researching black and minority ethnic students' experiences of physical education teacher education. *Sport Educ. Soc.* 17, 571–589. doi: 10.1080/13573322.2011.553951

Flintoff, A., Dowling, F., and Fitzgerald, K. (2015). Working through whiteness, race and (anti) racism in physical education teacher education. *Phys. Educ. Sport Peda.* 20, 559–570. doi: 10.1080/17408989.2014.962017

Flower, A., Burns, M. K., and Bottsford-Miller, N. A. (2007). Meta-analysis of disability simulation research. *Remedial Spec. Educ.* 28, 72–79.

Frankel, E. B., Gold, S., and Ajodhia-Andrews, A. (2010). International preschool inclusion: Bridging the gap between vision and practice. *Young Except. Child.* 13, 2–16. doi: 10.1177/1096250610379983

Gaab, J. (2009). PASA – Primary Appraisal Secondary Appraisal - Ein Fragebogen zur Erfassung von situations-bezogenen kognitiven Bewertungen. [A questionnaire to record situation-related cognitive assessments]. *Verhaltenstherapie* 19, 114–115.

Giese, M., and Ruin, S. (2018). "Forgotten bodies - an examination of physical education from the perspective of ableism," in *Sport and Disability. From Integration Continuum To Inclusion Spectrum*, ed. F. Kiuppis (London: Routledge), 152–165.

Giese, M., and Weigelt, L. (2017). *Inklusiver Sportunterricht in Theorie und Praxis. [Inclusive Physical Education In Theory And Practice].* Aachen: Meyer & Meyer.

Gieß-Stüber, P. (2012). "Geschlechterforschung und Sportdidaktik" [Gender Studies and Sport Didactics], in *Handbuch Geschlechterforschung und Fachdidaktik [Handbook of Gender Studies and Subject Didactics]*, eds M. Kampshoff and C. Wiepcke (Wiesbaden: Springer), 273–286.

Gilor, O., and Katz, M. (2017). Teaching inclusive classes: What pre-service teachers in Israel think about their training. *J. Cogn. Educ. Psychol.* 16, 293–303. doi: 10.1891/JCEP.16.3

Grimminger, E. (2013). Sport Motor competencies and the experience of social recognition among peers in physical education – a video-based study. *Phys. Educ. Sport Peda.* 18, 506–519. doi: 10.1080/17408989.2012.690387

Grimminger, E. (2014). Getting into teams or pairs in physical education and exclusion processes among students – a mixed-method study. *Pedagog. Int. J.* 9, 155–171. doi: 10.1080/1554480X.2014.899546

Grimminger-Seidensticker, E., and Möhwald, A. (2017). Intercultural education in physical education: Results of a quasi-experimental intervention study with secondary school students. *Phys. Educ. Sport Peda.* 22, 445–458. doi: 10.1080/17408989.2016.1225030

Grube, D., Ryan, S., Lowell, S., and Stringer, A. (2018). Effective classroom management in physical education: Strategies for beginning teachers. *J. Phys. Educ. Recreat. Dance* 89, 47–52. doi: 10.1080/07303084.2018.1503117

Haegele, J. A., Hodge, S., Filho, P. J. B. G., and de Rezende, A. L. G. (2018). Brazilian physical education teachers' attitudes toward inclusion before and after participation in a professional development workshop. *Eur. Phys. Educ. Rev.* 24, 21–38. doi: 10.1177/1356336X16662898

Hall, T. E., Meyer, A., and Rose, D. H. (2012). *Universal Design For Learning In The Classroom*. New York, NY: The Guilford Press.

Harant, M. (2017). "Inklusion im Widerstreit." [Inclusion in a Conflict]," in *Bildung und Teilhabe [Education and Participation]*, eds I. Miethe, A. Tervooren, and N. Ricken (Wiesbaden: Springer), 111–130.

Hinternesch, S., Schwier, J., and Seyda, M. (2021). Digitalisierung meets internationalisierung. blended-learning im praxissemester [digitalisation meets internationalization. blended-learning during the practical semester]. *Z. Stud. und Lehre in der Sportwissenschaft* 4, 19–25.

Hopkins, S., Round, P., and Barley, K. (2018). Preparing beginning teachers for inclusion: Designing and assessing supplementary fieldwork experiences. *Teach. Teach. Theory Pract.* 24, 915–930. doi: 10.1080/13540602.2018.1495624

Hutzler, Y., Meier, S., Reuker, S., and Zitomer, M. (2019). Attitudes and self-efficacy of physical education teachers toward inclusion of children with disabilities: A narrative review of international literature. *Phys. Educ. Sport Peda.* 24, 249–266. doi: 10.1080/17408989.2019.1571183

Hutzler, Y., Zach, S., and Gafni, O. (2005). Physical education students' attitudes and self-efficacy towards the participation of children with special needs in regular classes. *Eur. J. Spec. Needs Educ.* 20, 309–327. doi: 10.1080/08856250500156038

Jamieson, J. (2004). Analysis of covariance (ANCOVA) with difference scores. *Int. J. Psychophysiol.* 52, 277–283. doi: 10.1016/j.ijpsycho.2003.12.009

Kastrup, V., and Kleindienst-Cachay, C. (2016). Reflective co-education' or male-oriented physical education' Teachers' views about activities in co-educational PE classes at German secondary schools. *Sport Educ. Soc.* 21, 963–984. doi: 10.1080/13573322.2014.984673

Kerner, C., Haerens, L., and Kirk, D. (2018). Understanding body image in physical education: Current knowledge and future directions. *Eur. Phys. Educ. Rev.* 24, 255–265. doi: 10.1177/1356336X17692508

Killoran, I., Woronko, D., and Zaretsky, H. (2014). Exploring pre-service teachers' attitudes towards inclusion. *Int. J. Incl. Educ.* 18, 427–442. doi: 10.1080/13603116.2013.784367

King, B. (2010). "Analysis of variance," in *International Encyclopedia of Education*, 3rd Edn, eds P. Peterson, E. Baker, and B. McGaw (Oxford: Elsevier), 32–36.

Ko, B., and Boswell, B. (2013). Teachers' perceptions, teaching practices, and learning opportunities for inclusion. *Phys. Educ.* 70, 223–242.

Kullmann, H., Lütje-Klose, B., and Textor, A. (2014). "Eine Allgemeine Didaktik für inklusive Lerngruppen – fünf Leitprinzipien als Grundlage eines Bielefelder Ansatzes der inklusiven Didaktik" [A general didactics for inclusive learning groups – five principles as foundation of an approach developed in Bielefeld]," in *Fachdidaktik inklusiv – Auf der Suche nach didaktischen Leitlinien für den Umgang mit Vielfalt in der Schule [Didactics of school subjects – In search of didactical principles for handling diversity at school]*, eds B. Amrhein and M. Dziak-Mahler (Münster: Waxmann), 89–107.

Lautenbach, F., and Heyder, A. (2019). Changing attitudes to inclusion in pre-service teacher education: A systematic review. *Educ. Res.* 61, 231–253. doi: 10.1080/00131881.2019.1596035

Lautenbach, F., Korte, J., Möhwald, A., Heyder, A., and Grimminger-Seidensticker, E. (2020). A 14-week intervention study on changing pre-service teachers' psychological perspectives on inclusion: Explicit and implicit attitudes, self-efficacy, and stress perception toward inclusion. *Front. Educ.* 5:7. doi: 10.3389/feduc.2020.00007

Lazarus, R. S., and Folkman, S. (1984). *Stress, Appraisal, And Coping.* New York, NY: Springer.

Leineweber, H., and Thomas, M. (2017). Heterogene Lerngruppen als Herausforderung für Sportlehrkräfte' [Heterogenous learning groups as a challenge for physical education teachers']. *Impulse* 23, 28–33.

Lüke, T., and Grosche, M. (2018b). What do I think about inclusive education' It depends on who is asking. Experimental evidence for a social desirability bias in attitudes towards inclusion. *Int. J. Incl. Educ.* 22, 38–53. doi: 10.1080/13603116.2017.1348548

Lüke, T., and Grosche, M. (2018a). Konstruktion und Validierung der Professionsunabha'ngigen Einstellungsskala zum Inklusiven Schulsystem (PREIS). [Design and validation of the profession-independent attitude scale for the inclusive school system]. *Empirische Sonderpädagogik.* 10, 3–20.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., and Jetten, J. (1994). Out of mind but back in sight: Stereotypes on the rebound. *J. Pers. Soc. Psychol.* 67, 808–817. doi: 10.1037/0022-3514.67.5.808

Martins, M., Costa, J., and Onofre, M. (2015). Practicum experiences as sources of pre-service teachers' self-efficacy. *Eur. J. Teach. Educ.* 38, 263–279. doi: 10.1080/02619768.2014.968705

McGowan, J. P. (1999). *The Effects Of Disability Simulations On Attitudes Toward Persons With Disabilities*, Ph.D. thesis. South Orange, NJ: Seton Hall University.

McNeish, D. (2018). Thanks Coefficient Alpha, we'll take it from here. *Psychol. Methods* 23, 412–433. doi: 10.1037/met0000144

Morgan, P., and Hansen, V. (2008). The relationship between PE biographies and PE teaching practices of classroom teachers. *Sport Educ. Soc.* 13, 373–391. doi: 10.1080/13573320802444994

Morley, D., Bailey, R., and Tan, J. (2005). Inclusive physical education: Teachers' views of including pupils with special educational needs and/or disabilities in physical education. *Eur. Phys. Educ. Rev.* 11, 84–107. doi: 10.1177/1356336X05049826

O'Neil, K., and Richards, K. A. R. (2018). Breaking from Traditionalism: Strategies for the recruitment of physical education teachers. *J. Phys. Educ. Recreat. Dance* 89, 34–41. doi: 10.1080/07303084.2017.1404511

O'Brien, D., Kudláček, M., and Howe, P. (2009). A contemporary review of english language literature on inclusion of students with disabilities in physical education: A european perspective. *Eur. J. Adapt. Phys. Act.* 2, 46–61. doi: 10.5507/euj.2009.004

Overton, H., Wrench, A., and Garrett, R. (2017). Pedagogies for inclusion of junior primary students with disabilities in PE. *Phys. Educ. Sport Peda.* 22, 414–426. doi: 10.1080/17408989.2016.1176134

Qi, J., and Ha, A. S. (2012). Inclusion in physical education: A review of literature. *Int. J. f Disabil. Dev. Educ.* 59, 257–281. doi: 10.1080/1034912X.2012.697737

Reina, R., Haegele, J. A., Pérez-Torralba, A., Carbonell-Hernández, L., and Roldan, A. (2021). The influence of a teacher designed and -implemented disability awareness programme on the attitudes of students towards inclusion. *Eur. Phys. Educ. Rev.* 27, 837–853. doi: 10.1177/1356336X21999400

Reina, R., Healy, S., Roldán, A., Hemmelmayr, I., and Klavina, A. (2019). *Incluye-T*: A professional development program to increase the self-efficacy of physical educators towards inclusion. *Phys. Educ. Sport Peda.* 24, 319–331. doi: 10.1080/17408989.2019.1576863

Reuker, S., Rischke, A., Kämpfe, A., Schmitz, B., Teubert, H., Thissen, A., et al. (2016). Inklusion im Sportunterricht [Inclusion in physical education]. *Sportwissenschaft* 46, 88–101.

Richards, K. A. R., Iannucci, C., McEvoy, E., and Simonton, A. (2020). "The professional socialisation challenge: Teacher education for a preferable future for physical education," in *School Physical Education And Teacher Education: Collaborative Redesign For The 21st Century*, eds A. MacPhail and H. A. Lawson (Abingdon: Routledge), 70–81. doi: 10.1186/s12913-016-1423-5

Richards, K. A. R., Templin, T. J., and Graber, K. (2014). The socialization of teachers in physical education: Review and recommendations for future works. *Kinesiol. Rev.* 3, 113–134. doi: 10.1123/kr.2013-0006

Savolainen, H., Malinen, O.-P., and Schwab, S. (2020). Teacher efficacy predicts teachers' attitudes towards inclusion – a longitudinal cross-lagged analysis. *Int. J. Incl. Educ.* [Epub ahead of print]. doi: 10.1080/13603116.2020.1752826

Schwarzer, R., and Jerusalem, M. (1999). *Skalen Zur Erfassung Von Lehrer-Und Schülermerkmalen [Scales For Recording Teacher And Student Characteristics]*. Available Online at: http://www.psyc.de/skalendoku.pdf. (accessed Mar 30, 2022).

Seifried, S., and Heyl, V. (2016). "Konstruktion und Validierung eines Einstellungsfragebogens zu Inklusion für Lehrkräfte (EFI-L)". [Design and validation of a questionnaire on attitudes towards inclusion for teachers]. *Empirische Sonderpädagogik* 8, 22–35.

Sharma, U., Forlin, C., and Loreman, T. (2008). Impact of training on pre-service teachers' attitudes and concerns about inclusive education and sentiments about persons with disabilities. *Disabil. Soc.* 23, 773–785. doi: 10.1080/09687590802469271

Sharma, U., Forlin, C., Loreman, T., and Earle, C. (2006). Pre-service teachers' attitudes, concerns and sentiments about inclusive education: An international comparison of the novice pre-service teachers. *Int. J. Spec. Educ.* 21, 80–93.

Shelley, K., and McCuaig, L. (2018). Close encounters with critical pedagogy in socio-critically informed health education teacher education. *Phys. Educ. Sport Peda.* 23, 510–523. doi: 10.1080/17408989.2018.1470615

Specht, J. A., and Metsala, J. L. (2018). Predictors of teacher efficacy for inclusive practice in pre-service teachers. *Except. Educ. Int.* 28, 67–82. doi: 10.5206/eei.v28i3.7772

Taliaferro, A. R., Hammond, L., and Wyant, K. (2015). Preservice physical educators' self-efficacy beliefs toward inclusion: The impact of coursework and practicum. *Adapt. Phys. Act. Q.* 32, 49–67. doi: 10.1123/apaq.2013-0112

Taliaferro, A., and Harris, N. P. (2014). The Effects of a one-day workshop on physical educators' self-efficacy towards inclusion of students with autism. *Palaestra* 28, 38–43.

Tant, M., and Watelain, E. (2016). Forty years later, a systematic literature review on inclusion in physical education (1975–2015): A teacher perspective. *Educ. Res. Rev.* 19, 1–17. doi: 10.1016/j.edurev.2016.04.002

Tiemann, H. (2015). "Inklusiven Sportunterricht gestalten – didaktisch-methodische Überlegungen" [Designing inclusive physical education – didactical-methodical considerations]," in *Inklusiver Sportunterricht in Theorie und Praxis [Inclusive physical education in theory and practice]*, eds M. Giese and L. Weigelt (Aachen: Meyer &Meyer), 53–66.

Tiemann, H. (2016). Konzepte, Modelle und Strategien für den inklusiven Sportunterricht – internationale und nationale Entwicklungen und Zusammenhänge [Concepts, models and strategies for inclusive physical education – international and national processes and contexts]. *Z. für Inklusion* 10:3.

Tschannen-Moran, M., Woolfolk Hoy, A., and Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Rev. Educ. Res.* 68, 202–248. doi: 10.3102/00346543068002202

United Nations [UN] (2014). *Convention On The Rights Of Persons With Disabilities*. Available Online at: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html. (accessed Mar 30, 2022).

Wilson, C., Woolfson, L. M., and Durkin, K. (2020). School environment and mastery experience as predictors of teachers' self-efficacy beliefs towards inclusive teaching. *Int. J. Incl. Educ.* 24, 218–234. doi: 10.1080/13603116.2018.1455901

Yeo, L. S., Chong, W. H., Neihart, M. F., and Huan, V. S. (2014). Teachers' experience with inclusive education in Singapore. *Asia Pac. J. Educ.* 36, 69–83. doi: 10.1080/02188791.2014.934781

Zach, S., Harari, I., and Harari, N. (2012). Changes in teaching efficacy of pre-service teachers in physical education. *Phys. Educ. Sport Peda.* 17, 447–462. doi: 10.1080/17408989.2011.582491

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/feduc.2022.976346

Check for updates

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University, Germany

REVIEWED BY
Zhi Liu,
Central China Normal University, China
Colin Cramer,
University of Tübingen, Germany

*CORRESPONDENCE
Maria Zimmermann
maria.zimmermann@hu-berlin.de

SPECIALTY SECTION
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

RECEIVED 23 June 2022
ACCEPTED 23 August 2022
PUBLISHED 15 September 2022

CITATION
Zimmermann M, Engel O and Mayweg-Paus E (2022) Pre-service teachers' search strategies when sourcing educational information on the Internet.
Front. Educ. 7:976346.
doi: 10.3389/feduc.2022.976346

Pre-service teachers' search strategies when sourcing educational information on the Internet

Maria Zimmermann*, Ole Engel and Elisabeth Mayweg-Paus

Digital Knowledge Management, Department of Education Studies, Faculty of Humanities and Social Science, Humboldt-University of Berlin, Berlin, Germany

Teachers need to be able to inform and justify their teaching practice based on available research knowledge. When searching for research knowledge, the Internet plays a crucial role as it allows teachers to search for and access evidence long after their own education at university. On the Internet, however, educational information can have varying levels of scientific groundedness (e.g., science articles or blogs from colleagues), and research indicates that (pre-service) teachers struggle to find, select, and evaluate online educational information. It is precisely for this reason that it is important to educate (pre-service) teachers on how to competently source online information. This study describes pre-service teachers' search strategies when sourcing online educational information about the topic "students' use of mobile phones in class." It sheds light on their use of (1) basic or advanced search strategies and (2) the role of Internet-specific epistemological beliefs (ISEBs). N = 77 pre-service teachers conducted a realistic search on the Internet and selected those web items (WI) that they perceived relevant for justifying whether mobile phones should be used in class. Their sourcing behavior was screen-recorded and analyzed. Most selected WI were found via search engines of Google LLC (91.4%). Advanced search strategies were defined as (1) using two or more search engines (performed by 62.3% of participants), (2) adapting search terms and/or formulating new search terms (90.9%), (3) selecting at least one WI that was not listed among the first four ranks on the first search engine results page (54.7%), and (4) checking for the trustworthiness of the author/source (14.3%) or the quality of the content (13%). Binary logistic regressions were used to analyze the relationship between ISEBs and (1) search strategies and (2) science-relatedness of WI as dependent variables. The predictor ISEB did not contribute to the models, meaning that differences in participants' ISEBs did not significantly relate to their search strategies nor to the science-relatedness of WI, all $\beta \leq |0.36|$, *Wald* ≤ 0.64, $p \geq 0.43$. The role of pre-service teachers' search strategies is discussed with respect to teachers' evidence-informed reasoning and its implications for teacher education.

KEYWORDS

pre-service teachers, evidence, search strategies online, sourcing competencies online, Internet-specific epistemological beliefs

Introduction

During their professional lives, teachers are confronted with a broad range of pedagogical questions and problems, such as questions like, "Should I allow my students to use their mobile phones in classes?" "What's best for their learning, to allow mobile phones or to ban them from classes?" Teachers, like many other practitioners, need to deal with such questions and problems in an evidence-informed manner. This means that they need to justify their answers, decisions, and practices professionally by engaging in complex epistemic processes, including searching for, interpreting, and using the evidence that is "(a) most relevant to the [pedagogical] decision and (b) has the highest degree of certainty" (Spencer et al., 2012, p. 133) (Fischer et al., 2014). Accordingly, educational research and policy standards of teacher education demand pre-service teachers to ground their decisions and actions in evidence from educational research rather than base them on gut feelings (e.g., Bauer and Prenzel, 2012; Bromme et al., 2014; Häkkinen et al., 2017; Thomm et al., 2021b). In this sense, sourcing relevant evidence is a crucial part of (pre-service) teachers' evidence-informed practices.

In this endeavor, the Internet plays a crucial role as it allows pre-service and in-service teachers to search for up-to-date educational information and access evidence easily, long after they have left university (Williams and Coles, 2007; Bromme et al., 2014; Caena and Redecker, 2019). For example, during their entire professional lives, teachers have to decide which teaching methods are best suited to achieve newly defined learning goals, such as helping their students develop media skills. This was particularly evident under the specific circumstances of the COVID-19 Pandemic when teachers around the world had to teach at a distance and therefore had to adapt their regular practices to the opportunities and challenges of the new online (or hybrid) teaching methods. In addition to these society developments which require new forms of teaching and learning, scientific evidence on teaching and learning are also constantly changing (Bromme and Goldman, 2014), and teachers can use the Internet to become informed about the current educational scientific findings on a particular topic.

Pre-service and in-service teachers seem to use the Internet frequently to extract educational information related to pedagogical problems (Williams and Coles, 2007; Bougatzeli et al., 2017), but sourcing relevant evidence from the Internet can come with several hurdles: Teachers can access a variety of information easily, but this information may have varying levels of scientific groundedness (e.g., open access education science journals, open educational resources, science-related blogs, or blogs from colleagues) and may be inaccurate due to unavailable gatekeeping mechanisms (Metzger and Flanagin, 2013; Hendriks et al., 2015). Thus, (pre-service) teachers have to evaluate the relevance and quality of a considerable amount of educational information. In this context, (pre-service) teachers report frustration and worry about being unable to find accurate information or evaluate it appropriately—even if they are intrinsically motivated to explore further and connect the information to other scientific sources during online searches (synthesizing evidence constitutes an appropriate scientific practice: Rousseau and Gunia, 2016) (Chen et al., 2019; Iding et al., 2009). Furthermore, when sourcing relevant educational evidence from the Internet, (pre-service) teachers need to deal with the affordances offered by the search engines and other media they are using online (e.g., blogs or video platforms). In this sense, it is becoming increasingly important to educate pre-service and in-service teachers on the skills they need to find, select, evaluate, and use science-related online information (e.g., European Digital Competence Framework for Educators [DigCompEdu]: Caena and Redecker, 2019). Often discussions on online sourcing competencies are based on the complex and interrelated constructs of information literacy (Duke and Ward, 2009; Häkkinen et al., 2017; Caena and Redecker, 2019). So, although (pre-service) teachers should be trained to be competent at sourcing evidence from the Internet—first, to keep themselves up-to-date in the sense of informal lifelong learning and, second, to teach their students how to source online information competently (e.g., Wilson et al., 2011; Caena and Redecker, 2019)—little is known about how they actually source online information and what strategies they actually use. Thus, it is important to understand how pre-service teachers source evidence from the Internet when aiming to find a solution for a pedagogical problem. The present study aims at describing pre-service teachers' sourcing behaviors on the Internet by focusing on (1) behaviors related to basic or advanced search strategies and (2) understanding the role that pre-service teachers' epistemological beliefs about knowledge from the Internet (i.e., Internet-specific epistemological beliefs [ISEBs]) play in their searches.

Approaches to searching for information on the Internet

Several theories and approaches from diverse research fields (e.g., communication science, information science, and psychology) exist that aim to describe how individuals search for information. While some models consider information searches to be iterative, stepwise processes (e.g., Kuhlthau, 1993), empirical research supports models that consider information searches as dynamic and gradual processes with cognitive, affective, as well as behavioral dimensions (e.g., Griffin et al., 1999; for an overview of information searching models, see, e.g., Joseph et al., 2013; Ghasemaghaei and Hassanein, 2019) (e.g., Wilson et al., 2002; Hyldegård, 2006; Jiang et al., 2015; Orlu, 2016). The Risk Information Search and Processing model (RISP; Griffin et al., 1999), for instance, is based on the Heuristic-Systematic Model of information processing (HSM: Chaiken, 1980; Eagly and Chaiken, 1993) and the theory of planned behavior (Ajzen, 1991). It argues that searching for and evaluating information (e.g., the relevance and quality of information) are dependent on each other, meaning that cognitive processes related to both overlap simultaneously. Aside from the assumption that the processes of searching for and evaluating information go hand in hand, the RISP model, at its core, focuses on the psychological need for information sufficiency, which drives any search for information (Griffin et al., 2004; Yang et al., 2014). Furthermore, the model points out that personal characteristics, subjective norms, channel beliefs (i.e., beliefs about information channels, such as the Internet, which is considered a mediated information channel: Dunwoody and Griffin, 2014), and one's self-efficacy in sourcing information are important factors that drive the extent to which one performs critical elaboration while sourcing information. Similarly, many approaches on sourcing online information assume that individuals either process information in a heuristic or a systematic way depending on several factors (e.g., motivation or epistemic beliefs) (Metzger and Flanagin, 2013; Bromme and Goldman, 2014; Stadtler et al., 2017), and numerous empirical studies have found that individuals use various cues (e.g., the rank of a search result: Haas and Unkel, 2017) when selecting and evaluating online information (Sundar, 2008; Choi and Stvilia, 2015) (for an overview of discussed heuristics in evaluation of online information, see Sundar, 2008; Metzger and Flanagin, 2013). However, research about whether these cues are actually processed in a heuristic rather than a systematic way when searching for online information is still in its infancy (Schemer et al., 2008; Yang et al., 2014; Meinert and Krämer, 2022).

In addition to the RISP model, Brand-Gruwel et al. (2009) describe the processing of information when aiming to make informed decisions about scientific issues by focusing on the specific conditions of searching for information on the Internet. They describe five components for successfully solving information problems related to online searching behaviors: (1) Similar to the initial steps in the RISP model (Griffin et al., 1999), first, individuals define the information problem at hand; (2) they formulate corresponding search queries that are submitted to the search engine; (3) they evaluate search results presented on the search engine result page (SERP) to determine which information to access; (4) then, as in the RISP model, they evaluate the information provided by websites by considering aspects such as source parameters, their own prior knowledge, and information from other sources; and, finally, (5) they integrate information across multiple websites to reach a solution to the information problem (i.e., build a comprehensive mental representation of the problem and plan interventions; see also the processing of multiple documents, as considered in the Multiple Document Task-based Relevance Assessment and Content Extraction model (MD-TRACE) (Rouet and Britt, 2011). Accordingly, the model highlights the relevance of individuals' use of search engines and their formulation of search terms, which may affect the selection as well as the evaluation of online information. During web searches, for instance, selecting from the search results presented on a SERP requires choosing between a high number of alternative search results that usually only display sparse information (i.e., a title, short excerpt of the web page, and the URL). According to both models, a (pre-service) teacher's decision about which search results to click on (e.g., to check for further information) also depends on other factors, such as their prior knowledge, beliefs, or time capacities.

While pre-service teachers may use different strategies when sourcing online educational information (e.g., depending on their online sourcing competencies or on their individual epistemic beliefs, as will be outlined in 1.2 and 1.3), other factors, such as how they enter search terms, browse information, and select search results, are also impacted by media affordances. In this vein, media affordances (e.g., the algorithm a search engine uses) determine not only how specific media are used but also the ways in which individuals can engage with the technology (Evans et al., 2016). For example, when acquiring (scientific) information, (pre-service) teachers, like other information seekers, tend to use only one type of search engine (i.e., Google) (Bougatzeli et al., 2017), such that their sourcing will be limited by the default characteristics of the search engine and its SERPs, such as the algorithm the engine uses to present search results, the interface it offers for individuals to manually filter search results, or the sparsity of information it displays. The uncritical use of only one search engine means that one risks selecting search results in a biased way, as the results are predetermined by the affordances of the search engine; Kammerer et al. (2009), for instance, investigated individuals' interaction with two different search engines, one being a traditional query-based search system and one being an exploratory tag-based search system wherein individuals could interactively tag related search results. While the findings indicated that individuals' prior knowledge affected how many keywords they used for their inquiry, the use of the tag-based search interface was found to possibly compensate for differences in prior knowledge, as individuals who used the tag-based search engine used the tagging feature to give feedback on the relevance of search results, spent more time and were more engaged with the interface, and summarized their search inquiry by giving more arguments.

Furthermore, the algorithm a search engine uses to determine the order of results may influence whether a (pre-service) teacher selects any of the search results and whether they perform any further search queries. Research indicates that individuals would rather view/select the highest-ranked search results on a SERP (e.g., Eysenbach and Köhler, 2002; Pan et al., 2007; Wirth et al., 2007; Salmerón et al., 2013; Haas and Unkel, 2017). However, by selecting information only because of its rank on the SERP, one risks choosing information of low relevance or even low quality. In two experiments, Kammerer and Gerjets (2014) varied not only the trustworthiness and rank of the search results that were displayed on the SERP but also the interface of the search engine; they did this to investigate whether and how individuals select higher-ranked results even when they are less trustworthy. The students in the first experiment were highly impacted by the rank of the search results: when the search results on the top of the page were the less trustworthy ones, the students selected more of the least trustworthy search results and spent more time on them, and, vice versa, students selected fewer of the most trustworthy results and spent less time on them, which led students to list fewer arguments from the most trustworthy sources. In a follow-up experiment, this effect (namely, that individuals selected and viewed search results according to their ranking by the search engine, not their relevance) were highly decreased when the search engine's interface displayed the results in a three-by-three grid; thus, the affordances offered by the search engine matter.

Thus, it is reasonable that different search engine affordances not only lead to different search results but also impact how individuals conduct their search queries. Likewise, pre-service teachers' selection and evaluation of information depend both on how they conduct their searches (e.g., formulation of search keywords) (e.g., Hinostroza et al., 2018) and on personal factors (e.g., epistemic beliefs) (Kammerer and Gerjets, 2012).

Basic vs. advanced strategies of sourcing online information

We conclude that a variety of media affordances play a crucial role when sourcing online information. Furthermore, searching for information on the Internet is considered a complex process involving several searching behaviors (e.g., formulating search terms, evaluating search results presented on the SERP) (Brand-Gruwel et al., 2009). In this context, (pre-service) teachers can use several strategies to achieve a search task, e.g., search for relevant, appropriate, complete, and correct information on the Internet to ground an evidence-informed decision about an educational issue. However, defining which search strategies reflect competency (and which do not) is challenging, as defining the success of online information sourcing may differ depending on personal desires (e.g., one's epistemic aim in relieving uncertainty about a topic) or normative standards (e.g., achieving understanding about a topic in alignment with the requirement of a search task) (Hendriks et al., 2020).

While standards and policies for teacher education in general demand that (pre-service) teachers ground their decisions and actions in science-related evidence (e.g., Bauer and Prenzel, 2012), teacher competence frameworks also exist that aim to describe (pre-service) teachers' competencies in sourcing online information. For instance, the European Framework for the Digital Competencies of Educators uses a progress scale to define levels of searching strategies (i.e., Newcomer, Explorer, Integrator, Expert, Leader, Pioneer); it considers the use of other sources (e.g., official repositorium) in addition to a search engine as a very advanced strategy (i.e., Leader) for identifying and assessing relevant information and resources (Redecker and Punie, 2017). Similarly, the framework considers that evaluating the reliability of online information and resources and their suitability for an educational issue is an Expert-level search strategy.

Furthermore, research so far has used several indicators to describe behaviors related to searching strategies that likely lead to more relevant and appropriate search results and, thus, are considered more advanced search strategies. In this sense, the aspects of advanced search strategies most often focused on by researchers include the decision to use a certain search engine, the formulation of keywords to find, scan, evaluate, and select relevant search engine results, and the selection of the most relevant information (Hinostroza et al., 2018).

As described above, the affordances of media may crucially impact the results of one's search task. In this sense, it seems in particularly alarming that for any decision about what type of search engine to use, Google has the strongest dominance, with a global market proportion of over 92% (StatCounter, 2022). At the same time, different studies highlight the potential threat of search engines' biases, especially in terms how they may shape people's opinions, as individuals seem to over-rely on Google search results (Ballatore, 2015; Salehi et al., 2018). Hence, a skilled search strategy that limits the risk of search engines' biases might entail using multiple search engines, as this might decrease the outsized effects that one search engine's algorithm can have when pre-selecting and ranking search results.

Similarly, another crucial aspect of one's search strategy is selecting relevant search results. Again, several studies indicate that individuals have problems critically reading and scanning the lists of results on a SERP and tend to simply select the search results at the top of the list (e.g., Salmerón et al., 2013; Rieh et al., 2016). This means that selecting lower-ranked links on a SERP or links that are not on the first SERP might indicate a more critical consideration of more search results and, thus, might indicate advanced search strategies.

Furthermore, formulating search terms plays an important role in the search strategy, as writing complete sentences or using very few different search terms and synonyms could lead to results that are too general and irrelevant (Hinostroza et al., 2018). Accordingly, adapting and using new search terms is considered an advanced search strategy that might help (pre-service) teachers retrieve more relevant search results regarding their search task.

Finally, another important aspect of search strategies is evaluating the information and source quality (e.g., Bromme and Goldman, 2014; Redecker and Punie, 2017). Research indicates that selecting information is often influenced by criteria that are less relevant for actually evaluating the quality and relevance of information (e.g., design and usability; Hinostroza et al., 2018; rank on SERP; Haas and Unkel, 2017) than by criteria being more relevant for evaluating the information quality itself or the trustworthiness of the authors/sources [e.g., authors'/sources' expertise that may at least help pre-service teachers decide whether they can rely on the information provider, especially when they are not able to critically elaborate the quality of information, such as when they do not have enough time to do so (Bromme and Goldman, 2014)]. Thus, checking for the quality of information and for the trustworthiness of its authors/sources is considered an advanced strategy that likely helps pre-service teachers to retrieve relevant, appropriate, correct, and complete information (e.g., Bromme and Goldman, 2014).

(Internet-specific) epistemological beliefs and their role for pre-service teachers' searching behavior and selection of scientific evidence

According to Schommer (1990), epistemic beliefs consist of several dimensions that are relatively independent of each other and are conceptualized as beliefs about knowledge and how knowledge emerges. In general, one's beliefs about the nature of scientific knowledge—as part of epistemic cognition (Chinn et al., 2014)—may directly influence which strategies and practices they employ during online sourcing (Muis, 2007; Barzilai and Zohar, 2016; Hendriks et al., 2020). Bråten et al. (2005) were the first to investigate special aspects of epistemic beliefs regarding the Internet (which is considered an information channel in the RISP model: Dunwoody and Griffin, 2014). They argue that because "hypermedia technologies such as the Internet allow for new ways of presenting knowledge and new ways of knowing, measures of personal epistemology should probably focus specifically on beliefs about the nature of knowledge and knowing in such technological environments" (Bråten et al., 2005, p. 147). As such, they invented a measurement of epistemic beliefs that focuses specifically on beliefs about the nature of knowledge and knowing in Internet-based environments (i.e., ISEBs).

Like epistemic beliefs in general, an individual's ISEBs may also influence how they search for information on the Internet. In this sense, one study found that students with more educational years tended to have advanced epistemic beliefs regarding the uncertainty of Internet-based knowledge (i.e., constructivist oriented), which made them more likely to suspect that the Internet is a good source containing accurate knowledge, as well as more inclined to justify and evaluate Internet-based knowledge with other sources (Chiu et al., 2016). Research on (pre-service) teachers' ISEBs and their sourcing behavior seems inconsistent, as some research indicates that in-service teachers' advanced epistemological beliefs could mean that they use more sophisticated online search strategies (i.e., selecting less irrelevant information) to filter and organize information than those with less advanced beliefs (Tsai et al., 2011); yet, other research indicates that pre-service teachers' ISEBs did not have a significant impact on their online search strategies (Yilmaz and Çakmak, 2016).

When it comes to (pre-service) teachers' preferences for scientific or anecdotal evidence (e.g., experiences of colleagues), research has indicated that (pre-service) teachers tend to prefer anecdotal evidence (e.g., Bråten and Ferguson, 2015; Kiemer and Kollar, 2021). Of course, relying on anecdotal evidence can be important in, for example, determining the practicability of certain teaching methods in specific situations. However, anecdotal evidence rarely meets the systematic standards for knowledge generation that forms of scientific evidence often do (e.g., Spencer et al., 2012). In a recent study by Hendriks et al. (2021), pre-service teachers judged the trustworthiness of a researcher vs. An experienced teacher depending on what epistemic aims the pre-service teachers held (i.e., their aims at achieving epistemic ends, such as gathering knowledge or getting practical explanations; see also Chinn et al., 2014); when pre-service teachers aimed for a theoretical explanation about schooling, they judged the researcher to be more trustworthy. Thus, it seems as though pre-service teachers select evidence according to their epistemic beliefs, as their judgment of certain sources of information depending on their epistemic aims indicates that they have assumptions about how the source can help fulfill their epistemic aims. In this sense, pre-service teachers' epistemic beliefs about the nature of knowledge and knowing on the Internet may affect not only their actual searching behavior but also whether they select scientific vs. anecdotal evidence.

Rationale of this study

Given this theoretical and empirical background, we wanted to describe pre-service teachers' sourcing strategies by focusing on behavioral processes related to their selection as well as evaluation of online information (Griffin et al., 1999). We focused on aspects of search strategies that are considered specific to sourcing information on the Internet and play a crucial role in sourcing relevant information (e.g., use of search engines; Brand-Gruwel et al., 2009; Kammerer and Gerjets, 2014). As pre-service teachers' ISEBs may also relate to how they source information on the Internet (e.g., Tsai et al., 2011; Dunwoody and Griffin, 2014), we additionally assessed participants' ISEBs.

The goals of the present research were twofold: First, we aimed at describing pre-service teachers' searching strategies when sourcing online educational information (Research Question 1). Therefore, we investigated whether participants used basic vs. advanced search strategies and, thus, analyzed several behavioral aspects that are considered crucial for the competencies in sourcing information on the Internet; these aspects included (1) the frequencies of types of search engines used, (2) the number of used search engines, (3) the adaptation or formulation of (new) search terms, (4) the selected information's rank on SERP and the SERP page number it came from, and (5) the instance of any type of quality check (e.g., Salmerón et al., 2013; Bromme and Goldman, 2014; Hinostroza et al., 2018). Second, we investigated the relation of pre-service teachers' ISEBs to their searching strategies as well as to the science-relatedness of their selected information (Research Question 2).

> RQ1: How do pre-service teachers search for online educational information, and what strategies do they use when sourcing these?

> RQ2: Are pre-service teachers' Internet-specific epistemic beliefs related to their searching strategies and/or to their selection of evidence?

Materials and methods

Participants

Study participants included 91 pre-service teachers from three universities in Germany who were studying at the bachelor's or master's degree level to become secondary school teachers. Participation was voluntary, and participants received an allowance of 20€. Data from 12 participants were excluded (1) due to issues in recording their search behaviors via screen video, (2) due to issues with the Internet connection during the investigation, or (3) because the time they spent on conducting the experiment differed more than one standard deviation from the mean duration. This resulted in a final sample of $N = 77$ participants (51 females and 1 diverse) aged 18–41 years ($M = 25.29$, $SD = 5.06$). The participants' average length of study at the time of the survey was 4.4 semesters ($SD = 2.99$). The time spent on conducting the search task was $M = 15.76$ min ($SD = 8.64$). Of the sample, $n = 27$ participants were studying at the master's level and $n = 50$ were at the bachelor's level.

Participants reported that they used a computer, notebook, or et for an average of $M = 4.01$ ($SD = 2.51$) hours per week. The average time spent on the Internet was reported to be $M = 5.01$ ($SD = 3.05$) hours per week. The weekly time for information seeking on the Internet was reported to be $M = 1.99$ ($SD = 1.55$) hours per week, and for online information seeking about educational topics they reported to invest an average of $M = 1.56$ ($SD = 1.22$) hours per week. Participants rated their self-perceived prior knowledge about the topic "students' use of mobile phones in class," as neither very low nor very high (i.e., based on four items: $M = 2.48$; $SD = 0.82$). Participants' attitudes toward banning mobile phones was balanced (i.e., based on four items: $M = 2.96$; $SD = 1.04$).

Procedure

The investigation was conducted from November 2019 to January 2020. As the investigation was performed on-site at the university, participants had access to the network and freely accessible licensed scientific books and sources of the university. Each participant worked alone in front of the computer, at their own pace, guided by the instructions of the online survey (by Questback EFS Surveys) (i.e., without verbal instructions from the investigators). In the beginning of the study, an open web browser window (i.e., Mozilla Firefox) was on display showing all participants the same university website.

In the beginning of the survey, the demographic variables were assessed, as were participants' self-reported ISEBs. In the next step, the following fictional scenario was constructed: All participants were asked to imagine themselves as teachers. They had the task of searching for information about mobile phone use in class in preparation for a fictional school conference about this topic (see **Supplementary material 1**). For this, they were asked to select relevant web content (henceforth called *web items*, WI) that would allow them to build an opinion about the topic. Then, all participants were asked to search for about 20 min for educational information about the topic and to select two or four WI (see also, section "The number of selected web items related to the search task as control variable"). Accordingly, all participants sought pedagogical information on the same topic of "students' use of mobile phones in class" on the Internet. Based on their actual search results, they selected online WI that they perceived to be relevant for forming opinions and making decisions. Participants were allowed to select any type of WI (e.g., scientific articles, videos, and blog entries). During the search task, participants' search behavior was captured by recording their screens.

The educational topic

Students' use of mobile phones in class is a highly debated topic. It is not only a question of school administration but also a topic that is frequently addressed by the media and academics. Schools in Europe, and in Germany's federal states, regulate the use of mobile phones in classes very differently. The research within the field of educational sciences deals with the advantages and disadvantages of mobile phone use in class regarding students' attention and learning outcomes, as well as students' social and digital competencies (e.g., Sung et al., 2016). This topic was selected for the search task because it has practical relevance and because diverse and conflicting educational evidence can be found on the Internet.

The number of selected web items related to the search task as control variable

The data in this study were collected as part of a larger online experiment with the hypothesis that participants would reason their selection of WI differently depending on whether they reason in an individual or collaborative setting (Zimmermann and Mayweg-Paus, 2021). After participants were told about the search task (which is reported here and was nearly the same for all participants regardless of which experimental condition they were eventually assigned to) participants were divided into two groups according to the experimental conditions. As part of the experiment, the only aspect that differed between experimental conditions during the search task was that participants were asked to select either four WI in the individual reasoning condition or two WI in the collaborative condition. Thus, $n = 33$ participants were part of $group_{4WebItems}$, and 50 participants were placed in $group_{2WebItems}$.

Since in this study we exclusively focused on the information search process (i.e., the search task of the experiment), in the following we do not differentiate between individual and collaborative reasoning settings. However, as participants either selected two or four WI during the search task, in this study we can control for any effects due to the number of selected WI related to the search task (i.e., two vs. four WI). In this sense, the number of search results/links that students were told to select might also have impacted their searching behavior. Thus, in a preparatory analysis, we analyzed whether the number of selected WI related to the search task had any influence on participants' searching strategies, i.e., (1) the number of search engines they used, (2) whether they adapted or formulated new search terms; (3) the rank and number of SERP associated with the WI they selected; and (4) whether they performed a quality check.

Four binary logistic regressions with the categorial independent variable two vs. four WI were analyzed. The dependent variables were defined as binary variables (i.e., as in the main analysis). The independent variable was not found to contribute to the models, meaning that selecting two vs. four WI did not significantly influence participants' search behavior: (1) $\beta = -0.16$, $SE = 0.48$, $Wald = 0.11$, $p = 0.74$; (2) $\beta = 1.45$, $SE = 1.11$, $Wald = 1.71$, $p = 0.19$; (3) $\beta = 0.21$, $SE = 0.34$, $Wald = 0.39$, $p = 0.53$, and (4) $\beta = -0.33$, $SE = 0.54$, $Wald = 0.38$, $p = 0.54$. All together, this means that the number of WI that the participants were told to select during the search task (i.e., whether participants selected two vs. four WI) had no significant influence on the dependent measures and, thus, was not included in our main analyses. Therefore, the results reported below come from analyzing the first two selected WI of participants in the $group_{4WebItems}$ as well as the two selected WI of participants in the $group_{2WebItems}$.

Measurements

Science-relatedness of selected web items

In sum, we analyzed 154 WI that were selected by participants (i.e., two WI for each participant). We considered those WI to be science related if the content referred to primary or secondary scientific sources (i.e., scientific journal articles, scientific reports, monographs, scientific blogs, school textbooks, or university theses). They were considered not to be science related if the content referred to journalistic sources or anecdotal evidence (i.e., online news portals, information platforms, or blogs or YouTube videos by teachers) (first author and last author, 2021).

Basic vs. advanced search strategies

To shed light on participants' search strategies, we followed in line with the literature and analyzed the following aspects: (1) the frequencies of the search engine types used among all participants and for each participant, (2) the number of search engines participants used during the entire searching process, (3) whether participants adapted or formulated new search terms during the searching process, (4) the rank on the SERP and the SERP page number associated with the WI they selected, and (5) whether they checked the quality of sources during the search process.

The *frequencies of the search engine types used* (e.g., Google or Ecosia) refers to the two selected WI (i.e., the type of search engine that was ultimately used to find the selected WI). The frequencies represent how often the search engines were used among all participants.

The *number of search engines used* indicates the quantity of different search engines a participant used during his/her search task (i.e., even if the used search engine did not lead to the final selected WI). In cases where a participant used at least two search engines to retrieve web results (e.g., to compare results or to conduct further research), the search strategy (in terms of the number of used search engines) was considered advanced.

The variable *adapting or formulating new search terms* indicates whether a participant (1) specified searches by using variations of the same search term, (2) used new search terms, or (3) used a mixed strategy that included specified terms and new search terms. In the case that a participant used one or both strategies, the search strategy (in terms of search term adaptation) was considered advanced.

The variables *number of SERP* and *rank of WI on SERP* indicate, respectively, whether the WI was selected on the first, second, or subsequent SERP and, when it was selected from the first SERP, what its rank was. When participants did not simply select one of the first four WI on the first SERP but instead selected at least one of the two WI from a lower rank or from one of the following SERPs, this was considered an advanced strategy, as it indicates that participants considered more than only the highest-ranked WI.

The variable *quality check* indicates whether a participant checked for quality (i.e., the trustworthiness of the author/source or the credibility of the statements via hyperlinks). For instance, we considered it a quality check when a participant examined a prior search result by using the name of the provider as a search term. In the case that a participant checked for quality during the search process, this was considered an advanced search strategy.

Lastly, to examine each participant's overall search strategy for the entire search process, an *overall index* was calculated to give insights into participants' competencies in sourcing. The index was calculated based on the four aspects of a search strategy, namely whether participants (1) used more than one search engine, (2) formulated new or adapted the search terms, (3) did not select WI from the first four ranks on the first SERP, and (4) checked for quality of sources and content. Thus, for each participant, an index of $i = 0$ was calculated, and the value was added by $i + 1$ if one of the search strategy aspects was fulfilled. Thus, the index depicts five competence levels of information searching (from 0 to 4) that are described as follows: basic search strategy, advanced search strategy, intermediate search strategy, proficient search strategy, expert search strategy.

Internet-specific epistemological beliefs

We assessed participants' ISEBs based on the questionnaire by Bråten et al. (2005). The questionnaire addresses dimensions concerning web-based knowledge (what one believes that knowledge is like on the web) and web-based knowing (how one comes to know on the web). The 14 items yielded an internal consistency of Cronbach's $\alpha = 0.87$.

Results

Science-relatedness of selected web items

Of the total 154 WI participants selected, 32 WI (20.8%) were determined to be science related (see first author and last author, 2021, for a list of all selected WI, incl. hyperlinks and how often they were selected among all participants). We also analyzed whether any (and how many) of the science-related WI were among those WI that participants considered relevant for building an opinion about the search topic. Fifty out of all 77 participants did not select any science-related WI; the other 27 participants selected at least one WI that was determined to be science related. Interestingly, these findings indicate that while most participants did not select any science-related WI, still about one-third of the participants at least considered scientific evidence in addition to other forms of information (e.g., anecdotal evidence from teacher colleagues in blogs, or journalistic information) for building an opinion about a pedagogical problem.

The type of search engines

Google was by far the most frequently used search engine. More than three-quarters of the selected WI (78.0%) were found via Google. Furthermore, Google Scholar (6.0%), YouTube (2.7%), Google Videos (2.0%), Google News (2.0%), and Google Books (0.7%) were used for WI selection. Thus, over 90% of the selected WI were reached via search engines of Google LLC. The second most often used search engines were the university's library search engine (i.e., Primus) and Google Scholar, considered two scientific search engines. The only used commercial search engine that was not associated with Google LLC was Ecosia (2.7%), which is a non-scientific search engine. Table 1 displays the frequencies of all used search engine types by referring to participants' selected WI.

Results in terms of search strategies

Use of more than one search engine

We analyzed the use of more than one search engine, as this is considered an advanced search strategy. Most participants (62.3%) did use more than one search engine: About one-third of participants used either two or three search engines, and a small group (5.2%) even used four search engines. In contrast, still more than one-third of all participants used only one search engine (37.7%) (Table 2). Even though a rather high proportion of participants used an advanced search strategy, namely using at least two search engines, it is important to again highlight that even different search engines might be associated with a single company (i.e., as in this study: Google LLC).

Adaptation or formulation of new search terms

Another important aspect of a skilled search strategy is formulating new search terms or adapting the search terms during the search process. The strategy adapting search terms was used by almost one-third of participants (29.9%), while almost one-fifth (19.5%) formulated new search terms. Both strategies were used by $n = 32$ participants (41.6%), while $n = 7$ (9.1%) used only one search term (Table 3). Accordingly, $n = 70$ participants (90.9%) used an advanced search strategy regarding formulating new or adapting search terms.

Position of selected web items on search engine result page

In terms of 152 WI, almost all the selected WI (148 WI, 97.4%) stemmed from the first SERP (Table 4). Furthermore, in terms of 151 WI, most of the selected WI (104 WI, 68.8%) were selected from those WI that were highly ranked (i.e., in first four listed results) (Table 5: The coding of two –respectively, three—WI was not possible due to technical issues). As described above, we also analyzed whether participants considered lower-ranked WI for their sourcing process. Results showed that for 54.7% of all participants, at least one of their WI (of the two selected) stemmed neither from the first four ranks nor from the first SERP, indicating an advanced search strategy.

TABLE 1 Frequencies of used search engine types.

Search engines	WI*	Percentage
Google	117	78.0
University library search engine	9	6.0
Google scholar	9	6.0
Ecosia	4	2.7
YouTube	4	2.7
Google videos	3	2.0
Google news	3	2.0
Google books	1	0.7
Total	150	100

*All participants selected two WI. Coding of four WI was not possible due to technical issues.

Quality check

Another important aspect of skilled search strategies is checking the quality of sources and content during a search process. For participants' analysis of their search processes, we considered two different forms of quality checks. First,

TABLE 2 Number of used search engines.

Number of search engines used	Frequency	Percentage
Only one search engine used	29	37.7
Two search engines used	23	29.9
Three search engines	21	27.3
More than three search engines	4	5.2
Total	77	100

TABLE 3 Adaption of or using of new search term.

Form of adaption	Frequency	Percentage
Search term specified	23	29.9
New search term	15	19.5
Specified and new search term	32	41.6
Only one term for all WI	7	9.1
Total	77	100

TABLE 4 Search engine's result page.

Selected result page	WI*	Percentage
First page	148	97.4
Second page	3	2.0
Third page	1	0.7
Total	152	100

*All participants selected two WI. Coding of two WI was not possible due to technical issues.

TABLE 5 Rank of WI on the result page.

Rank result page	Responses	
	WI*	Percentage
1	37	24.5
2	23	15.2
3	26	17.2
4	18	11.9
5	7	4.6
6	10	6.6
7	10	6.6
8	7	4.6
9	6	4.0
10	7	4.6
Total	151	100

*All participants selected two WI. Coding of three WI was not possible due to technical issues.

TABLE 6 Quality check.

Form of quality check	Frequency	Percentage
No quality check	56	72.7
Author/source has been checked	11	14.3
Source was checked	10	13.0
Total	77	100

checking the author/source on other websites was done by $n = 11$ participants (14.3%). Second, checking sources or statements in relation to the content of the WI was done by $n = 10$ participants (13.0%) (**Table 6**). Conclusively, $n = 21$ participants (27.3%) carried out at least one quality check and, therewith, used an advanced search strategy.

Participants' overall search strategies

Lastly, we calculated an overall index to describe an advanced search strategy; the index was calculated based on the aforementioned aspects: (1) use of more than one search engine, (2) new/adaption of search terms, (3) no selection of a WI from the first four ranks on the first SERP, and (4) quality check. Only a small group of $n = 4$ participants (5.2%) performed a very basic search strategy, meaning that no criteria were met. This group only entered one search term, only used one search engine, selected their two WI from the first four ranks on the first SERP, and did not check for quality. Another small group of $n = 8$ participants (10.4%) pursued an advanced search strategy in which only one of the four advanced search strategies was applied. The largest group, $n = 32$ participants (41.6%), performed an intermediate search strategy using two of the four advanced search strategies. Almost one-third of participants, $n = 24$ (31.6%), used a proficient search strategy in which three of the four advanced strategies were applied. The group that used an expert search strategy, in which all four advanced search strategies were used, consisted of $n = 9$ participants (11.7%) (see **Table 7**). This group used more than one search engine, adapted the search term (or used new search terms), selected at least one WI that was not found on the first SERP within the first four ranks, and, finally, checked for quality.

The relation between Internet-specific epistemological beliefs and basic vs. advanced search strategies as well as science-relatedness of web items

Four binary logistic regressions were used to analyze the relationship between ISEB and use of the four aspects of search strategies as dependent variables. The dependent variables were defined as binary variables, i.e., participants either (1) used more than one search engine, (2) used new search terms, (3) did not select a WI from the first four ranks on the SERP, or (4) performed quality checks. The predictor ISEB did not contribute to the models, meaning that differences in participants' ISEBs did not significantly relate to their search behavior, (1) $\beta = 0.21$, $SE = 0.39$, $Wald = 0.29$, $p = 0.59$; (2) $\beta = 0.04$, $SE = 0.65$, $Wald = 0.005$, $p = 0.95$; (3) $\beta = 0.21$, $SE = 0.38$, $Wald = 0.31$, $p = 0.58$, and (4) $\beta = -0.20$, $SE = 0.42$, $Wald = 0.24$, $p = 0.63$.

To investigate whether participants' ISEBs related to their selection of WI, two further binary logistic regressions were conducted with the first as well as the second selected WI as dependent binary variables (science-related vs. not science-related). Again, the predictor ISEB was not found to contribute to both models, meaning that differences in participants' ISEBs did not significantly relate to whether the first or second selected WI was science related (first selected WI: $\beta = 0.36$, $SE = 0.45$, $Wald = 0.64$, $p = 0.43$; second selected WI: $\beta = -0.05$, $SE = 0.44$, $Wald = 0.01$, $p = 0.91$).

Discussion

Pre-service teachers' search strategies

With respect to describing pre-service teachers' search strategies in sourcing online educational information (RQ1), we analyzed several aspects as indicators of participants' search behaviors in terms of their use of different search engines, their adaptation of keywords to find WI, as well as their selection and evaluation of relevant WI. Interestingly, more than 90% of participants used search engines from Google LLC (e.g., Google's search engine or YouTube) to select their WI. As one aspect of an advanced search strategy, 29.9% of participants used two, 27.3% used three, and 5.2% used four search engines to combine their search results. However, 37.7% of all participants limited their searches to only one search engine. While it is promising to see that the other 62.4% of participants used more than one search engine, most of these search engines were associated with Google LLC. In terms of the position of WI, an alarming majority of selected WI (68.8%) were ranked among the first four search results of the first SERP. Thus, participants' selections of WI that they perceived relevant for building an opinion about the educational topic might be influenced by the specific media affordances of Google (e.g., search engine algorithm, interface) (Pan et al., 2007; Kammerer et al., 2009; Haas and Unkel, 2017).

In terms of adapting search terms as an aspect of a skilled search strategy, almost every participant (90.9%) either adapted the preliminary search term or formulated new search terms during their sourcing process, which likely helped them to retrieve more relevant WI (e.g., Hinostroza et al., 2018).

Checking the quality of sources and content is considered an important aspect of search strategies related to the evaluation of

TABLE 7 Sum index about applied advanced search strategies.

Index about applied advanced search strategies	Frequency	Percentage
Basic search strategy	4	5.2
Advanced search strategy	8	10.4
Intermediate search strategy	32	41.6
Proficient search strategy	24	31.2
Expert search strategy	9	11.7
Total	77	100

For all tables, the sum of the percentages may deviate slightly from 100% due to rounding.

information (e.g., Bromme and Goldman, 2014). In this sense, we analyzed (1) whether participants used the name of sources as a subsequent search term to indicate any checking of source's trustworthiness and (2) whether participants clicked on further hyperlinks to indicate whether they checked for relevance and the quality of the online information. It turned out that only about one-quarter of all participants carried out such strategies in evaluating the information (27.3%).

To give participants an overall competency score for their information sourcing, we combined all the single aspects into an index; this allowed us to see how many participants achieved more than one criterion of an advanced search strategy. In this vein, a promising number of n = 33 participants (43.3%) used a proficient or expert search strategy, in which they applied three or four, respectively, out of the four criteria for advanced strategies. However, still n = 12 participants (15.6%) did not fulfill any or more than one criterion of advanced search strategies.

(Pre-service) teachers are encouraged to base their decisions and practices on evidence from educational research (e.g., Bauer and Prenzel, 2012; Fischer et al., 2014), as scientific evidence is most relevant to pedagogical issues and has a high degree of certainty (Spencer et al., 2012). In a direct comparison, (pre-service) teachers seem to prefer anecdotal rather than scientific evidence (e.g., Bråten and Ferguson, 2015; Kiemer and Kollar, 2021). Accordingly, it is interesting to see that in this study, pre-service teachers indeed selected science-related WI in addition to non-science-related WI (e.g., with anecdotal evidence in blogs from teachers), whereby about one out of every five WI was science related (i.e., primary scientific sources).

With respect to the relation of pre-service teachers' ISEB to their searching strategies as well as their selection of evidence (RQ2), participants' ISEB did not significantly relate to their search strategies nor the science-relatedness of their selected WI. While this result is in line with some previous research (Yilmaz and Çakmak, 2016), it is still not possible to make a conclusive statement about any possible relation, since some research findings indicate that there is a relation between ISEB and search strategies (Tsai et al., 2011).

Limitations

With respect to measuring pre-service teachers' search strategies, it is important to mention that the described aspects (i.e., used search engine types, number of used search engines, adaptation of search terms, the rank of the selected WI on the SERP, and the quality check) were determined by using the screen-recorded videos that showed participants' search behavior during a realistic search task, where participants were allowed to select those WI that they perceived relevant for building an opinion about the educational topic. Thus, even though this study's search task represents a relatively externally valid investigation, the analyzed aspects only serve as indicators for pre-service teachers' actual competencies in sourcing online educational information. In this sense, these aspects only serve as hints on certain manifestations of sourcing competencies, as we only assessed participants' sourcing behaviors that were visible on their screens (e.g., selected WI, formulation of search terms). As such, we were not able to indicate, for instance, what information participants' read while browsing through the SERP's search results. In this sense, our use of the mere presence or absence of certain searching aspects represents only a first insight into pre-service teachers' actual sourcing competencies. For example, a participant may not have needed to check for a source's trustworthiness (i.e., use the name of the author/source as new search term during the search task)—no matter whether this checking refers to anecdotal evidence (e.g., of a teacher colleague in a YouTube video) or to scientific evidence (e.g., a journal article)—if they were already familiar with the author/source and, thus, might have already been aware of their level of expertise. Similarly, with respect to calculating the index for a participant's overall sourcing competency, we must note the four aspects used to calculate the index may have different levels of relevance for assessing sourcing competency. For example, regarding effectively and efficiently selecting relevant online information, the number of used search engines may not be quite as important as checking the quality check of a WI. Despite this limitation, the index provides a useful heuristic that gives an overview of pre-service teachers' overall search strategies by revealing the presence (or absence) of the single aspects related to skilled search strategies.

Regarding pre-service and in-service teachers' sourcing of educational information, one of the most important reasons that they may search for information on the Internet is simply because information on the Internet is accessible. However, in terms of searching for primary scientific evidence (e.g., scientific articles), even the Internet has obstacles, and teachers usually do not have the same access to such content that individuals at universities do. In our study, the participating pre-service teachers had access to the university network and, thus, were able to access scientific journal articles. Future research, thus, might investigate whether (pre-service) teachers' selection of science-related evidence might differ when they do not have access to scientific journals or how the search strategies may differ between pre-service and in-service teachers, for example because in-service teachers may have fewer capacities to spend time on intensive sourcing of online information. In this vein, we must also note that in our study, future teachers were explicitly asked to source online information to build their opinions. In the everyday life of a teacher, this prerequisite does not necessarily exist and teachers perhaps not engage in any research reception and evidence-informed practices at all, for instance, because they may perceive a lack of sourcing skills or a lack of time (Thomm et al., 2021a).

While the topic of students' mobile phone use in class is highly relevant for (pre-service) teachers, this study only focused on a single educational topic. As such, we cannot necessarily generalize pre-service teachers' search strategies to other search tasks and topics. Hence, future research may expand these findings by focusing on different topics. In this context, it might be interesting to investigate whether the degree of scientific certainty related to a topic might lead to differences in pre-service teachers' searching behavior as well as to different results in terms of any impact of pre-service teachers' ISEBs on these searching behaviors. Similarly, it would be valuable for future research to explicitly examine whether the sourcing behavior of (pre-service) teachers changes over time and whether both technical and societal developments have an impact on their search strategies (e.g., with respect to developments of the search engine's adjustments to filter search results or a pandemic-related situation where teachers may have had to search more frequently for online information to adjust their teaching). In this context, we assume that the results of the present study can also be transferred to today's search strategies of (pre-service) teachers, at least to a large extent, since the findings indicate above all that (pre-service) teachers' awareness of the relevance of any influence of the search engines' affordances could be increased.

Conclusion and implications

In this study, we described pre-service teachers' search strategies when sourcing online educational information about the topic "students' mobile phone use in class" by focusing on several aspects of their sourcing behavior that indicate skilled search strategies.

As these aspects are only considered indicators for competency in sourcing online information (e.g., Hinostroza et al., 2018), future research my expand this study's description of pre-service teachers' search strategies by investigating whether any *advanced* strategy (e.g., adaptation of search terms or use of more than one search engine) indeed leads to more relevant information that is also of good quality. In this sense, any attempt to educate pre-service teachers on the skills related to advanced search strategies would also benefit from knowing which aspects of an advanced search strategy are the most important for retrieving relevant information. Yet, defining what exactly constitutes success in sourcing online information is challenging (e.g., Hendriks et al., 2021), so future research may benefit from considering not only pre-service teachers' self-perceived relevance of information but also objective judgments about this information's quality and relevance (e.g., ratings from educational research experts). In line with the RISP model (Griffin et al., 1999)—and other approaches that consider multiple ways of processing (online) information—determining success in sourcing online may also be related to personal factors, such as ones' motivation to find relevant information, and, thus, could be considered in future research. Similarly, considering a pre-service teacher's degree of motivation for retrieving relevant online information, future research should consider their personal background knowledge about search engines at a declarative as well as a procedural level (e.g., to assess whether they understand how search engines and their algorithms work and how interfaces can be used), as differences in such knowledge may also influence their sourcing behaviors. Lastly, with respect to in-service teachers' evidence-informed practices, future studies may examine how any selection of relevant online educational information (e.g., scientific evidence when it comes to theoretical explanations or reports from colleagues when it comes to practical tips: Hendriks et al., 2021) indeed is used, for instance, for teachers' actual lesson preparation and whether teachers' awareness about potentially influences caused by the media affordances on the Internet indeed leads to, for instance, finding and selecting the most relevant online information. Overall, searching for, selecting, and evaluating relevant information is, of course, not only important for the teaching profession, but sourcing information competently is of special relevance for teachers, as it relates not only to their own lifelong learning and evidence-informed practices, but it may also influence how they teach their students to competently source relevant online information (e.g., Caena and Redecker, 2019).

Taken together, the findings of this study indicate that most pre-service teachers adapted their search terms to retrieve more relevant information. However, at the same time, it is alarming that most of the pre-service teachers were likely influenced by the affordances of Google's search engines, as they often only used one search engine and almost always selected information that was ranked highly on the SERP, even though the search task was to select information that they perceived to be relevant for building an opinion about the topic. In this sense, the study clearly emphasizes that it is important to increase (pre-service) teachers' awareness about media affordances and about their own search strategies when sourcing online educational information. This has practical implications for teacher trainings and in-service teachers' sourcing practices. Also, the findings raise the question of whether increasing awareness about the potential influence of search engines' affordances may help (pre-service) teachers overcome the biases of search engines or go beyond using only basic search strategies that enforce these biases (e.g., only selecting high-ranked information). So far, it is unclear whether merely alerting pre-service teachers about the potential influence of search engines helps to increase their awareness of how search engine algorithms work; alternatively, it may also be necessary to implement specific interventions into teacher trainings that educate them about not only why but also how to use more advanced search strategies to retrieve relevant online educational information efficiently and effectively (e.g., how to adapt search terms effectively or why and when to use more than one search engine). In this context, it seems promising to implement interventions into teacher trainings that focus on fostering pre-service teachers' critical reflection about their own search strategies (e.g., settings in which pre-service teachers collaboratively reflect on their own search strategies; first author and last author, 2021). In particular, promoting (pre-service) teachers' critical questioning of their own search strategies and, for instance, whether selected online information indeed is relevant—as metacognitive strategies (Kuhn, 1999)—may allow them to face the challenges of an ever-evolving media environment on the Internet.

By describing pre-service teachers' use of basic vs. advanced strategies, this study provides a foundation for further in-depth investigations into the strategies pre-service teachers use to search for online educational information. In this context, individual aspects like motivation, knowledge about search engines, or epistemic beliefs about knowledge from the Internet should be considered. Furthermore, the study points at the importance of considering the complex array of media aspects that influence online information searches (i.e., media affordances, such as search engines' algorithms) as well as the importance of fostering pre-service teachers' awareness about potential biases caused by these aspects. All in all, it seems important to foster pre-service teachers' critical reflection about their own search strategies and to additionally promote their knowledge about search engines' affordances and the potential biases caused by using certain search engines; giving them this type of knowledge may increase their *critical* search strategies (e.g., reflecting critically about whether the high-ranked information indeed is the most relevant information, or whether authors are trustworthy). This seems particularly important, as this study's findings indicate that pre-service teachers need to improve their sourcing competencies regarding using search engines, selecting information on SERPs, and checking for information quality and sources' trustworthiness.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MZ and EM-P conceived the idea of the study. MZ collected the data and took the lead in writing the manuscript. MZ and OE analyzed the data and engaged in writing sections of the manuscript. All authors provided feedback and ideas.

Funding

This research was supported by the Einstein Center Digital Future (ECDF). The Einstein Center Digital Future had no involvement in study design, data collection, analysis and interpretation, and the decision to submit the article for publication. The article processing charge was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—491192747 and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

Acknowledgments

We thank Thu Trang Phi, Jonas Stampka, and Lisa Tinkl for their help with data collection and processing. The data reported in this manuscript stems from an experiment reported and published as: First author, and last author (2021). Anonymization for blind-review. The presented data and findings in this manuscript were never published before and serve additional explanatory analyses that go beyond the examination of the assumed hypotheses of the experimental investigation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.976346/full#supplementary-material

References

Ajzen, I. (1991). The theory of planned behavior. *Theor. Cogn. Self Regul.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Ballatore, A. (2015). Google chemtrails: A methodology to analyze topic representation in search engine results. *First Monday* 20:5597. doi: 10.5210/fm.v20i7.5597

Barzilai, S., and Zohar, A. (2016). "Epistemic (Meta) cognition: Ways of Thinking about knowledge and knowing," in *Handbook of Epistemic Cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (London: Routledge).

Bauer, J., and Prenzel, M. (2012). European teacher training reforms. *Science* 336, 1642–1643. doi: 10.1126/science.1218387

Bougatzeli, E., Douka, M., Bekos, N., and Papadimitriou, E. (2017). Web literacy practices of teacher education students and in-service teachers in Greece: A descriptive study web reading practices of teacher education students and in-service teachers in Greece: A descriptive study. *Presch. Prim. Educ.* 97–109. doi: 10.12681/ppej.10336

Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004

Bråten, I., and Ferguson, L. (2015). Beliefs about sources of knowledge predict motivation for learning in teacher education. *Teach. Teach. Educ.* 50, 13–23. doi: 10.1016/j.tate.2015.04.003

Bråten, I., Strømsø, H. I., and Samuelstuen, M. S. (2005). The relationship between internet-specific epistemological beliefs and learning within internet technologies. *J. Educ. Comput. Res.* 33, 141–171. doi: 10.2190/E763-X0LN-6NMF-CB86

Bromme, R., and Goldman, S. R. (2014). The Public's bounded understanding of science. *Educ. Psychol.* 49, 59–69. doi: 10.1080/00461520.2014.921572

Bromme, R., Prenzel, M., and Jäger, M. (2014). Empirische bildungsforschung und evidenzbasierte bildungspolitik. *Z. Für Erzieh.* 17, 3–54. doi: 10.1007/s11618-014-0514-5

Caena, F., and Redecker, C. (2019). Aligning teacher competence frameworks to 21st century challenges: The case for the European Digital Competence Framework for Educators (Digcompedu). *Eur. J. Educ.* 54, 356–369. doi: 10.1111/ejed.12345

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Pers. Soc. Psychol.* 39, 752–766. doi: 10.1037/0022-3514.39.5.752

Chen, Y.-J., Chien, H.-M., and Kao, C.-P. (2019). Online searching behaviours of preschool teachers: A comparison of pre-service and in-service teachers' evaluation standards and searching strategies. *Asia Pac. J. Teach. Educ.* 47, 66–80. doi: 10.1080/1359866X.2018.1442556

Chinn, C., Rinehart, R., and Buckland, L. (2014). "Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition," in *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*, eds D. Rapp and J. Braasch (Cambridge, MA: MIT press), 425–453.

Chiu, Y.-L., Liang, J.-C., and Tsai, C.-C. (2016). Exploring the roles of education and internet search experience in students' internet-specific epistemic beliefs. *Comput. Hum. Behav.* 62, 286–291. doi: 10.1016/j.chb.2016.03.091

Choi, W., and Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *J. Assoc. Inf. Sci. Technol.* 66, 2399–2414. doi: 10.1002/asi.23543

Duke, T. S., and Ward, J. D. (2009). Preparing information literate teachers: A metasynthesis. *Libr. Inf. Sci. Res.* 31, 247–256. doi: 10.1016/j.lisr.2009.04.003

Dunwoody, S., and Griffin, R. (2014). "The role of channel beliefs in risk information seeking," in *Effective risk communication*, eds J. Arvai and L. Rivers III (London: Taylor & Francis), 220–233.

Eagly, A. H., and Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

Evans, S., Pearce, K., Vitak, J., and Treem, J. (2016). Explicating affordances: A conceptual framework for understanding affordances in communication research: Explicating affordances. *J. Comput. Mediat. Commun.* 22, 35–52. doi: 10.1111/jcc4.12180

Eysenbach, G., and Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 324, 573–577. doi: 10.1136/bmj.324.7337.573

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i3.96

Ghasemaghaei, M., and Hassanein, K. (2019). Dynamic model of online information quality perceptions and impacts: A literature review. *Behav. Inf. Technol.* 38, 302–317. doi: 10.1080/0144929X.2018.1531928

Griffin, R. J., Dunwoody, S., and Neuwirth, K. (1999). Proposed model of the relationship of risk information seeking and processing to the development of preventive behaviors. *Environ. Res.* 80, S230–S245. doi: 10.1006/enrs.1998.3940

Griffin, R. J., Neuwirth, K., Dunwoody, S., and Giese, J. (2004). Information sufficiency and risk communication. *Media Psychol.* 6, 23–61. doi: 10.1207/s1532785xmep0601_2

Haas, A., and Unkel, J. (2017). Ranking versus reputation: Perception and effects of search result credibility. *Behav. Inf. Technol.* 36, 1285–1298. doi: 10.1080/0144929X.2017.1381166

Häkkinen, P., Järvelä, S., Mäkitalo-Siegl, K., Ahonen, A., Näykki, P., and Valtonen, T. (2017). Preparing teacher-students for twenty-first-century learning practices (PREP 21): A framework for enhancing collaborative problem-solving and strategic learning skills. *Teach. Teach.* 23, 25–41. doi: 10.1080/13540602.2016.1203772

Hendriks, F., Kienhues, D., and Bromme, R. (2015). Measuring Laypeople's trust in experts in a digital age: The Muenster epistemic trustworthiness inventory (METI). *PLoS One* 10:e0139309. doi: 10.1371/journal.pone.0139309

Hendriks, F., Mayweg-Paus, E., Felton, M., Iordanou, K., Jucks, R., and Zimmermann, M. (2020). Constraints and affordances of online engagement with scientific information—A literature review. *Front. Psychol.* 11:572744. doi: 10.3389/fpsyg.2020.572744

Hendriks, F., Seifried, E., and Menz, C. (2021). Unraveling the "smart but evil" stereotype: Pre-service teachers' evaluations of educational psychology researchers versus teachers as sources of information. *Z. Für Pädagog. Psychol.* 35, 157–171. doi: 10.1024/1010-0652/a000300

Hinostroza, J. E., Ibieta, A., Labbé, C., and Soto, M. T. (2018). Browsing the internet to solve information problems: A study of students' search actions and behaviours using a 'think aloud' protocol. *Educ. Inf. Technol.* 23, 1933–1953. doi: 10.1007/s10639-018-9698-2

Hyldegård, J. (2006). Collaborative information behaviour—exploring Kuhlthau's information search process model in a group-based educational setting. *Form. Methods Inf. Retr.* 42, 276–298. doi: 10.1016/j.ipm.2004.06.013

Iding, M. K., Crosby, M. E., Auernheimer, B., and Barbara Klemm, E. (2009). Web site credibility: Why do people believe what they believe? *Instr. Sci.* 37, 43–63. doi: 10.1007/s11251-008-9080-7

Jiang, T., Liu, F., and Chi, Y. (2015). Online information encountering: Modeling the process and influencing factors. *J. Doc.* 71, 1135–1157. doi: 10.1108/JD-07-2014-0100

Joseph, P., Debowski, S., and Goldschmidt, P. (2013). Models of information search: A comparative analysis. *Inf. Res.* 18:562.

Kammerer, Y., and Gerjets, P. (2012). Effects of search interface and Internet-specific epistemic beliefs on source evaluations during web search for medical information: An eye-tracking study. *Behav. Inf. Technol.* 31, 83–97. doi: 10.1080/0144929X.2011.599040

Kammerer, Y., and Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *Int. J. Hum. Comput. Interact.* 30, 177–191. doi: 10.1080/10447318.2013.846790

Kammerer, Y., Nairn, R., Pirolli, P., and Chi, E. H. (2009). "Signpost from the Masses: Learning effects in an exploratory social tag search browser," in *Proceedings of the SIGCHI conference on human factors in computing systems CHI '09*, (New York, NY: Association for Computing Machinery), 625–634. doi: 10.1145/1518701.1518797

Yilmaz, K. F. G., and Çakmak, E. K. (2016). Internet-specific epistemological beliefs and online information searching strategies of pre-service teachers: Gender and department differences. *Particip. Educ. Res.* 3, 63–80. doi: 10.17275/per.16.10.3.2

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Z. Für. Pädagog. Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Able.

Kuhn, D. (1999). A developmental model of critical thinking. *Educ. Res.* 28, 16–46. doi: 10.3102/0013189X028002016

Meinert, J., and Krämer, N. C. (2022). How the expertise heuristic accelerates decision-making and credibility judgments in social media by means of effort reduction. *PLoS One* 17:e0264428. doi: 10.1371/journal.pone.0264428

Metzger, M. J., and Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *J. Pragmat.* 59, 210–220. doi: 10.1016/j.pragma.2013.07.012

Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. *Educ. Psychol.* 42, 173–190. doi: 10.1080/00461520701416306

Orlu, A. (2016). *Information seeking behaviour of masters students: Affective and behavioural dimensions*. Lincoln, NE: University of Nebraska-Lincoln.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., and Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *J. Comput. Mediat. Commun.* 12, 801–823. doi: 10.1111/j.1083-6101.2007.00351.x

Redecker, C., and Punie, Y. (2017). *European framework for the digital competence of educators: DigCompEdu*. Luxembourg: Publications Office of the European Union.

Rieh, S. Y., Collins-Thompson, K., Hansen, P., and Lee, H.-J. (2016). Towards searching as a learning process: A review of current perspectives and future directions. *J. Inf. Sci.* 42, 19–34. doi: 10.1177/0165551515615841

Rouet, J., and Britt, M. (2011). *Relevance processes in multiple document comprehension*. Available online at: /paper/Relevance-processes-in-multiple-document-Rouet-Britt/3915f5c4f3fd4b6d60f396410913ff846a66f05c (Accessed April 23, 2021).

Rousseau, D. M., and Gunia, B. C. (2016). Evidence-based practice: The psychology of EBP implementation. *Annu. Rev. Psychol.* 67, 667–692. doi: 10.1146/annurev-psych-122414-033336

Salehi, S., Du, J. T., and Ashman, H. (2018). Use of web search engines and personalisation in information searching for educational purposes. *Inf. Res. Int. Electron. J.* 23, 1–13.

Salmerón, L., Kammerer, Y., and García-Carrión, P. (2013). Searching the web for conflicting topics: Page and user factors. *Comput. Hum. Behav.* 29, 2161–2171. doi: 10.1016/j.chb.2013.04.034

Schemer, C., Matthes, J., and Wirth, W. (2008). Toward improving the validity and reliability of media information processing measures in surveys. *Commun. Methods Meas.* 2, 193–225. doi: 10.1080/19312450802310474

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *J. Educ. Psychol.* 82, 498–504. doi: 10.1037/0022-0663.82.3.498

Spencer, T. D., Detrich, R., and Slocum, T. A. (2012). Evidence-based practice: A framework for making effective decisions. *Educ. Treat. Child.* 35, 127–151. doi: 10.1353/etc.2012.0013

Stadtler, M., Winter, S., Scharrer, L., Thomm, E., Krämer, N., and Bromme, R. (2017). Selektion, integration und evaluation. *Psychol. Rundsch.* 68, 177–181. doi: 10.1026/0033-3042/a000361

StatCounter (2022). *Search engine market share worldwide*. Available online at: https://gs.statcounter.com/search-engine-market-share (accessed June 21, 2022).

Sundar, S. S. (2008). The main model: A heuristic approach to understanding technology effects on credibility. *Digit. Media* 28, 2–73.

Sung, Y.-T., Chang, K.-E., and Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Comput. Educ.* 94, 252–275. doi: 10.1016/j.compedu.2015.11.008

Thomm, E., Seifried, E., and Bauer, J. (2021b). Informing professional practice: (Future) teachers' choice, use and evaluation of (non-)scientific information sources about educational topics. *Z. Für. Pädagog. Psychol.* 35, 121–126. doi: 10.1024/1010-0652/a000309

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021a). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Z. Für. Pädagog. Psychol.* 35, 173–184. doi: 10.1024/1010-0652/a000301

Tsai, P.-S., Tsai, C.-C., and Hwang, G.-J. (2011). The correlates of Taiwan teachers' epistemological beliefs concerning Internet environments, online search strategies, and search outcomes. *Spec. Issue Internet Teach. Educ. Asian Exp.* 14, 54–63. doi: 10.1016/j.iheduc.2010.03.003

Williams, D., and Coles, L. (2007). Teachers' approaches to finding and using research evidence: An information literacy perspective. *Educ. Res.* 49, 185–206. doi: 10.1080/00131880701369719

Wilson, C., Grizzle, A., Tuazon, R., Akyempong, K., and Cheung, C. K. (2011). *Media and information literacy curriculum for teachers*. Paris: United Nations Educational, Scientific and Cultural Organization (UNESCO).

Wilson, T., Ford, N., Ellis, D., Foster, A., and Spink, A. (2002). Information seeking and mediated searching: Part 2. Uncertainty and its correlates. *J. Am. Soc. Inf. Sci. Technol.* 53, 704–715. doi: 10.1002/asi.10082

Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., et al. (2007). A process model of the formation of spatial presence experiences. *Media Psychol.* 9, 493–525. doi: 10.1080/15213260701283079

Yang, Z. J., Aloe, A. M., and Feeley, T. H. (2014). Risk information seeking and processing model: A meta-analysis. *J. Commun.* 64, 20–41. doi: 10.1111/jcom.12071

Zimmermann, M., and Mayweg-Paus, E. (2021). The role of collaborative argumentation in future teachers' selection of online information. *Z. Pädagog. Psychol. German J. Pedagog. Psychol.* 1–14. doi: 10.1024/1010-0652/a000307

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/feduc.2022.976791

Check for updates

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University, Germany

REVIEWED BY
Kris-Stephen Besa,
University of Münster, Germany
Michael Sailer,
Ludwig Maximilian University of Munich, Germany

*CORRESPONDENCE
Thamar Voss
thamar.voss@ezw.uni-freiburg.de

SPECIALTY SECTION
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

RECEIVED 23 June 2022
ACCEPTED 30 August 2022
PUBLISHED 23 September 2022

CITATION
Voss T (2022) Not useful to inform teaching practice? Student teachers hold skeptical beliefs about evidence from education science.
Front. Educ. 7:976791.
doi: 10.3389/feduc.2022.976791

Not useful to inform teaching practice? Student teachers hold skeptical beliefs about evidence from education science

Thamar Voss*

Department of Educational Science, University of Freiburg, Freiburg, Germany

A goal of teacher education is to promote evidence-based teaching. Teacher beliefs are assumed to act as facilitators or barriers to evidence-based thinking and practices. In three sub-studies with a total of N = 346 German student teachers, the extent of student teachers' beliefs about education science and their consequences and sources were investigated. First, the results of questionnaire data indicated that student teachers held skeptical beliefs about education science: On average, they perceived education science as less complex than their subject disciplines and as less important for successful teaching than their subject didactics. Additionally, they endorsed myths about learning and teaching. Second, the more skeptical the student teachers' beliefs, the lower their engagement in education science courses within teacher education. Third, hypotheses about potential sources of these skeptical beliefs were experimentally tested as starting points for changing beliefs. The results showed that the "soft" research methods typical of education science and a general tendency to perceive research findings as trivial (hindsight bias) might contribute to this devaluation. Furthermore, students studying the natural sciences and students with little experience with education science held more skeptical beliefs.

KEYWORDS

evidence-based practice, pre-service teachers', beliefs and assumptions, education science, misconceptions

Introduction

There are increasing demands for teaching to evolve into a more evidence-based profession (Slavin, 2002; Bauer and Prenzel, 2012; Ferguson, 2021). Such demands relate not only to educational policy but also to teachers' professional thinking and behavior (Davies, 1999; Bauer and Prenzel, 2012; Niemi, 2016). Many countries require in their teacher education standards that teachers should be able to plan and design lessons based on scientific evidence (e.g., see Bauer and Prenzel, 2012, for an overview). Orientation toward the best available knowledge is standard in other professions, such as medicine (Sackett et al., 1996; Helmsley-Brown and Sharp, 2003). However, teachers rarely draw on evidence when planning, analyzing, and reflecting on job-related situations (Dagenais et al., 2012; Patry, 2019). A central reason for this lack of evidence-based thinking lies in teachers' beliefs (e.g., Patry, 2019). There is evidence that people in general hold skeptical beliefs about disciplines such as psychology or education science (Lilienfeld, 2012). The content of such disciplines is commonly perceived as less complex than that of the natural sciences (Keil et al., 2010). In particular, the research methods in disciplines such as education science seem to be perceived as "soft" (Munro and Munro, 2014, p. 533), providing less valid results than "hard" (Munro and Munro, 2014, p. 533) methods in the natural sciences. Similarly, (prospective) teachers often belief that findings from education science are of little relevance to their teaching in the classroom (Broekkamp and Hout-Wolters, 2007). These beliefs contradict research findings highlighting the importance of teachers' knowledge of empirical evidence from education science – for example, knowledge about effective teaching methods or effective classroom management strategies (e.g., König and Pflanzl, 2016; Ulferts, 2019; Voss et al., 2022) – for their professional success.

Consequently, in a series of three studies, I investigated whether student teachers held skeptical beliefs about education science and examined the consequences and sources of such skeptical beliefs. First, I implemented questionnaires capturing the extent of student teachers' beliefs about the importance and complexity of education science compared to the subject disciplines and the subject didactics. Second, I examined the consequences of such skeptical beliefs for engagement in education science courses within the teacher education program by using correlational data. Third, I conducted experimental studies to test hypotheses about sources of student teachers' skeptical beliefs.

Beliefs about evidence from education science

Conceptualizations of teachers' professional knowledge typically distinguish three domains of teachers' professional knowledge (Shulman, 1987): content knowledge (Krauss et al., 2008), pedagogical content knowledge (Krauss et al., 2008), and pedagogical-psychological knowledge (Voss et al., 2011). In accordance with this topology, in many countries, teacher education curricula (e.g., the curricula of the German teacher education system) require pre-service teachers to take courses in three disciplines (Bauer and Prenzel, 2012), namely in the *subject disciplines* to acquire content knowledge (i.e., the knowledge about the subject matter they will be teaching), *subject didactics* to acquire pedagogical content knowledge (i.e., how to make the subject matter accessible to their future students), and *education science* to acquire knowledge about pedagogical-psychological phenomena relevant for students' learning in general (e.g., knowledge about learning strategies, students' motivation, assessment).

Prior research has indicated that beliefs about education science as a discipline within teacher education are rather skeptical (Cramer, 2013; Siegel and Daumiller, 2021). Beliefs are personal views about the self and the world that are thought to be true (Richardson, 1996). Beliefs are organized in a complex mental network and often termed subjective theories (Patry, 2019). Analogously to scientific theories, people use such subjective theories or beliefs to describe, explain, and predict phenomena, but such beliefs have a different epistemic status. While scientific theories are based on objectifiable, justifiable bodies of knowledge, beliefs are based on experience. Thus, beliefs do not meet the criterion of being objectifiable through scientific evidence (Richardson, 1996). Therefore, systems of beliefs typically contain elements that are not based on scientific evidence and not consistent with scientific theories. The skeptical beliefs about education science primarily manifest themselves with respect to two aspects.

First, teachers often doubt that empirical evidence from education science is important for teachers (Beycioglu et al., 2010; Dagenais et al., 2012; Cain, 2016; Thomm et al., 2021a). Many teachers assume that such evidence is not applicable to their practice (e.g., Merk et al., 2017; Joram et al., 2020; review by van Schaik et al., 2018) and perceive a gap between education science research and daily challenges in the classroom (Broekkamp and Hout-Wolters, 2007; Merk et al., 2017). There is also some evidence that student teachers perceive the education science even as less important than the subject disciplines (e.g., Cramer, 2013).

Second, research on epistemological beliefs has shown that student teachers often hold unsophisticated epistemological beliefs about education science topics (Guilfoyle et al., 2020; Moser et al., 2021). Epistemological beliefs are subjective theories about knowledge and knowing (Hofer, 2001). Existing conceptualizations identify various dimensions of epistemological beliefs. However, most conceptualizations include complexity of knowledge as one such dimension (e.g., Schommer, 1990; Hofer and Pintrich, 1997). Research results have indicated that student teachers hold rather unsophisticated beliefs (e.g., Brownlee et al., 2001) and, for instance, believe that education science content is not particularly complex (Lilienfeld, 2012). Even children rate psychological questions as easier to answer than questions from disciplines like chemistry or physics (Keil et al., 2010).

In addition to general beliefs about education science as a discipline of teacher education, student teachers often hold misconceptions about specific educational topics. Misconceptions are beliefs contradicted by established research findings in a discipline (Bensley and Lilienfeld, 2017).

Research has shown that teachers and student teachers often endorse misconceptions (e.g., Dekker et al., 2012; Bensley and Lilienfeld, 2017; Pieschl et al., 2021). Such misconceptions are: instruction needs to be adapted to specific learning styles (Macdonald et al., 2017; Eitel et al., 2021), it is primarily the teacher's personality that matters for teaching success (Darling-Hammond, 2006), having more experience automatically makes one a better teacher, smaller class sizes automatically lead to better student learning (Menz et al., 2021a). Existing research exposes these four statements as misconceptions. For instance, (1) there is no solid evidence that there is any benefit to adapting instruction to learning styles (e.g., Kirschner and van Merriënboer, 2013). Several studies have (2) shown that teachers' profession-specific competencies (rather than general personality traits) are important for teaching success (e.g., Kunter et al., 2013). Evidence also indicates that (3) teachers with more experience are not automatically better teachers, but that it depends on how teachers leverage their experiences (Friedrichsen et al., 2009; Kleickmann et al., 2013). Several studies (4) have indicated that a smaller class size is no guarantee for better learning (Hattie, 2009). Such misconceptions that receive widespread endorsement are also called *myths* (e.g., neuromyths, Dekker et al., 2012).

Consequences of student teachers' beliefs for their learning

It is assumed that beliefs have consequences for people's motivation and behavior (Fives and Buehl, 2012; Buehl and Beck, 2015), because they are thought to serve as a filter for interpreting new experiences (Pajares, 1992): Humans always perceive situations through the lens of their existing beliefs, which affect how they select and process information and how they make decisions in a given situation (Fives and Buehl, 2012; Patry, 2019).

This filtering effect of beliefs is assumed to be particularly important in the context of teaching and teacher education (Yadav et al., 2011; Brownlee et al., 2017). Beliefs are formed very early during our school careers (for an overview, see, for example, Pajares, 1992; for an empirical study, see, for example, Haney and McArthur, 2002). Accordingly, student teachers bring a set of fixed beliefs based on their experiences with them into their teacher education program. It is assumed that these beliefs shape what and how pre-service teachers learn during teacher education (Fives and Buehl, 2012; Stark, 2017; Ferguson, 2021). For instance, inadequate beliefs or misconceptions may lead to an oversimplification of complex information and result in poor learning outcomes (Schommer-Aikins, 2004; Moser et al., 2021). Accordingly, the results of interview studies with rather small samples (Holt-Reynolds, 1992; Bondy et al., 2007) have shown that students interpret situations in line with their beliefs. Furthermore, in a study with Norwegian student teachers, Bråten and Ferguson (2015) found evidence that more positive beliefs by student teachers about the importance of formalized sources of knowledge (such as research articles or textbooks) are associated with higher motivation to learn from formal teacher training courses (see also Chan, 2003 for a study with student teachers from Hong Kong and Siegel and Daumiller, 2021 for a mixed-method study with a relatively small sample).

Sources of student teachers' beliefs

Because beliefs about teaching and learning are based on years of one's own school experiences, these beliefs are thought to have multifarious sources (Lilienfeld, 2012).

(1) Disciplinary culture: People often assume that the impact of science on society strongly differs across disciplines (Janda et al., 1998; Richardson and Lacroix, 2021). Physics and mathematics typically have the highest prestige, whereas social sciences such as psychology, sociology, or education science have the lowest (e.g., Simonton, 2006; Klavans and Boyack, 2009). This low prestige may contribute to the devaluation of education science among student teachers. Furthermore, research on epistemological beliefs has revealed interindividual differences in students' beliefs by disciplinary culture: Students studying "soft" disciplines (e.g., psychology, education science) held more sophisticated epistemological beliefs than students of "hard" sciences (e.g., mathematics, physics, biology, Paulsen and Wells, 1998; Karimi, 2014). However, there are also contradictory research results. For instance, Rosman et al. (2020) found hardly any differences in epistemological beliefs between biology and psychology students.

(2) Experience with the academic discipline: Based on experiences from their own school days, student teachers enter teacher education with a fixed set of beliefs about teaching and learning (Richardson, 1996; Fives and Buehl, 2012). These beliefs are often at odds with the scientific theories taught at universities (Joram, 2007; Fives and Buehl, 2012). As they gain more experience with the academic discipline of education science, student teachers should develop a more appropriate representation of it. Accordingly, (limited) research results indicate that students with more experience (e.g., students with more courses in the discipline or students in a master's degree program vs. bachelor's degree program) hold more positive beliefs about the discipline than students with less experience (Bartels et al., 2009, for psychology students; Moser et al., 2021, for pre-service teachers).

(3) "Soft" research methods of the discipline: Many people have an unfavorable opinion of psychology's scientific quality (Lilienfeld, 2012). It is interesting to note that neuropsychological evidence is perceived more like "hard" sciences than the other "softer" subdisciplines of psychology (Keil et al., 2010). For instance, there is manifold evidence

on the *seductive allure effect* (Weisberg et al., 2008): People judge explanations of psychology findings as better when those explanations contain logically irrelevant neuroscience information (e.g., Weisberg et al., 2008; Hopkins et al., 2016). Consequently, distrust in the reliability of evidence from soft sciences such as education science may also be rooted in the typical research methods of education science as opposed to "harder" sciences such as chemistry, physics, or neuropsychology (Lilienfeld, 2012). For instance, Munro and Munro (2014) found evidence in a scenario-based approach that students evaluated the quality of evidence generated with brain magnetic resonance imaging (e.g., MRI) more favorably than evidence from cognitive tests.

(4) Preference for information from anecdotal sources: Drawing upon their own experiences in school, teachers often prefer anecdotal information from practitioners to inform their practice compared to evidence from scientific sources (e.g., Ferguson, 2020; Kiemer and Kollar, 2021). Research results have demonstrated that the preference for such non-scientific (i.e., anecdotal) sources contributes to shaping student teachers' misconceptions about topics from education science (Menz et al., 2021b).

(5) Hindsight bias: Many people believe that most knowledge from "soft" disciplines is obvious (Lilienfeld, 2012). This tendency to view outcomes as foreseeable once we know them is termed hindsight bias (the "I knew it all along" effect, Lilienfeld, 2012, p. 120). Research results have shown that this tendency is pronounced among human beings in general (e.g., in political elections, Blank et al., 2003), as well as among student teachers concerning topics from education science (Wong, 1995). Hindsight bias concerning evidence from education science may thus also contribute to student teachers' skeptical beliefs about education science.

Thus, there is evidence for potential sources of the skeptical beliefs, but studies with actual samples of student teachers that systematically explore these different sources are lacking.

The present study

Data come from three sub-studies from a research program investigating German secondary school student teachers' beliefs about education science. In Germany, secondary school teacher training programs are divided into a bachelor's degree program (six semesters) and a master's degree program (four semesters). Student teachers study at least two subjects and take courses in the subject didactics of these two subjects as well as in education science. Bachelor's degree programs at most universities have a clear focus on the two subject disciplines, with the most credits awarded in the subject disciplines, while master's degree programs have a stronger focus on education science and subject didactics. Although German universities are organized in a federal system with differences across the federal states, this overall structure is found in each state.

In the present study, I first compared student teachers' beliefs about education science with their beliefs about their subject disciplines and subject didactics. I assumed a devaluation of education science in terms of beliefs about the *importance* of education science for teaching and the *complexity* of education science (Sub-Study 1). Furthermore, I expected that student teachers, on average, would endorse myths about educational topics (Sub-Study 3). Second, consequences of these beliefs were investigated with the assumption that skeptical beliefs about education science would be associated with a lower *engagement* with research from education science and a lower *openness* to scientific evidence (Sub-Study 2). Third, possible sources of the devaluation of education science were examined. Specifically, I investigated (a) the impact of the *subjects* students were studying as an indicator of the disciplinary culture [i.e., natural science subjects (STEM) vs. other subjects; Sub-Study 1], (b) the impact of students' *level of experience* with education science (i.e., students in the bachelor's vs. master's degree program; Sub-Study 1). As further possible sources, I investigated in two experimental studies (c) whether student teachers tend to devaluate evidence from studies using *soft* research methods in comparison to *hard* research methods (Sub-Study 2), (d) whether student teachers prefer information from *anecdotal* (vs. *scientific*) sources (Sub-Study 2), and (e) whether student teachers tend to believe that evidence from education science is trivial (*hindsight bias*; Sub-Study 3).

Sub-Study 1

Hypotheses

I assumed that students would hold skeptical beliefs about the importance and complexity of education science. Furthermore, I expected moderating effects of the subject the student teachers were studying and their degree program (bachelor's versus master's level).

(1) Hypotheses on the importance of the disciplines for professional success:

- Student teachers believe that education science is less important for professional success than their subject disciplines and subject didactics (*importance devaluation hypothesis*).
- The tendency to devalue the importance of education science is more pronounced among student teachers studying a STEM subject than among students not studying a STEM subject (*importance-by-subject hypothesis*).
- The tendency to devalue the importance of education science is more pronounced among student teachers in the bachelor's degree program than among student teachers

in the master's degree program (*importance-by-degree hypothesis*).

(2) Hypotheses on the complexity of the disciplines:

- Student teachers evaluate education science as less complex than their subject disciplines and subject didactics (*complexity devaluation hypothesis*).
- Student teachers studying a STEM subject devalue the complexity of education science more strongly than student teachers not studying a STEM subject (*complexity-by-subject hypothesis*).
- The tendency to devalue the complexity of education science is more pronounced among student teachers in the bachelor's degree program than among student teachers in the master's degree program (*complexity-by-degree hypothesis*).

I conducted an *a priori* power analysis in G*Power (Faul et al., 2009) for analyses of variance with the between-subject factors degree program (bachelor's vs. master's degree) and subject (STEM vs. non-STEM), with beliefs as the within-subject factor (three levels: beliefs about education science, subject disciplines, and subject didactics; expected medium-sized intercorrelations), and their interaction ($\alpha = 0.05$, power $\beta = 0.80$). The results indicated that a sample size of $N = 36$ would be sufficient to detect the expected medium-sized main effects for the within-subject factor, and a sample size of $N = 206$ would be sufficient to detect the expected small interaction effects between the within-subject factor and the between-subject factors.

Materials and methods

Sample

A total of $N = 210$ student teachers from the University of Freiburg participated in Sub-Study 1 (**Table 1**). About 50% of the participants were enrolled in a master's degree program (i.e., *Master of Education*), the others were enrolled in a bachelor's degree program with the option to subsequently pursue a Master of Education. Among participants, 65% were studying at least one STEM subject.

Instruments

Beliefs about the disciplines

Student teachers answered questions about the importance of education science, the students' subject disciplines, and subject didactics for professional success and the complexity of the topics covered in these disciplines. The item wording was parallel for the three disciplines. Student teachers indicated their agreement with statements about the importance and complexity of each discipline on 6-point Likert scales ranging from 1 (=*completely disagree*) to 6 (=*completely agree*). An example item measuring beliefs about the importance of the discipline for professional success is: *Comprehensive knowledge of theories and concepts from education science/my subject disciplines/my subject didactics helps to cope with the daily challenges of being a teacher* (7 items for each discipline). An example item measuring beliefs about the complexity of the topics in the discipline is: *You have to think hard to understand the topics in education science/my subject disciplines/my subject didactics* (4 items for each discipline; see **Table 2** for descriptive statistics and **Table 3** for the intercorrelations among the scales).

Degree program and subjects

Additionally, student teachers indicated their degree program (master's or bachelor's degree program) and the subjects they were studying. The subjects were coded as STEM (students studying at least one STEM subject; i.e., mathematics, physics, chemistry, computer science, or geography) versus non-STEM (students studying two subjects in the linguistics, humanities, or social sciences).

Results

Do student teachers evaluate education science as less important than their subject disciplines and subject didactics?

I computed an analysis of variance with discipline as the within-subject factor (i.e., education science, subject disciplines, and subject didactics) and beliefs regarding their importance as the dependent variable. Furthermore, I included the students' subjects (STEM vs. non-STEM) and degree program (master vs. bachelor) as between-subject factors to investigate the moderator hypotheses. The results showed a significant large main effect of discipline, $F(2, 352) = 59.166$, $p = 0.000$, $\eta^2 = 0.25$ [controlling for gender with no significant effect: $F(1, 352) = 1.051$, $p = 0.351$, $\eta^2 = 0.00$]. To uncover the significant main effect of discipline, I computed a planned contrast following the *importance*

TABLE 1 Overview of the three sub-samples.

	Sub-Study 1	Sub-Study 2	Sub-Study 3
n	210	87	49
Age: *M (SD)*	22.74 (3.05)	24.52 (2.35)	–
Gender: % female	73	67	53
Subjects: % with at least one STEM	65	18	–
Degree program: % master's level	49	100	100
Semester (first subject): mode	2	2	1

TABLE 2 Descriptive statics for the instruments measuring student teachers' beliefs about the three disciplines (Sub-Study 1 and Sub-Study 2).

	Sub-Study 1				Sub-Study 2			
	No	*M*	*SD*	α	*No*	*M*	*SD*	α
Importance of the discipline for professional success								
Education science	7	3.80	0.97	0.87	7	4.11	1.14	0.94
Subject disciplines	7	3.87	0.97	0.81	7	3.65	0.82	0.74
Subject didactics	7	4.89	0.87	0.89	7	5.21	0.66	0.82
Complexity of the discipline								
Education science	4	3.20	0.99	0.76	4	3.17	0.99	0.82
Subject disciplines	4	4.85	0.94	0.84	4	4.84	0.90	0.77
Subject didactics	4	3.18	0.90	0.82	4	2.97	0.84	0.83

No = number of items per scale; *M* = mean; *SD* = standard deviation; α = Cronbach's alpha, ratings vary between 1 (=*completely disagree*) and 6 (=*completely agree*).

TABLE 3 Intercorrelations of the instruments (Sub-Study 1).

		1	2	3	4	5	6
1	Importance education science	1					
2	Importance subject disciplines	0.05	1				
3	Importance subject didactics	0.36*	0.16*	1			
4	Complexity education science	−0.01	0.02	−0.13	1		
5	Complexity subject disciplines	0.02	−0.24*	0.25*	−0.20*	1	
6	Complexity subject didactics	0.09	0.03	−0.16*	0.34*	0.02	1

*$p < 0.05$.

devaluation hypothesis that students rated education science as less important than the subject disciplines and subject didactics. Consequently, I specified the contrast with the weights: subject disciplines = +1, subject didactics = +1, education science = −2. The contrast was statistically significant, $F(1, 176) = 34.240$, $p = 0.000$, $\eta^2 = 0.16$. Thus, the results supported the *importance devaluation hypothesis*: Student teachers evaluated subject didactics and subject disciplines as more important for professional success than education science. Descriptively, student teachers rated their subject didactics as particularly important compared to the other two disciplines (Table 2).

The interaction with subject was not statistically significant [$F(2, 352) = 1.822$, $p = 0.163$, $\eta^2 = 0.01$]. Thus, the results did not support the *importance-by-subject hypothesis* that devaluation of education science is more pronounced among students studying a STEM subject.

In contrast, the interaction effect with degree program was statistically significant [$F(2, 352) = 5.173$, $p = 0.006$, $\eta^2 = 0.03$; small effect, Figure 1]. A simple effects analysis with Bonferroni correction of the alpha level revealed that student teachers at the master's level rated education science ($p = 0.001$) – and subject didactics ($p = 0.000$) – as significantly more important than students at the bachelor's level (Figure 1). In contrast, the difference for subject disciplines by degree program ($p = 0.992$) was not statistically significant. Hence, the results partially supported the *importance-by-degree hypothesis* that devaluation of education science is more pronounced among students at the bachelor's compared to master's level.

Do student teachers evaluate education science as less complex than their subject disciplines and subject didactics?

An analysis of variance with discipline as the within-subject factor (i.e., education science, subject disciplines, and subject didactics), subject and degree program as between-subject factors, and beliefs about complexity as the dependent variable revealed a significant, large main effect, $F(2, 348) = 150.982$, $p = 0.000$, $\eta^2 = 0.47$ [again controlling for gender with no significant effect: $F(1, 348) = 0.170$, $p = 0.8431$, $\eta^2 = 0.00$]. To test the *complexity devaluation hypothesis* that student teachers rate education science as less complex than their subject disciplines and subject didactics, I computed a planned contrast with the weights: subject disciplines = +1, subject didactics = +1, education science = −2. The contrast was statistically significant, $F(1, 174) = 78.269$, $p = 0.000$, $\eta^2 = 0.31$. Hence, in line with the hypothesis, student teachers evaluated the complexity of education science significantly lower than the complexity of their subject disciplines and subject didactics.

Furthermore, the results indicated a significant interaction effect for both moderators, that is, subject [$F(2, 348) = 5.366$, $p = 0.005$, $\eta^2 = 0.03$] and degree program [$F(2, 348) = 7.235$,

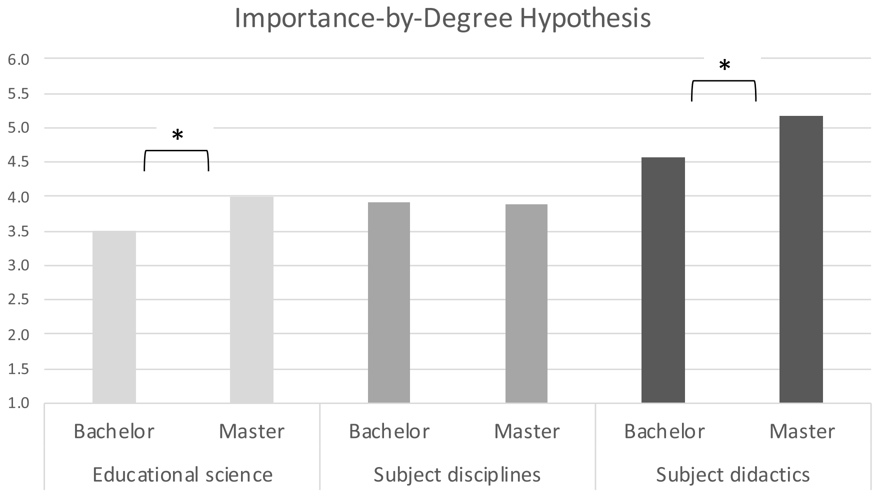


FIGURE 1
Interaction effects: Importance of the disciplines for professional success as a function of degree program (Sub-Study 1). The figure displays beliefs about the importance of the three disciplines for professional success among students at the bachelor's vs. master's level. *$p < 0.05$ in the simple effects analyses.

$p = 0.001$, $\eta^2 = 0.04$; **Figure 2**]. A simple effects analysis with Bonferroni correction of the alpha level revealed that students with and without a STEM subject differed significantly on all three variables: In line with the *complexity-by-subject hypothesis,* students with a STEM subject rated the complexity of education science significantly lower ($p = 0.021$), the complexity of their subject disciplines significantly higher ($p = 0.019$), and the complexity of subject didactics lower ($p = 0.028$) than students without a STEM subject. Regarding the *complexity-by-degree hypothesis*, the simple effects analysis indicated that students at the bachelor's level rated the complexity of their subject disciplines significantly lower ($p = 0.003$) and the complexity of subject didactics significantly higher ($p = 0.009$) than students at the master's level. Contradicting the hypothesis, students from the two programs did not differ significantly with regard to their ratings of the complexity of education science ($p = 0.160$).

Summary

In line with the assumptions, I found evidence for a devaluation of education science among student teachers: On average, they perceived education science as less important for professional success than subject didactics and as less complex than their subject disciplines. The results also yielded moderating effects: The tendency to devalue the importance of education science was more pronounced among students with less experience with education science (i.e., bachelor's degree students) than among students with more experience with education science (i.e., master's degree students). In addition, the tendency to devalue the complexity of education science was more pronounced among students with a STEM subject than among students without a STEM subject.

Sub-Study 2

Hypotheses

I investigated the consequences of the skeptical beliefs for engagement and possible sources (*soft* versus *hard* research methods and *anecdotal* versus *scientific* sources of evidence) with the following assumptions.

(1) I hypothesized that more skeptical beliefs about education science would be associated with lower engagement with research from education science:

- Students with more negative beliefs are less willing to exert effort in educational science courses (*beliefs engagement hypothesis*).
- Students with more negative beliefs are less open to evidence-based practices (*beliefs openness hypothesis*).

(2) As potential sources for the devaluation of education science, I expected:

- Students consider research findings from studies using *soft* research methods (typical methods from education science, such as surveys, systematic observations, standardized tests) to be less trustworthy than findings from studies using *hard* research methods (typical research methods

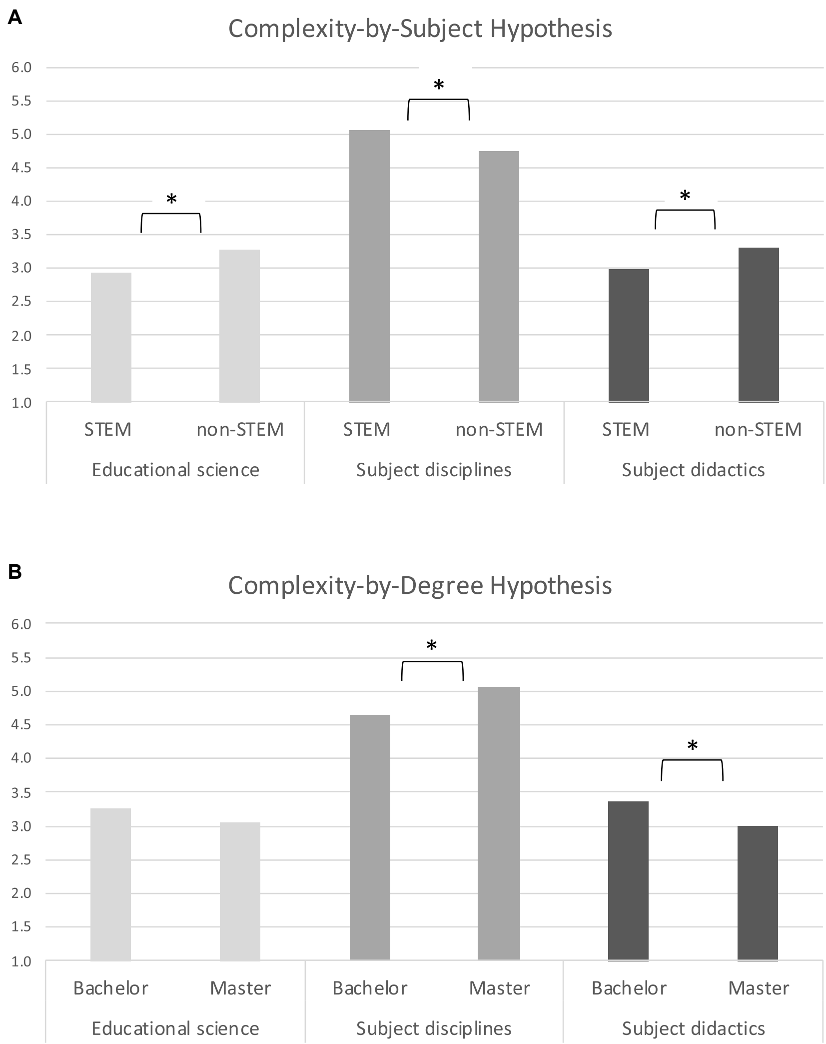


FIGURE 2
Interaction effects: Complexity of the disciplines as a function of subject and degree program (Sub-Stud 1). The figure displays beliefs about the complexity of the three disciplines among students with a STEM subject (STEM) versus no STEM subjects (non-STEM; **A**) and for students at the bachelor's versus master's level **(B)**. $^*p < 0.05$ in the simple effects analyses.

from the natural sciences, such as EEG, fMRI; *research method hypothesis*).

- Student teachers perceive information reported by colleagues (i.e., anecdotal source) as more trustworthy than information from empirical educational research (scientific source; *source of evidence hypothesis*).

An *a priori* power analysis in G*Power for a linear multiple regression analysis to test the hypotheses on the engagement (α = 0.05, power β = 0.80, 9 predictors) indicated that a sample size of N = 74 would be sufficient to detect a medium-sized effect. For the hypotheses on the sources, I conducted two experimental manipulations (*soft* versus *hard* research methods; *anecdotal* versus *scientific* source). The power analysis for an ANOVA with a within-subject factor with two levels indicated that a sample size of N = 46 would be sufficient to detect a medium-sized effect.

Materials and methods

Sample

In total, N = 87 student teachers participated in Sub-Study 2 (**Table 1**). All participants were in the first to fourth semester of the Master of Education.

Instruments

Beliefs about the disciplines

Student teachers completed the same instrument as in Sub-Study 1 on their beliefs about the importance of the three disciplines for professional success and the complexity of the disciplines (**Table 2**).

Engagement with education science

Two aspects of engagement with education science were measured using 15 6-point Likert scale items (*completely disagree* to *completely agree*). First, eight items captured student teachers' *willingness to make an effort* in education science (Cronbach's α = 0.88; adapted from Jonkmann et al., 2013). An example item is: *I do my best in education science courses.* Second, seven items were adapted from Aarons (2004) to measure *openness to evidence-based practices* (Cronbach's α = 0.70). An example item is: *I would use new methods that have been proven effective in research, even if they were very different from what I am used to doing.*

Vignettes about research findings

A total of 11 short vignettes about research findings were constructed. For each vignette, participants indicated how trustworthy the research findings were (6-point Likert scales, *very trustworthy* to *not at all trustworthy*). They were instructed that trustworthiness means the extent to which they would rely on and trust these findings. The vignettes were included in two versions of the questionnaire with two experimental within-subject variations:

The (1) *research method* was experimentally varied for five research findings. Each finding existed in two versions: one based on a typical research method from neuroscience (research method = hard; e.g., *the recording of brain activity/results from EEG/results from fMRI showed...*) and one based on a typical research method from education science (research method = soft, e.g., *the results of a standardized survey/of a systematic observation/of a standardized test showed...*). The research findings were counterbalanced across the two questionnaire versions.

The (2) *source of evidence* was experimentally varied for the other six research findings: Again, each finding existed in two versions that were identical except that one was based on a report by a colleague (source = anecdotal: e.g., *In my daily school life, I often observe that...*) and one on a scientific source (source = science; e.g., *Research results have shown that...*). The vignettes were also counterbalanced across questionnaire versions, and each participant rated three findings from an anecdotal source and three findings from a scientific source.

Other than these two experimental variations (source and research method), the vignettes were parallelized in terms of content (e.g., testing effect, self-regulated learning, homework), length (word count), and readability (Flesch, 1948). Neither the vignettes experimentally varying the research method nor those varying the source of information differed significantly from one another in terms of length and readability (**Table 4**).

Results

Are beliefs about education science related to engagement with education science?

I conducted a regression analysis of *willingness to make an effort* in education science on beliefs about the importance and complexity of education science, the students' subject disciplines, and subject didactics (controlling for gender, subject, and additionally for university, because students in this sub-study studied either at the University of Freiburg or the University of Education in Freiburg). Importance of education science and complexity of education science (and subject) were significant predictors of willingness to make an effort, whereas beliefs about the importance and complexity of the students' subject disciplines and subject didactics did not explain differences in willingness to make an effort in education science courses (**Table 5**). Thus, in line with the *beliefs engagement hypothesis,* stronger beliefs that education science is important for professional success and that education science is complex were associated with students being more willing to make an effort in their education science courses.

An analogously computed regression analysis of *openness to evidence-based practices* on beliefs about the importance and complexity of education science, subject disciplines, and subject didactics (also controlling for gender, subject, and university) showed that importance was a significant predictor, but complexity was not. Thus, partly in line with the hypothesis, more strongly believing that education science is important was associated with students being more open to evidence-based practice.

Do soft research methods and students' preference for anecdotal information contribute to the devaluation of education science?

The analysis of variance with research method as the within-subject factor (i.e., soft vs. hard), trustworthiness of the findings as the dependent variable (and gender, subject, and university as control variables) revealed a significant small to medium-sized main effect of research method [$F(1, 69) = 7.127$, $p = 0.009$, $\eta^2 = 0.09$]. This main effect supported the *research method hypothesis*: Student teachers rated findings obtained with hard research methods as more trustworthy than findings obtained with soft research methods. With the exception of university [$F(1, 69) = 4.982$, $p = 0.029$, $\eta^2 = 0.06$], none of the covariates showed a significant effect.

An analogous analysis of variance with source of evidence as the within-subject factor (i.e., anecdotal vs. scientific) also revealed a significant medium-sized main effect of source [$F(1, 68) = 4.324$, $p = 0.041$, $\eta^2 = 0.06$, no significant effects of any of the covariates]. This result indicated that – contradicting the *source of evidence hypothesis* – student teachers rated findings

TABLE 4 Means, standard deviations, and one-way analyses of variance in length and readability of the vignettes (Sub-Study 2 and Sub-Study 3).

Research method (Sub-Study 2)							
	Soft		**Hard**		***F*(1, 8)**	***p***	***d***
	M	***SD***	***M***	***SD***			
Word count	46.67	16.37	48.83	16.63	0.05	0.825	0.13
Flesch index	11.67	10.65	11.33	12.31	1.73	0.096	−0.03
Source of information (Sub-Study 2)							
	Anecdotal		**Science**		***F*(1, 10)**	***p***	***d***
	M	***SD***	***M***	***SD***			
Word count	71.67	14.71	70.50	17.85	0.02	0.904	−0.07
Flesch index	26.83	5.64	27.33	5.54	0.02	0.880	0.09
Correctness of research findings (Sub-Study 3)							
	Correct		**Incorrect**		***F*(1, 10)**	***p***	***d***
	M	***SD***	***M***	***SD***			
Word count	49.50	11.83	50.50	11.31	0.02	0.884	0.09
Flesch index	15.17	10.30	17.67	9.29	0.19	0.668	0.25
Myths (Sub-Study 3)							
	Correct		**Incorrect**		***F*(1, 6)**	***p***	***d***
	M	***SD***	***M***	***SD***			
Word count	33.25	9.84	32.00	5.72	0.05	0.833	−0.15
Flesch index	27.75	10.87	26.75	17.80	0.00	0.927	−0.0.7

from scientific sources as more trustworthy than findings from anecdotal sources.

Summary

Regarding the consequences of student teachers' beliefs, in line with the hypothesis, the results indicated that more skeptical beliefs about education science were related to lower engagement with research from education science and – at least for skeptical beliefs about the importance of education science – to lower openness to scientific evidence.

Furthermore, regarding possible sources of these skeptical beliefs, the experimental manipulation results indicated that student teachers, on average, evaluated empirical findings from studies with soft research methods as less trustworthy than equivalent empirical findings from studies with hard research methods – but with a rather small effect size. Thus, education science's typical soft research methods might contribute to its devaluation. Against the expectation, findings from anecdotal sources were given lower trustworthiness ratings than equivalent findings from scientific sources. However, this effect was also small.

Sub-Study 3

Research questions

Do (1) student teachers tend to perceive evidence from education science as trivial and (2) do they believe in myths regarding education science?

Materials and methods

Sample

A total of N = 49 student teachers participated in Sub-Study 3 (Table 1). All participants were in the first semester of their Master of Education program at the University of Freiburg.

Instruments: Vignettes about research findings and myths

The participants were also presented with short vignettes about research findings from education science, again with an experimental manipulation, but this time regarding the correctness of the research findings: Each of the six findings overall was presented in two versions, one as actually found in the research (e.g., *research findings showed that performance improved*), and one not in line with the research results (e.g., *research findings showed that performance did not improve*). The correctness of the findings was also counterbalanced across questionnaires and randomly assigned to subjects. Participants had to indicate on a 6-point Likert scale how obvious these findings were to them (*very obvious* to *not at all obvious*). They were instructed to rate whether the findings were expectable and not surprising (i.e., obvious) or surprising and contradicted what they would have expected (i.e., not obvious).

Furthermore, four similar vignettes on typical misconceptions (i.e., myths) about phenomena from education science were developed in two versions (incorrect = myth vs. correct) and randomly assigned to the students. The four myths were (1) the need to adapt instruction to students' learning styles (need to adapt vs. do not need to adapt), (2) impact of teacher personality (it is primarily a teacher's personality that matters for teaching success vs. a teacher's personality is not the primary factor for teaching success), (3) impact of teaching experience (having more experience automatically makes one a better teacher vs. does not automatically make one a better teacher), (4) impact of class size on student learning (smaller class sizes automatically lead to better student learning vs. do not automatically lead to better learning). These vignettes were also parallelized and did not significantly differ in terms of length or readability (Flesch, 1948; **Table 4**).

TABLE 5 Results of the regression analysis predicting engagement with education science (Sub-Study 2).

Predictors	Willingness to make an effort β	Openness β
Importance of the discipline for professional success		
Education science	0.34**	0.56**
Subject disciplines	−0.11	0.05
Subject didactics	0.01	−0.07
Complexity of the discipline		
Education science	0.36*	−0.12
Subject disciplines	0.02	−0.17
Subject didactics	−0.07	0.18
Control variables		
University[a]	−0.08	0.09
Subject[b]	0.26*	0.06
Gender[c]	−0.15	−0.04
R^2	0.36	0.22

$N = 76$; $^*p < 0.05$, $^{**}p < 0.01$.
[a]1 = University of Freiburg, 2 = University of Education in Freiburg. [b]1 = at least one STEM subject, 2 = no STEM subjects. [c]1 = female, 2 = male.

Results

The *a priori* power analysis for an analysis of variance with a within-subject factor with two levels (incorrect vs. correct) indicated that a sample size of $N = 46$ would be sufficient to detect a medium-sized effect.

Do student teachers perceive evidence from education science as trivial?

In the analysis of variance with correctness as the within-subject factor (i.e., correct vs. incorrect) and obviousness as the dependent variable (controlling for gender), the main effect was not significant [$F(1, 43) = 1.918$, $p = 0.173$, $\eta^2 = 0.04$]. The means were above the theoretical midpoint of 3.5 (**Table 6**), indicating that, in line with the *hindsight assumption*, students on average tended to evaluate findings from education science as rather obvious – independent of whether the findings were correct or incorrect. Descriptively, the mean was higher for the incorrect than the correct research findings (**Table 6**).

Do student teachers endorse myths about education science?

The analogously computed analysis of variance for myths with the within-subject factor correctness and obviousness as the dependent variable (controlling for gender) revealed a large main effect for correctness [$F(1, 43) = 40.434$, $p = 0.000$, $\eta^2 = 0.49$]. This result suggested that students strongly believe

TABLE 6 Means and standard deviations of vignettes (Sub-Studies 2 and 3).

	M	*SD*
Research method (Sub-Study 2)[a]		
Soft	4.34	0.93
Hard	4.47	0.87
Source of evidence (Sub-Study 2)[a]		
Anecdotal	4.05	0.89
Scientific	4.41	0.72
Research findings (hindsight bias, Sub-Study 3)[b]		
Correct	3.89	0.89
Incorrect	4.09	0.95
Myths (Sub-Study 2)[b]		
Correct	3.09	1.21
Incorrect	5.01	1.38

[a]$N = 77$, dependent variable = trustworthiness of the finding. [b]$N = 47$, dependent variable = obviousness of the finding.

in the myths: They rated the incorrect findings (i.e., the myths) as much more obvious and expected than the correct findings (see **Table 6** for descriptive statistics).

Summary

The results of Sub-Study 3 suggest that, consistent with the hindsight bias, student teachers retrospectively evaluated research findings from education science as trivial ("I knew it all along"). Furthermore, on average, student teachers strongly believe in myths about learning and teaching.

Overall discussion

I examined whether student teachers devaluate education science compared to their subject disciplines and subject didactics. Additionally, I investigated the consequences and potential sources of the devaluation.

Do student teachers hold skeptical beliefs about education science?

The results of three sub-studies indicated a pronounced devaluation of education science among student teachers in a German sample. I found evidence for this devaluation based on different research approaches (quasi-experimental questionnaire data and data from experimental studies) and reflecting several aspects.

First, in the questionnaire Sub-Study 1 with a large sample, student teachers perceive education science as *less complex* than their subject disciplines on average.

Second, student teachers perceive education science as *less important for teaching success* than subject didactics on average.

These skeptical beliefs about the importance of education science for teaching success are not in line with the empirical evidence: Research results indicate that teachers' knowledge about topics from education science is related to teaching success in terms of instructional quality and achievement: Higher pedagogical-psychological knowledge of teachers is associated with higher learning support of the students (Voss et al., 2014, 2022), a more efficient classroom management (Voss et al., 2014, 2022), and higher students' achievement (König and Pflanzl, 2016). This importance of teacher knowledge is not limited to pedagogical-psychological knowledge. It has also been shown that pedagogical content knowledge is related to the teaching success (e.g., Baumert et al., 2010), and content knowledge has been shown to be an essential basis for developing pedagogical content knowledge (e.g., Friedrichsen et al., 2009; Kleickmann et al., 2017). Thus, the three domains of teacher knowledge are important, pedagogical-psychological knowledge, pedagogical content knowledge, and content knowledge. This is in line with the national standards or guidelines for teacher education of many countries. Such standards describe what teachers should know and be able to do and typically cover aspects of all three domains of teacher knowledge (Bauer and Prenzel, 2012). Consequently, in Germany, student teachers must take courses in the three disciplines – subject disciplines, subject didactics, and education science – to acquire pedagogical-psychological knowledge, pedagogical content knowledge, and content knowledge. Thus, the devaluation of education science's importance for teaching found in the present study on average among student teachers is not consistent with either education policy standards or empirical research. Therefore, it seems necessary to target such inappropriate beliefs for change during teacher education. The devaluation tendency is also mirrored in motivational constructs, as research, for instance, indicated that the subject interest of teachers is higher than the interest in education science (e.g., Pozas and Letzel, 2021).

Third, in the experimental Sub-Study 3, student teachers turned out to strongly believe in myths about learning. Although plenty of research debunks the myths as myths (e.g., Kirschner and van Merriënboer, 2013; Macdonald et al., 2017; Eitel et al., 2021), the student teachers in the present sample believe in these myths. Thus, the results suggest a need to break down misconceptions of student teachers (Menz et al., 2021a; Prinz et al., 2021). In the present study, four such myths were examined as examples. However, other myths exist, such as that some students are information-savvy digital natives and that learners can multitask (Kirschner and De Bruyckere, 2017). Future research should address such other myths.

Are skeptical beliefs related to engagement in education science courses?

The devaluation of education science compared to subject disciplines and subject didactics is especially relevant in light of the assumed filtering effect of beliefs (e.g., Pajares, 1992; Patry, 2019) and thus the assumed importance of beliefs for future teachers' professional thinking and learning (e.g., Fives and Buehl, 2012). The results of the present study indicate that these beliefs matter for student teachers' motivation: More skeptical beliefs about the complexity and importance of education science were associated with lower engagement with research from education science and less openness to scientific evidence (the latter statistically significant only for skeptical beliefs about the importance of education science). Thus, the results suggest that the devaluation of education science is crucial, as it is related to the quality of students' uptake of learning opportunities. This result is in line with the assumed importance of beliefs for learning and the uptake of learning opportunities during teacher education (Fives and Buehl, 2012; Stark, 2017; Ferguson, 2021). The questionnaire data of the present study thus complement the results of smaller interview studies (e.g., Holt-Reynolds, 1992; Bondy et al., 2007). However, further research is needed to examine other indicators of uptake of learning opportunities in teacher training. In the present study, self-report data on motivation and openness were used. An important next step would be to investigate the associations of student teachers' beliefs with alternative measures of motivation (Fulmer and Frijters, 2009). Possible alternative approaches include the use of observational data (e.g., tasks chosen or persistence when engaging in tasks) or analyses of students' authentic learning materials (e.g., learning protocols, lesson plans).

Furthermore, more research is needed to examine the assumed detrimental effect of skeptical beliefs about evidence from education science on teaching success as prior research is ambiguous. For instance, some studies found that teachers with more skeptical beliefs about evidence-based practices *do not* differ in the frequency of the use of evidence-based practices from teachers with less skeptical beliefs (e.g., McNeill, 2019, see also Krammer et al., 2021, for believing in neuromyths). At the same time, other studies showed that positive beliefs about evidence *are* related to more frequent use of evidence-based practices (Combes et al., 2016). Additionally, evidence on beliefs, in general, indicated congruencies between teacher beliefs and teaching practices (e.g., overview from Buehl and Beck, 2015).

What are potential sources of skeptical beliefs about education science?

Knowledge of the sources of student teachers' skeptical beliefs may serve as starting points for breaking down dysfunctional beliefs and misconceptions.

Therefore, first, in the questionnaire Sub-Study 1, I investigated whether the devaluation of education science depends on students' selected subject disciplines and degree programs as potential sources of skeptical beliefs about education science. The results indicated that the tendency to devaluate the complexity of education science was more pronounced among students of STEM subjects than students with no STEM subjects. Thus, disciplinary culture obviously plays a role in shaping the tendency to devaluate the complexity of education science compared to students' subject disciplines. Furthermore, the devaluation of the importance of education science compared to subject didactics was moderated by student teachers' experience with education science: Student teachers at the bachelor's level devaluate the importance of education science for teaching success more strongly than student teachers at the master's level. The participants of the present study were student teachers from Freiburg University. During the bachelor's degree program at Freiburg University, student teachers have to complete only one module on education science, whereas in the master's degree program, significantly more credit hours are devoted to education science. Thus, on average, bachelor's students have less experiences with education science. The results of the present study suggest they are more prone to dysfunctional beliefs about education science, whereas student teachers with more experience (i.e., in the master's degree program) appear less prone to such devaluations. This might be because students in the bachelor's degree program have little knowledge about education science as a professional discipline and thus might lack the awareness of the importance of the discipline. This explanation would also have parallels to the Dunning-Kruger effect (Kruger and Dunning, 1999), a prominent effect in metacognitive research indicating that people with little knowledge in a domain tend to be unaware of their deficient knowledge. As a consequence of the moderating effect of experience, it seems vital to create learning opportunities early in teacher training programs that support students in reflecting on their skeptical beliefs about education science and forming a more appropriate conception of the discipline and its importance for teaching success. In light of research on typical gender differences (e.g., women are less likely to choose STEM subjects than men; Roloff Henoch et al., 2015), it is interesting to note that I found no effect of the covariate gender.

Second, experimental evidence indicated in the present study a dysfunctional pattern in the reception of research findings from education science. This also sheds light on potential sources for the devaluation of education science. The results of the experimental Sub-Study 2 indicate that the *soft research methods* typical of education science might contribute to its devaluation: Student teachers on average evaluated empirical findings from studies with soft research methods as less trustworthy than equivalent empirical findings from studies with hard research methods. Thus, student teachers need more knowledge about research methods and their validity (Voss et al., 2020; Thomm et al., 2021b) to reduce this potentially biased perception of the quality of research findings based on different methods.

Contrary to the assumption, findings from *anecdotal sources* were given lower trustworthiness ratings than equivalent findings from scientific sources. Prior research indicates that student teachers prefer teachers as sources of information compared to researchers (Menz et al., 2021b), have more positive beliefs about the utility of anecdotal information compared to educational research, and use anecdotal sources more frequently than scientific evidence (Kiemer and Kollar, 2021). Together with the results of the present study, this may indicate that even though student teachers evaluate information from scientific sources as more trustworthy, they use scientific sources less frequently than anecdotal sources. Additionally, Hendriks et al. (2021) found that the perceived trustworthiness of teachers and researchers depends on students' specific epistemic goal. Further research should shed light on this apparent contradiction by investigating beliefs about different sources together with concrete use of these sources in teaching and students' goals.

Finally, student teachers in the experimental Sub-Study 3 tended to believe that evidence from education science is trivial. This hindsight bias is a general phenomenon that has also been investigated in areas other than education (e.g., Blank et al., 2003). I found the tendency to perceive evidence about learning and teaching as common sense and foreseeable in the study among student teachers. Thus, this tendency may also contribute to the evolution of skeptical beliefs about education science.

Strengths and limitations

A main strength of this study was the combination of a quasi-experimental sub-study with experimental sub-studies. In doing so, evidence to *describe* phenomena was generated, such as the average level of student teachers' beliefs about education science compared to their beliefs about their subject disciplines and subject didactics. In addition, evidence to *explain* phenomena was generated in the experimental sub-studies, such as why student teachers perceive research findings from education science as less trustworthy than research findings from STEM subjects.

The samples of the three sub-studies consisted of a total of 346 student teachers from two universities in Germany. They studied different subject disciplines and were enrolled in different degree programs (bachelor's and master's). These results might be generalized to countries with similar conditions (e.g., culture, teacher education system). However, the results cannot be generalized to differently structured teacher education systems. As the *a priori* power analyses indicated, the sample sizes were sufficient to detect the expected effects. Nevertheless, larger samples would be desirable in future research to examine the importance of further moderators, such as the amount of student teachers' teaching experience, the length of time they have been in a teacher education program, or the prior knowledge about the disciplines.

Another limitation is the use of cross-sectional data. In the present study, the degree programs (bachelor's and master's) served as a proxy for experiences with education science. However, longitudinal data would be necessary to draw conclusions about the development of student teachers' beliefs over the course of teacher training. For example, a longitudinal study with student teachers in Germany found a decline in beliefs about the importance of education science over time on average (Cramer, 2013). As these results contradict the results of the present cross-sectional study, with more positive beliefs among students with more experience with education science, further research is needed to elucidate this contradiction.

Another limitation is how the consequences of skeptical beliefs were measured in Sub-Study 2, as the participants provided self-reports on engagement with education science. The self-report instrument was adapted from validated instruments (e.g., Aarons, 2004; Jonkmann et al., 2013). However, further studies with alternative measures, such as observational data or analyses of authentic learning materials, are needed.

Furthermore, the experimental sub-studies provided evidence for potential sources of skeptical beliefs about education science. For instance, the results indicate that student teachers rate research findings based on hard research methods as more trustworthy than equivalent research findings based on soft research methods like those typical of education science. However, as a further limitation, the direct impact of these sources on the formation of student teachers' beliefs was not examined and should be addressed in future research. Additionally, future research should investigate other potential influencing factors. For instance, student teachers' professional roles might also affect their beliefs about education science. Similar to research on motives for teaching (e.g., Watt et al., 2012), some student teachers might see themselves primarily as experts on their subject matter. Those student teachers might be highly interested in the subject and pass on the subject matter to the students. On the contrary, for other student teachers, educating students might be much more part of their professional role. As a result, those students might be more interested in education science and might also have more positive beliefs about the discipline.

Conclusion

Overall, the findings of the sub-studies indicate that student teachers, on average, held skeptical beliefs about education science. This poses a challenge to those involved in teacher education: Student teachers bring inappropriate beliefs and misconceptions into their teacher education program, and these beliefs are related to their engagement with the learning content. The findings support the assumption that beliefs are facilitators or barriers to the use of evidence in instructional situations (Fischer, 2021). Thus, it seems important to address beliefs early in teacher education programs (Stark, 2017), encourage students to reflect on their beliefs, and create specific learning opportunities to break down misconceptions and inappropriate beliefs.

The results of the present study on the potential sources of student teachers' skeptical beliefs can provide information on where to start. They should be considered alongside theoretical models (e.g., Gregoire, 2003) and evidence on how to successfully change beliefs (Gill et al., 2004; Kleickmann et al., 2016; Prinz et al., 2021).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

TV contributed to resource, conceptualization, methodology, analysis, and writing.

Acknowledgments

Many thanks go to Jörg Wittwer for his feedback and to Keri Hartman for proofreading the article. Furthermore, the author acknowledge support by the Open Access Publication Fund of the University of Freiburg.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aarons, G. A. (2004). Mental health provider attitudes toward adoption of evidence-based practice: the evidence-based practice attitude scale (EBPAS). *Ment. Health Serv. Res.* 6, 61–74. doi: 10.1023/B:MHSR.0000024351.12294.65

Bartels, J. M., Hinds, R. M., Glass, L. A., and Ryan, J. J. (2009). Perceptions of psychology as a science among university students: the influence of psychology courses and major of study. *Psychol. Rep.* 105, 383–388. doi: 10.2466/PR0.105.2.383-388

Bauer, J., and Prenzel, M. (2012). European teacher training reforms. *Science* 336, 1642–1643. doi: 10.1126/science.1218387

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *Am. Educ. Res. J.* 47, 133–180. doi: 10.3102/0002831209345157

Bensley, D. A., and Lilienfeld, S. O. (2017). Psychological misconceptions: recent scientific advances and unresolved issues. *Curr. Dir. Psychol. Sci.* 26, 377–382. doi: 10.1177/0963721417699026

Beycioglu, K., Ozer, N., and Ugurlu, C. T. (2010). Teachers' views on educational research. *Teach. Teach. Educ.* 26, 1088–1093. doi: 10.1016/j.tate.2009.11.004

Blank, H., Fischer, V., and Erdfelder, E. (2003). Hindsight bias in political elections. *Memory* 11, 491–504. doi: 10.1080/09658210244000513

Bondy, E., Ross, D., Adams, A., Nowak, R., Brownell, M., Hoppey, D., et al. (2007). Personal epistemologies and learning to teach. *Teach. Educ. Spec. Educ.* 30, 67–82. doi: 10.1177/088840640703000202

Bråten, I., and Ferguson, L. E. (2015). Beliefs about sources of knowledge predict motivation for learning in teacher education. *Teach. Teach. Educ.* 50, 13–23. doi: 10.1016/j.tate.2015.04.003

Broekkamp, H., and Hout-Wolters, B. v (2007). The gap between educational research and practice: a literature review, symposium, and questionnaire. *Educ. Res. Eval.* 13, 203–220. doi: 10.1080/13803610701626127

Brownlee, J. L., Ferguson, L. E., and Ryan, M. (2017). Changing teachers' epistemic cognition: a new conceptual framework for epistemic reflexivity. *Educ. Psychol.* 521, 242–252. doi: 10.1080/00461520.2017.1333430

Brownlee, J., Purdie, N., and Boulton-Lewis, G. (2001). Changing epistemological beliefs in pre-service teacher education students. *Teach. High. Educ.* 6, 247–268. doi: 10.1080/13562510120045221

Buehl, M. M., and Beck, J. S. (2015). "The relationship between teachers' beliefs and practices," in *International Handbook of Research on Teachers' Beliefs*, eds H. Fives and M. G. Gill (London: Routledge), 66–84.

Cain, T. (2016). Research utilisation and the struggle for the teacher's soul: a narrative review. *Eur. J. Teach. Educ.* 39, 616–629. doi: 10.1080/02619768.2016.1252912

Chan, K. (2003). Hong Kong teacher education students' epistemological beliefs and approaches to learning. *Res. Educ.* 69, 36–50. doi: 10.7227/RIE.69.4

Combes, B. H., Chang, M., Austin, J. E., and Hayes, D. (2016). The use of evidenced-based practices in the provision of social skills training for students with autism spectrum disorder among school psychologists. *Psychol. Sch.* 53, 548–563. doi: 10.1002/pits.21923

Cramer, C. (2013). Beurteilung des bildungswissenschaftlichen Studiums durch Lehramtsstudierende in der ersten Ausbildungsphase im Längsschnitt [Longitudinal assessment of educational science by student teachers in the first phase of their teacher education program]. *Z. Pädagog.* 59, 66–82. doi: 10.25656/01:11927

Dagenais, C., Lysenko, L., Abrami, P. C., Bernard, R. M., Ramde, J., and Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: a review of empirical research. *Evid. Policy* 8, 285–309. doi: 10.1332/174426412X654031

Darling-Hammond, L. (2006). *Powerful Teacher Education*. San Francisco, CA: Jossey-Bass.

Davies, P. (1999). What is evidence-based education? *Br. J. Educ. Stud.* 47, 108–121. doi: 10.1111/1467-8527.00106

Dekker, S., Lee, N. C., Howard-Jones, P., and Jolles, J. (2012). Neuromyths in education: prevalence and predictors of misconceptions among teachers. *Front. Psychol.* 3:429. doi: 10.3389/fpsyg.2012.00429

Eitel, A., Prinz, A., Kollmer, J., Niessen, L., Russow, J., Ludäscher, M., et al. (2021). The misconceptions about multimedia learning questionnaire: an empirical evaluation study with teachers and student teachers. *Psychol. Learn. Teach.* 20, 420–444. doi: 10.1177/14757257211028723

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/brm.41.4.1149

Ferguson, L. E. (2020). *Teacher Beliefs in Higher Education.* Norway: Læringsfestivalen.

Ferguson, L. E. (2021). Evidence-informed teaching and practice-informed research. *Z. Pädagog. Psychol.* 35, 199–208. doi: 10.1024/1010-0652/a000310

Fischer, F. (2021). Some reasons why evidence from educational research is not particularly popular among (pre-service) teachers: a discussion. *Z. Pädagog. Psychol.* 35, 209–214. doi: 10.1024/1010-0652/a000311

Fives, H., and Buehl, M. M. (2012). "Spring cleaning for the "messy" construct of teachers' beliefs: what are they? Which have been examined? What can they tell us?," in *APA Educational Psychology Handbook, Vol 2: Individual Differences and Cultural and Contextual Factors*, eds K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, and M. Zeidner (Washington, DC: American Psychological Association), 471–499. doi: 10.1037/13274-019

Flesch, R. (1948). A new readability yardstick. *J. Appl. Psychol.* 32, 221–233. doi: 10.1037/h0057532

Friedrichsen, P. J., Abell, S. K., Pareja, E. M., Brown, P. L., Lankford, D. M., and Volkmann, M. J. (2009). Does teaching experience matter? Examining biology teachers' prior knowledge for teaching in an alternative certification program. *J. Res. Sci. Teach.* 46, 357–383. doi: 10.1002/tea.20283

Fulmer, S. M., and Frijters, J. C. (2009). A review of self-report and alternative approaches in the measurement of student motivation. *Educ. Psychol. Rev.* 21, 219–246. doi: 10.1007/s10648-009-9107-x

Gill, M. G., Ashton, P. T., and Algina, J. (2004). Changing preservice teachers' epistemological beliefs about teaching and learning in mathematics: an intervention study. *Contemp. Educ. Psychol.* 29, 164–185. doi: 10.1016/j.cedpsych.2004.01.003

Gregoire, M. (2003). Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educ. Psychol. Rev.* 15, 147–179. doi: 10.1023/A:1023477131081

Guilfoyle, L., McCormack, O., and Erduran, S. (2020). The "tipping point" for educational research: the role of pre-service science teachers' epistemic beliefs in evaluating the professional utility of educational research. *Teach. Teach. Educ.* 90:103033. doi: 10.1016/j.tate.2020.103033

Haney, J. J., and McArthur, J. (2002). Four case studies of prospective science teachers' beliefs concerning constructivist teaching practices. *Sci. Educ.* 86, 783–802. doi: 10.1002/sce.10038

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London: Routledge.

Helmsley-Brown, J., and Sharp, C. (2003). How do teachers use research findings to improve their professional practice. *Oxf. Rev. Educ.* 29, 449–471.

Hendriks, F., Seifried, E., and Menz, C. (2021). Unraveling the "smart but evil" stereotype: pre-service teachers' evaluations of educational psychology researchers versus teachers as sources of information. *Z. Pädagog. Psychol.* 35, 157–171. doi: 10.1024/1010-0652/a000300

Hofer, B. K. (2001). Personal epistemology research: implications for learning and teaching. *J. Educ. Psychol. Rev.* 13, 353–383. doi: 10.1023/A:1011965830686

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.2307/1170620

Holt-Reynolds, D. (1992). Personal history-based beliefs as relevant prior knowledge in course work. *Am. Educ. Res. J.* 29, 325–349. doi: 10.2307/1163371

Hopkins, E. J., Weisberg, D. S., and Taylor, J. C. (2016). The seductive allure is a reductive allure: people prefer scientific explanations that contain logically irrelevant reductive information. *Cognition* 155, 67–76. doi: 10.1016/j.cognition.2016.06.011

Janda, L. H., England, K., Lovejoy, D., and Drury, K. (1998). Attitudes toward psychology relative to other disciplines. *Prof. Psychol.* 29, 140–143. doi: 10.1037/0735-7028.29.2.140

Jonkmann, K., Rose, N., and Trautwein, U. (2013). *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen [Tradition and Innovation: Developmental Trajectories at Hauptschulen and Realschulen in Baden-Württemberg and Mittelschulen in Saxony]*. Tübingen: Projektbericht an die Kultusministerien der Länder.

Joram, E. (2007). Clashing epistemologies: aspiring teachers', practicing teachers', and professors' beliefs about knowledge and research in education. *Teach. Teach. Educ.* 23, 123–135. doi: 10.1016/j.tate.2006.04.032

Joram, E., Gabriele, A. J., and Walton, K. (2020). What influences teachers'"buy-in" of research? Teachers' beliefs about the applicability of educational research to their practice. *Teach. Teach. Educ.* 88:102980. doi: 10.1016/j.tate.2019.102980

Karimi, M. N. (2014). Disciplinary variations in English domain-specific personal epistemology: insights from disciplines differing along Biglan's dimensions of academic domains classification. *System* 44, 89–100. doi: 10.1016/j.system.2014.03.002

Keil, F. C., Lockhart, K. L., and Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *J. Exp. Psychol.* 139, 1–15. doi: 10.1037/a0018319

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Z. Pädagog. Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kirschner, P. A., and De Bruyckere, P. (2017). The myths of the digital native and the multitasker. *Teach. Teach. Educ.* 67, 135–142. doi: 10.1016/j.tate.2017.06.001

Kirschner, P. A., and van Merriënboer, J. J. (2013). Do learners really know best? Urban legends in education. *Educ. Psychol.* 48, 169–183. doi: 10.1080/00461520.2013.804395

Klavans, R., and Boyack, K. W. (2009). Toward a consensus map of science. *J. Am. Soc. Inf. Sci. Technol.* 60, 455–476. doi: 10.1002/asi.20991

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., et al. (2013). Teachers' content knowledge and pedagogical content knowledge: the role of structural differences in teacher education. *J. Teach. Educ.* 64, 90–106. doi: 10.1177/0022487112460398

Kleickmann, T., Tröbst, S., Heinze, A., Anschütz, A., Rink, R., and Kunter, M. (2017). "Teacher knowledge experiment: conditions of the development of pedagogical content knowledge," in *Competence Assessment in Education: Research, Models and Instruments*, eds D. Leutner, J. Fleischer, J. Grünkorn, and E. Klieme (Berlin: Springer), 111–130. doi: 10.1007/978-3-319-50030-0_8

Kleickmann, T., Tröbst, S., Jonen, A., Vehmeyer, J., and Möller, K. (2016). The effects of expert scaffolding in elementary science professional development on teachers' beliefs and motivations, instructional practices, and student achievement. *J. Educ. Psychol.* 108, 21–42. doi: 10.1037/edu0000041

König, J., and Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *Eur. J. Teach. Educ.* 39, 419–436. doi: 10.1080/02619768.2016.1214128

Krammer, G., Vogel, S. E., and Grabner, R. H. (2021). Believing in neuromyths makes neither a bad nor good student-teacher: the relationship between neuromyths and academic achievement in teacher education. *Mind Brain Educ.* 15, 54–60. doi: 10.1111/mbe.12266

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., et al. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *J. Educ. Psychol.* 100, 716–725. doi: 10.1037/0022-0663.100.3.716

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037//0022-3514.77.6.1121

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., and Hachfeld, A. (2013). Professional competence of teachers: effects on instructional quality and student development. *J. Educ. Psychol.* 105, 805–820. doi: 10.1037/a0032583

Lilienfeld, S. O. (2012). Public skepticism of psychology: why many people perceive the study of human behavior as unscientific. *Am. Psychol.* 67, 111–129. doi: 10.1037/a0023963

Macdonald, K., Germine, L., Anderson, A., Christodoulou, J., and McGrath, L. M. (2017). Dispelling the myth: training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Front. Psychol.* 8:1314. doi: 10.3389/fpsyg.2017.01314

McNeill, J. (2019). Social validity and teachers' use of evidence-based practices for autism. *J. Autism Dev. Disord.* 49, 4585–4594. doi: 10.1007/s10803-019-04190-y

Menz, C., Spinath, B., and Seifried, E. (2021a). Misconceptions die hard: prevalence and reduction of wrong beliefs in topics from educational psychology among preservice teachers. *Eur. J. Psychol. Educ.* 36, 477–494. doi: 10.1007/s10212-020-00474-5

Menz, C., Spinath, B., and Seifried, E. (2021b). Where do pre-service teachers' educational psychological misconceptions come from?. *Z. Pädagog. Psychol.* 35, 143–156. doi: 10.1024/1010-0652/a000299

Merk, S., Rosman, T., Rueß, J., Syring, M., and Schneider, J. (2017). Pre-service teachers' perceived value of general pedagogical knowledge for practice: relations with epistemic beliefs and source beliefs. *PLoS One* 12:e0184971. doi: 10.1371/journal.pone.0184971

Moser, S., Zumbach, J., Deibl, I., Geiger, V., and Martinek, D. (2021). Development and application of a scale for assessing pre-service teachers' beliefs about the nature of educational psychology. *Psychol. Learn. Teach.* 20, 189–213. doi: 10.1177/1475725720974575

Munro, G. D., and Munro, C. A. (2014). "Soft" versus "hard" psychological science: biased evaluations of scientific evidence that threatens or supports a strongly held political identity. *Basic Appl. Soc. Psychol.* 36, 533–543. doi: 10.1080/01973533.2014.960080

Niemi, H. (2016). "Academic and practical: research-based teacher education in Finland," in *Do Universities have a Role in the Education and Training of Teachers? An International Analysis of Policy and Practice*, ed. B. Moon (Cambridge: Cambridge University Press), 19–33.

Pajares, M. F. (1992). Teachers' beliefs and educational research: cleaning up a messy construct. *Rev. Educ. Res.* 62, 307–332. doi: 10.3102/00346543062003307

Patry, J.-L. (2019). "Situation specificity of behavior: the triple relevance in research and practice of education," in *Progress in Education*, ed. R. V. Nata (Hauppauge, NY: Nova Publishers), 29–144.

Paulsen, M. B., and Wells, C. T. (1998). Domain differences in the epistemological beliefs of college students. *Res. High. Educ.* 39, 365–384. doi: 10.1023/A:1018785219220

Pieschl, S., Budd, J., Thomm, E., and Archer, J. (2021). Effects of raising student teachers' metacognitive awareness of their educational psychological misconceptions. *Psychol. Learn. Teach.* 20, 214–235. doi: 10.1177/1475725721996223

Pozas, M., and Letzel, V. (2021). Pedagogy, didactics, or subject matter? Exploring pre-service teachers' interest profiles. *Open Educ. Stud.* 3, 163–175.

Prinz, A., Kollmer, J., Flick, L., Renkl, A., and Eitel, A. (2021). Refuting student teachers' misconceptions about multimedia learning. *Instr. Sci.* 50, 89–110. doi: 10.1007/s11251-021-09568-z

Richardson, L., and Lacroix, G. (2021). What do students think when asked about psychology as a science? *Teach. Psychol.* 48, 80–89. doi: 10.1177/0098628320959924

Richardson, V. (1996). "The role of attitudes and beliefs in learning to teach," in *Handbook of Research on Teacher Education*, 2nd Edn, eds J. Sikula, T. Buttery, and E. Guyton (New York, NY: Macmillan), 102–106.

Roloff Henoch, J., Klusmann, U., Lüdtke, O., and Trautwein, U. (2015). Who becomes a teacher? Challenging the "Negative Selection" hypothesis. *Learn. Instr.* 36, 46–56.

Rosman, T., Seifried, E., and Merk, S. (2020). Combining intra-and interindividual approaches in epistemic beliefs research. *Front. Psychol.* 11:570. doi: 10.3389/fpsyg.2020.00570

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence-based medicine: what it is and what it is not. *Br. Med. J.* 312, 71–72. doi: 10.1136/bmj.312.7023.71

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *J. Educ. Psychol.* 82, 498–504. doi: 10.1037/0022-0663.82.3.498

Schommer-Aikins, M. (2004). Explaining the epistemological belief system: introducing the embedded systemic model and coordinated research approach. *Educ. Psychol.* 39, 19–29. doi: 10.1207/s15326985ep3901_3

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harv. Educ. Rev.* 57, 1–22. doi: 10.17763/haer.57.1.j463w79r56455411

Siegel, S. T., and Daumiller, M. (2021). Students' and instructors' understandings, attitudes and beliefs about educational theories: results of a mixed-methods study. *Educ. Sci.* 11:197. doi: 10.3390/educsci11050197

Simonton, D. K. (2006). Scientific status of disciplines, individuals, and ideas: empirical analyses of the potential impact of theory. *Rev. Gen. Psychol.* 10, 98–112. doi: 10.1037/1089-2680.10.2.98

Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educ. Res.* 31, 15–21. doi: 10.3102/0013189X031007015

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter Pädagogischer Praxis. *Z. Pädagog. Psychol.* 31, 99–110.

Thomm, E., Gold, B., Betsch, T., and Bauer, J. (2021a). When preservice teachers' prior beliefs contradict evidence from educational research. *Br. J. Educ. Psychol.* 91, 1055–1072. doi: 10.1111/bjep.12407

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021b). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Z. Pädagog. Psychol.* 35, 173–184. doi: 10.1024/1010-0652/a000301

Ulferts, H. (2019). *The Relevance of General Pedagogical Knowledge for Successful Teaching: Systematic Review and Meta-Analysis of the International Evidence from Primary to Tertiary Education*. Berlin: OECD Publishing, doi: 10.1787/ede8feb6-en

van Schaik, P., Volman, M., Admiraal, W., and Schenke, W. (2018). Barriers and conditions for teachers' utilisation of academic knowledge. *Int. J. Educ. Res.* 90, 50–63. doi: 10.1016/j.ijer.2018.05.003

Voss, T., Kunter, M., and Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *J. Educ. Psychol.* 103, 952–969. doi: 10.1037/a0025125

Voss, T., Kunter, M., Seiz, J., Hoehne, V., and Baumert, J. (2014). Die bedeutung des pädagogisch-psychologischen wissens von angehenden lehrkräften für die unterrichtsqualität [The significance of the pedagogical-psychological knowledge of future teachers for the quality of teaching]. *Z. Pädagogik* 60, 184–201.

Voss, T., Zachrich, L., Fauth, B., and Wittwer, J. (2022). The same yet different? Teaching quality differs across a teacher's classes, but teachers with higher knowledge make teaching quality more similar. *Learn. Instr.* 80:101614. doi: 10.1016/j.learninstruc.2022.101614

Voss, T., Zeeb, H., Dehmel, A., and Fauth, B. (2020). "Forschungsmethoden in der lehrerinnen-und lehrerbildung [Research methods in teacher education]," in *Handbuch lehrerinnen-und lehrerbildung*, eds C. Cramer, J. König, M. Rothland, and S. Blömeke (Stuttgart: UTB), 667–672.

Watt, H. M. G., Richardson, P. W., Klusmann, U., Kunter, M., Beyer, B., Trautwein, U., et al. (2012). Motivations for choosing teaching as a career: An international comparison using the FIT-Choice scale. *Teach. Teach. Educ.* 28, 791–805.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., and Gray, J. R. (2008). The seductive allure of neuroscience explanations. *J. Cogn. Neurosci.* 20, 470–477. doi: 10.1162/jocn.2008.20040

Wong, L. Y.-S. (1995). Research on teaching: process-product research findings and the feeling of obviousness. *J. Educ. Psychol.* 87, 504–511. doi: 10.1037/0022-0663.87.3.504

Yadav, A., Herron, M., and Samarapungavan, A. (2011). "Personal epistemology in preservice teacher education," in *Personal Epistemology and Teacher Education*, eds J. L. Brownlee, G. Schraw, and D. Berthelsen (London: Routledge), 25–39.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 October 2022
DOI 10.3389/feduc.2022.1006766

Check for updates

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University, Germany

REVIEWED BY
Marcela Pozas,
Humboldt University of Berlin,
Germany
Verena Letzel-Alt,
University of Trier, Germany

*CORRESPONDENCE
Holger Futterleib
holger.futterleib@uni-erfurt.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 29 July 2022
ACCEPTED 27 September 2022
PUBLISHED 19 October 2022

CITATION
Futterleib H, Thomm E and Bauer J
(2022) The scientific impotence
excuse in education – Disentangling
potency and pertinence assessments
of educational research.
Front. Educ. 7:1006766.
doi: 10.3389/feduc.2022.1006766

The scientific impotence excuse in education – Disentangling potency and pertinence assessments of educational research

Holger Futterleib*, Eva Thomm and Johannes Bauer

Faculty of Education, University of Erfurt, Erfurt, Germany

When facing belief-contradictory scientific evidence, preservice teachers tend to doubt the potency of science and consult scientific sources less frequently. Thus, individuals run the risk not only to maintain questionable assumptions but also to develop dysfunctional stances toward research as a reliable source of knowledge. In two studies, we (a) replicated findings on the so-called *scientific impotence excuse* (SIE) in education and (b) differentiated the effects on the potency and pertinence of science to investigate educational topics to better understand the nature of SIE-related science devaluation. Both studies followed a 2 × 2 mixed experimental design: Preservice teachers assessed their prior belief about an educational topic (i.e., effectiveness of grade retention) before and after reading either confirming or disconfirming scientific evidence concerning the topic. Study 1 (*N* = 147 preservice teachers; direct replication) confirmed the central prior findings of science devaluation when belief-evidence conflicts occur. In contrast, the results of Study 2 (*N* = 152; follow-up study) revealed no systematic devaluations of science when disentangling the facets of potency and pertinence. Despite partial devaluation tendencies, both studies revealed that preservice teachers adapted their prior beliefs to the evidence presented. These findings extend previous research by providing insights into the conditions of science devaluation.

KEYWORDS

preservice teacher education, prior beliefs, motivated reasoning, science devaluation, evidence-based practice

Introduction

When people are confronted with scientific evidence that contradicts their entrenched prior beliefs, they often engage in a variety of defensive responses in order to brush aside the evidence rather than to revise their personal assumptions (Kunda, 1990; Chinn and Brewer, 1993, 1998; Nauroth et al., 2014). One defensive response of such motivated reasoning can be to devalue the potency of science to study a given issue, at all. Munro (2010) dubbed this response the *scientific impotence excuse* (SIE): Facing belief-inconsistent scientific evidence, individuals dismiss the information on the grounds "that the topic of study is not amenable to scientific investigation" (Munro, 2010, p. 579). This tendency is highly problematic, because people do not only discount the particular piece of evidence; worse, they devalue the potency of science as an epistemic enterprise to attain valid knowledge about the topic. Employing the SIE may, thus, pave the way to generalized science denial (e.g., Nisbet et al., 2015; Lewandowsky and Oberauer, 2016; Hornsey, 2020).

Recently, Thomm et al. (2021a) showed that the SIE could also account for why preservice teachers develop critical stances toward educational research findings—a crucial barrier to implement evidence-informed practices as early as in initial teacher education (e.g., van Schaik et al., 2018). Despite reporting generally positive attitudes toward educational research, preservice teachers began to question its potency to examine the topic at stake when facing belief-inconsistent educational research findings. However, it remains unclear whether this devaluation concerns only participants' doubt on the *epistemic value* of research (i.e., its *potency* to provide valid knowledge) or extends to their doubt on the *pertinence* of research to investigate the topic at hand. The latter would be worse, because it strips educational research of its role as a relevant social institution to deliver reliable and valid knowledge on the topic.

In two experiments, we aimed to inspect the stability of prior findings and to disentangle the nature of devaluation in the context of preservice teachers' evaluation of educational research and its findings. Study 1 sought to replicate directly the main findings of Thomm et al. (2021a). The direct replication was chosen to test the previously found and surprising pattern showing that preservice teachers devalued the potency of science while adjusting their beliefs in the direction of the evidence presented. We also intended to corroborate the findings of the prior study, especially as it is one of the few to address preservice teachers' science devaluation. Study 2 complemented previous studies by investigating the potential effects of devaluation on both the potency and pertinence of educational research. This contribution extends prior research, as it evaluated the stability of the SIE as a mechanism of science devaluation and offers a differentiation of its effects on educational research's potency and pertinence. Thereby, the studies provide important insights into the pitfalls that must be considered when preservice teachers engage with scientific evidence in research-based teacher education.

Understanding devaluation of educational research

Research on motivated reasoning indicates that individuals tend to argue away information that threatens their prior beliefs (Kunda, 1990; Chinn and Brewer, 1993, 1998; Nauroth et al., 2014; Britt et al., 2019). Chinn and Brewer (1993, 1998) identified several ways in which people react to such belief-discrepant evidence. Instead of revising prior assumptions, individuals may turn to ignore, reject, or reinterpret anomalous information to protect their beliefs. The SIE, as posed by Munro (2010) on the basis of cognitive dissonance theory (Festinger, 1957), complements this array of protective mechanisms, but goes beyond the mere rejection of a piece of scientific evidence. Using the SIE, individuals resolve the belief-evidence discrepancy by devaluating the ability of science to study the topic, that is, science's *potency*. Thus, individuals justify devaluation by claiming that the issue cannot be investigated by the means of science, and run the risk to develop unfavorable, generalized attitudes toward it. Munro (2010) also suggests that people are particularly prone to employ the SIE if the scientific information is strong and, thus, cannot be argued away easily (e.g., by referring to flawed methods). Indeed, SIE arguments are apparent in public debates, for example, when proponents of homeopathy claim that its effects cannot be studied by standard scientific methods such as randomized controlled trials.

Across two experiments, Munro (2010) provided evidence for the SIE. After indicating their prior beliefs about a specific medical claim, participants read scientific evidence (i.e., short abstracts) that either confirmed or disconfirmed this claim. In line with the SIE, the results showed that participants systematically discounted the potency of science to study the topic if the read evidence contradicted their prior beliefs. Moreover, they generalized their doubt to the investigation of other unrelated scientific topics and were even less inclined to choose scientific sources to inform themselves. Complementing the overall picture, participants also resisted changing their prior beliefs. Drawing on these studies, Thomm et al. (2021a) examined whether the SIE could also be observed when preservice teachers faced belief-discrepant evidence from educational research. Education is a particularly interesting field to study the SIE, first, because there is a sharp contrast between the developments to make teaching a more research-based profession (Bauer and Prenzel, 2012; Rousseau and Gunia, 2016), and the empirical observations that teachers rarely draw on educational research and, rather, rely on personal observations and common sense (Dagenais et al., 2012; Lysenko et al., 2014; Pieschl et al., 2021).

Second, the social sciences may be particularly vulnerable to devaluation, as they are often perceived as "soft" and unreliable (Berliner, 2002). Hence, people may overestimate their own abilities in judging and explaining educational issues—a tendency exacerbated by the seeming verification of beliefs through everyday observations (Thomm et al., 2021b).

In line with Munro's (2010) results, Thomm et al. (2021a) found that preservice teachers facing belief-discrepant evidence on an educational issue (i.e., the effectiveness of grade retention on low-achieving students' academic progress) devalued the potency of science, and showed a lower preference for scientific sources to further inform themselves about this topic. This devaluation occurred even though participants reported overall positive attitudes toward educational research. However, unlike the evidence reported in Munro (2010), participants did not expand their doubt about scientific potency to the investigation of other educational or unrelated topics. Moreover, participants tended to change their beliefs in the direction of the presented evidence. Apparently, this evidence worked as a refutation of participants' prior beliefs (cf. Tippett, 2010; Kendeou et al., 2014), even though the scientific abstracts used in the study were not designed according to refutation text principles.

Though these studies corroborated the main hypotheses regarding the SIE, the stability and nature of the effect require further inquiry. First, given its significance in the educational context, replication of the effect is important to evaluate its consistency and strength. Second, different aspects of the devaluation need to be disentangled more thoroughly, as elaborated below.

Disentangling devaluation of potency and pertinence of educational research

Since many educational topics are accessible to one's own experiences and observations (Calderhead and Robson, 1991; Pajares, 1992; Richardson, 1996; Menz et al., 2021a), it may not be immediately obvious why these topics are subject to research, at all, and that research knowledge might be useful for teachers. Consequently, devaluation may not be confined to the potency of research to study educational topics; people may also contest that educational research is *pertinent* to do so (cf. Bromme and Thomm, 2016). While potency refers to the assigned epistemic value of research, pertinence represents a normative ascription of the relevant expertise to it and a mandate to contribute valuable knowledge about the domain at stake (Kitcher, 2011; Bromme and Gierth, 2021). Thus, questioning the pertinence of educational research is a more fundamental form of science rejection than doubting its potency. Discounting pertinence would allow dismissing educational research simply as "fishing in foreign waters," even if one had to admit that scientific methods principally can contribute valid knowledge.

Existing studies have not yet differentiated (preservice) teachers' appraisal of the potency and pertinence of educational research. However, some studies implicitly have addressed aspects related to pertinence. As mentioned above, there is a multitude of studies indicating that (preservice) teachers frequently judge educational research as irrelevant and detached from their practice (McIntyre, 2005; Hammersley, 2013; Winch et al., 2015; Farley-Ripple et al., 2018; van Schaik et al., 2018; Thomm et al., 2021b) and favor experience-based knowledge, instead (e.g., Bråten and Ferguson, 2015; van Schaik et al., 2018; Kiemer and Kollar, 2021). Cain (2017) found that teachers did not only question the validity of findings but assigned science "no greater authority than their own experiences or other forms of information" (p. 13). Though such findings shed some light on teachers' perceptions of the pertinence of educational research, it is still an open issue how such perceptions are influenced by belief-evidence conflicts and how this relates to potency appraisals.

Overview of the studies

The present contribution aimed to provide to a better understanding of the nature of the devaluation of educational research through preservice teachers by examining the SIE in two studies. Study 1 was a direct replication of Thomm et al. (2021a) that aimed to inspect the stability and strength of the effect of belief-evidence conflicts on the SIE. Study 2 was a follow-up (Schmidt, 2009) that aimed to differentiate the assessments of the perceived potency and pertinence of educational research as facets of devaluation, and to increase the external validity of prior findings. Both studies were preregistered[1]. To prevent possible cross-participation, they were conducted simultaneously with participants being assigned randomly to one of the respective studies.

Study 1

For the direct replication, we stated the same hypotheses as Thomm et al. (2021a). Though, as elaborated above, some of their results were inconsistent with the theoretical predictions and Munro's (2010) results (i.e., hypotheses H2a and H3, below), we decided to retain the original hypotheses because they reflect

1 Study 1: https://osf.io/prj87
Study 2: https://osf.io/m4eaj

the theoretical reasoning behind the SIE (Munro, 2010). Hence, we examined the following hypotheses:

> H1a: Preservice teachers are more critical about the potency of educational research to study a specific educational topic when scientific evidence contradicts rather than confirms preservice teachers' prior beliefs.

> H1b: Preservice teachers reading scientific evidence that contradicts rather than confirms their prior beliefs will show a decreased preference for scientific sources and, conversely, an increased preference for non-scientific sources.

> H1c: Preservice teachers choose scientific sources less often than non-scientific sources to seek additional information about the specific educational topic.

> H2a: Preservice teachers generalize their devaluation of educational research by doubting educational research's potency to study further educational topics.

> H2b: There are no carry-over effects of science devaluation to topics from other unrelated domains (medicine and pseudo-scientific).

> H3: Preservice teachers retain their prior beliefs although scientific evidence might contradict them.

Methods

For the direct replications, all methods followed the design, materials, and procedures of Thomm et al. (2021a).

Design

Study 1 was a 2 × 2 mixed experiment with the within-participants factor prior belief (*before vs. after reading the evidence*) and the between-participants factor evidence (*confirming vs. disconfirming the effectiveness of grade retention*). Accordingly, preservice teachers were randomly assigned to one of two conditions, reading either confirming or disconfirming scientific evidence on the effectiveness of grade retention (GR) to reduce potential deficits in students' school achievement. Before and after reading the assigned evidence, participants reported their respective beliefs about GR effectiveness.

Participants

Based on an *a priori* power analysis with GPower 3.1 (Faul et al., 2009), we aimed at an effective sample size of $N = 202$ preservice teachers to detect the effects of at least $f^2 = 0.087$ (cf. Thomm et al., 2021a) with 95% power ($\alpha = .05$) in a multiple linear regression[2].

Participants were recruited online through invitation via participant databases, advertisements at university lectures, and mailing lists in Germany. Participation was voluntary and participants could withdraw at any time without giving reasons or experiencing any consequences. As an incentive, participants could enroll in a lottery with winnings in the amount of 10–20€.

A total of $N = 237$ participants completed the study. During data cleaning, we deleted cases according to preregistered exclusion criteria as follows: $n = 15$ participants who had not provided informed consent or had withdrawn it; one case who was not enrolled in a teacher education program; $n = 30$ participants with unreasonable response times for completing the experiment (i.e., <5 or >120 mins); and $n = 44$ participants who had spent less than 1 min on the evidence reading task. This reflects an exclusion rate of 38%. The final sample of $N = 147$ still provided 85.4% power and allowed detecting effects as low as $f^2 = 0.076$ with 80% probability. Preservice teachers in the final sample were mostly female (83.7%) and $M = 21.9$ years old ($SD = 2.77$ years). They were mostly studying in a bachelor's degree program (49.7%) and had completed, on average, 3.82 semesters ($SD = 1.92$). Further, 31.3% were enrolled in a master's degree program with a duration of study of $M = 8.22$ semesters ($SD = 1.71$). The remaining 19% studied in traditional state examination programs (with no separate Bachelor's and Master's phase) with a duration of study of $M = 4.9$ semesters ($SD = 2.7$).

Procedure

The study was realized as an online experiment. After the general introduction and giving informed consent, participants reported demographic information (i.e., age, gender, study program). Subsequently, they rated their prior *belief about the effectiveness of GR*. Then, they read an introductory text on the topic of GR and were randomly assigned to one of two conditions presenting either confirming or disconfirming scientific evidence (i.e., five short abstracts) on the effectiveness of GR. After reading the evidence, participants rated the potency of science to study the topic at stake (*topic-specific potency*), as well as the potency of science to study further domain-related and domain-unrelated topics (*domain-related* and *domain-unrelated potency*). They further judged their preferences for scientific and non-scientific sources to learn more about GR effects (*source preference*) and were asked to select their most preferred source (*source choice*). Finally, participants reassessed their beliefs about the effectiveness of GR. Having finished the experiment, participants had the option to withdraw their consent for data usage and received a thorough debriefing.

2 Due to a technical error, the preregistered power analysis indicated a recommended sample size of $N = 181$, but $N = 202$ participants would have been needed to achieve the power of 95%.

Materials

Preservice teachers first read a short introductory text describing a recently published review summarizing scientific studies about the effects of GR on remedying low-performing students' deficits in school achievement. Next, each experimental group received five abstracts, each summarizing an empirical study on GR. The abstracts had been designed and tested by Thomm et al. (2021a) for the target group and to provide equivalent levels of length and complexity across the experimental conditions. The abstracts had a standardized form equivalent to typical study abstracts and ended with a clear final conclusion on the effectiveness of GR. Overall, they were representative of scientific research in this field regarding the applied methods, results, and conclusions. Across the evidence conditions, results and conclusions varied from confirming to disconfirming the effectiveness of grade retention. For more information on the development of the materials, see Thomm et al. (2021a); materials are available in the corresponding Appendix S1.

Measures

Unless indicated otherwise, the answer format for all measures described below was a 9-point rating scale (1 = do not agree at all, 9 = very much agree).

Prior belief was measured by rating the statement "Repeating a grade helps struggling students to compensate for their achievement deficits."

Topic-specific doubt on the potency of science to study the effectiveness of GR was assessed by rating the statement "The question whether GR helps struggling students to compensate their deficits in achievement is one that cannot be answered using scientific methods" (cf. Munro, 2010).

To measure the *generalization of science devaluation*, participants assessed the potency of science to study six additional educational topics (e.g., impact of class size on learning outcomes) and eight unrelated topics. The unrelated topics covered health issues (e.g., cell phone radiation causing cancer) and pseudo-scientific topics (e.g., astrology as a possible predictor of personality; Munro, 2010). Per topic, the participants rated the statement "How far can scientific methodologies be used to determine whether (e.g., computer-based learning supports students' knowledge acquisition)?" (1 = not at all, 9 = very well). Items were averaged per domain and resulted in sufficiently reliable scores (education: $\alpha = 0.68$, medicine: $\alpha = 0.81$, and pseudo-science: $\alpha = 0.73$).

As a measure of *source preferences* participants received a list of seven scientific (i.e., findings from scientific studies; educational scientist) and non-scientific sources (i.e., opinions and experiences of teacher, school student who repeated a class, teacher association, proponent and opponent of GR). They judged how likely they would be to seek information from each source. Both scales yielded acceptable reliability (preference for scientific sources, $\alpha = 0.62$; preference for non-scientific sources, $\alpha = 0.63$)[3]. In addition to their source preferences, participants had to choose one source from the list that they would finally consult (*source choice*).

Analyses

We performed all analyses using R 4.0.2 (R Core Team, 2022). To test hypotheses H1a, H1b, H2a, and H2b (i.e., moderating effects of prior belief on the relation of scientific evidence and doubt of scientific potency, respectively their source preferences), we conducted multiple regression analyses using Hayes' PROCESS macro 4.1 for R (Hayes, 2022). The evidence condition was dummy-coded (0 = confirming vs. 1 = disconfirming the effectiveness of GR), and prior belief was centered at the grand mean. For testing H2b, we set $\alpha = 0.20$ because the null hypothesis was the target (i.e., no generalization of doubting science to topics from unrelated domains).

To test H1c, we ran a binary logistic regression regarding whether belief-evidence conflicts decrease the choice of a scientific source (coded as 1) over a non-scientific one (coded as 0). H3 (i.e., belief change) was tested by a repeated-measures ANOVA.

As preregistered, we applied transformations to variables exhibiting highly asymmetric distributions prior to analysis in order to avoid biased standard errors and significance tests (Tabachnick and Fidell, 2014; Fox, 2016). Highly asymmetric distributions were characterized by both P-P plots and significant tests of skewness ($z \geq |2.58|$, indicating $p \leq 0.01$; Field et al., 2013). We applied log transformation to variables with positive moderate skew (i.e., topic-specific potency), inverse transformation to variables with extreme positive skew (i.e., pseudo-scientific topics potency), and a reflect-and-log transformation for variables exhibiting moderate negative skew (i.e., domain-specific potency, preference for scientific sources; Tabachnick and Fidell, 2014).

Results

Table 1 provides an overview of the descriptive statistics. Scale means indicated that participants altogether had favorable beliefs about the potency of science and a noteworthy preference for scientific sources to inform themselves about educational topics.

Devaluation of the potency of science and its sources (hypotheses 1a–1c)

To test H1a, we regressed doubt over the potency of science on evidence condition, prior beliefs, and their interaction. Analyses yielded a statistically significant overall effect, *F*(3,

3 Like Thomm et al. (2021a), we averaged the preference ratings for proponent and opponent of GR because of their high correlation ($r = 0.89$, $p < 0.001$).

TABLE 1 Study 1: Descriptive statistics by evidence condition.

	M (*SD*)	Skew (*SE*)	Kurtosis (*SE*)
Evidence confirming effectiveness of grade retention			
Belief about GR effectiveness (T1)	5.63 (2.08)	−0.53 (0.28)	−0.52 (0.55)
Belief about GR effectiveness (T2)	7.40 (1.36)	−1.27 (0.28)*	2.47 (0.55)
Doubt on science's potency to study GR	2.16 (1.55)	2.09 (0.28)*	4.38 (0.55)
Potency to study related educational topics	6.40 (1.38)	−1.21 (0.28)*	2.50 (0.55)
Potency to study unrelated topics			
Medicine	5.81 (1.84)	−0.67 (0.28)	0.06 (0.55)
Pseudo-science	2.11 (1.36)	1.79 (0.28)*	3.69 (0.55)
Source Preference			
Scientific sources GR	7.73 (1.00)	−0.76 (0.28)*	0.05 (0.55)
Non-scientific sources GR	6.06 (1.34)	−0.28 (0.28)	0.37 (0.55)
Evidence disconfirming effectiveness of grade retention			
Belief about GR effectiveness (T1)	5.90 (2.13)	−0.40 (0.28)	−1.03 (0.56)
Belief about GR effectiveness (T2)	3.35 (1.85)	0.90 (0.28)*	−0.10 (0.56)
Doubt on science's potency to study GR	3.11 (2.10)	1.04 (0.28)*	0.19 (0.56)
Potency to study related educational topics	6.18 (1.33)	−1.43 (0.28)*	4.37 (0.56)
Potency to study unrelated topics			
Medicine	5.45 (1.54)	−0.11 (0.28)	−0.14 (0.56)
Pseudo-science	2.65 (2.09)	1.56 (0.28)*	1.64 (0.56)
Source Preference			
Scientific sources GR	7.36 (1.52)	−0.96 (0.28)*	0.29 (0.56)
Non-scientific sources GR	5.71 (1.42)	−0.28 (0.28)	0.00 (0.56)

Values represent untransformed data; GR = grade retention; *P-P-plots and significant tests of skew ($z \geq |2.58|$, $p \leq 0.01$) indicated serious skewness.

143) = 4.90, $p = 0.003$, $R^2 = 0.09$. In line with our assumption, the interaction term of evidence condition and prior beliefs confirmed a significant moderation effect of prior belief, $b = 0.04$, $SE(b) = 0.02$, $t = 2.02$, $p = 0.046$. **Figure 1** depicts the crossover interaction entailed by H1a. Additionally, the regression model yielded a statistically significant main effect of evidence condition, $b = 0.14$, $SE(b) = 0.05$, $t = 3.15$, $p = 0.002$.

Probing the interaction, we conducted a simple slopes analysis. Findings revealed a statistically significant simple effect of evidence condition for participants with both high [i.e., 1 *SD* above the sample mean; $b = 0.23$, $SE(b) = 0.06$, $t = 3.66$, $p < 0.001$] and average prior belief [i.e., at the sample mean; $b = 0.14$, $SE(b) = 0.05$, $t = 3.15$, $p = 0.002$], but not for participants with low prior belief [i.e., 1 *SD* below the sample mean; $b = 0.05$, $SE(b) = 0.06$, $t = 0.80$, $p = 0.452$]. That is, particularly participants with strong or average prior beliefs in GR effectiveness tended to doubt the potency of science when evidence contradicted their prior beliefs.

To test H1b, we regressed the preferences for scientific sources on evidence condition, prior beliefs, and their interaction. The overall model was statistically significant, $F(3, 142) = 3.64$, $p = 0.014$, $R^2 = 0.07$, and yielded a statistically significant interaction, $b = 0.04$, $SE(b) = 0.02$, $t = 2.54$, $p = 0.012$ (**Figure 2**).

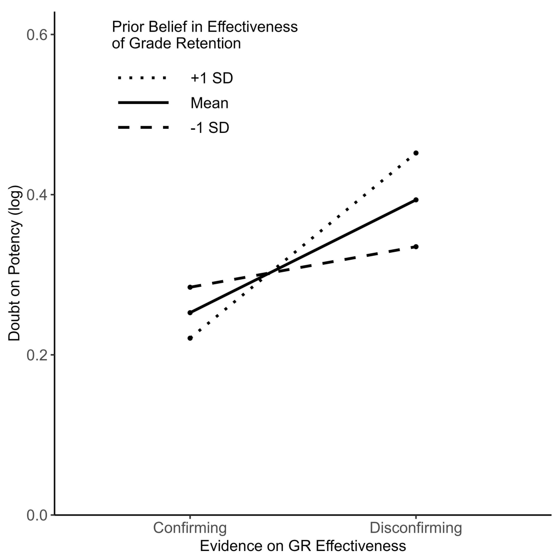


FIGURE 1
Doubt on the potency of science to study the effectiveness of grade retention.

Subsequent simple slopes analysis revealed a statistically significant effect of evidence on participants with high prior belief [i.e., 1 *SD* above the sample mean; $b = 0.13$, $SE(b) = 0.05$,

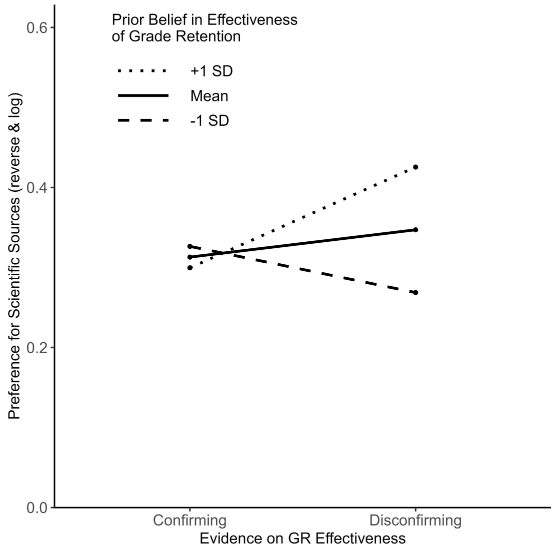


FIGURE 2
Preference for scientific soures. Higher values indicate lower source preference ratings due to transformation.

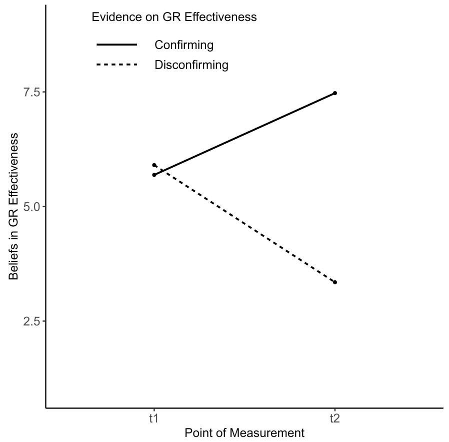


FIGURE 3
Belief change regarding the effectiveness of grade retention (Study 1).

t = 2.47, p = 0.015] but not for participants with average [i.e., at the sample mean; b = 0.03, $SE(b)$ = 0.04, t = 0.95, p = 0.346] and low prior belief [i.e., 1 SD below sample mean; b = −0.06, $SE(b)$ = 0.05, t = −1.13, p = 0.260]. Thus, facing belief-discrepant evidence, participants with high prior belief in the effectiveness of GR tended to have a decreased preference for scientific sources.

An analogously performed regression with preference for non-scientific sources did not attain significance, $F(3, 142) = 0.87$, $p = 0.456$, $R^2 = 0.02$.

The logistic regression for source choice (H1c) failed to attain statistical significance for the overall model [$\chi^2(3) = 7.60$, p = 0.055], even though the interaction effect entailed by the hypothesis yielded statistical significance [b = −0.40, $\chi^2(1) = 4.81$, $p = 0.028$].

Generalization of science devaluation to further domain-related and unrelated topics (hypotheses 2a and 2b)

Contrary to H2a, the regression model for the potency of science to investigate further educational topics (H2a) did not yield statistical significance, $F(3, 143) = 0.79$, p = 0.501, R^2 = 0.02. The same was true for the regression models for unrelated medical [$F(3, 143) = 0.63$, p = 0.600, R^2 = 0.01] and pseudo-scientific topics [$F(3, 143) = 0.52$, p = 0.671, R^2 = 0.01]. The latter two findings, however, are in line with the expected null effects for these dependent variables (H2b). Hence, participants did not generalize their doubt over the potency of science to other related or unrelated topics.

Belief change in the face of belief-contradictory evidence (hypothesis 3)

A mixed ANOVA with the within-participants factor belief (before vs. after reading the evidence) and the between-participants factor evidence condition (confirming vs. disconfirming evidence) yielded statistically significant main effects of evidence condition [$F(1, 144) = 54.14$, $p < 0.001$, η^2 = 0.22] and prior belief [$F(1, 144) = 6.92$, p = 0.009, η^2 = 0.01], as well as a statistically significant interaction [$F(1, 144) = 218.76$, $p < 0.001$, η^2 = 0.26] (**Figure 3**). Follow-up dependent samples t-tests indicated that participants changed their beliefs in the direction of the presented evidence. Participants who had read confirming evidence increased their belief in the effectiveness of GR, $t(73) = 9.40$, $p < 0.001$, $d = 1.09$, whereas participants reading disconfirming evidence decreased it, $t(71) = 11.38$, $p < 0.001$, $d = 1.34$.

Conclusion

Study 1 aimed to directly replicate the findings of Thomm et al. (2021a). We found the same pattern of findings as the original study with comparable (small) effect sizes, the only exception being the non-significant overall model of source choice. Both studies support the SIE (Munro, 2010): In the face of strong, belief-threatening evidence that cannot be brushed aside easily, people may turn to devalue the potency of science to study the issue to protect their beliefs (H1a). This discounting also affected preferences for scientific sources (H1b). However, concerning source choice (H1c), we only found a tendency in the expected direction: The significant interaction effect was invalidated by the non-significant overall test such that this result should not be interpreted. These discrepant findings suggest that a lower preference for scientific sources does not translate into increased preference for non-scientific sources, *per se*, hence, favoring the choice of the latter. Preservice teachers are, instead, more likely to consider scientific sources to answer a (scientific) question.

In contrast to the findings by Munro (2010), the results of this study and Thomm et al. (2021a) unanimously suggest that individuals do not necessarily transfer devaluation to further related and unrelated topics. Our results indicate that preservice teachers did not generalize their doubt about the potency of science to other educational research topics (H2a). There was also no evidence of carry-over effects to unrelated medical or pseudo-scientific topics. Thomm et al. (2021a) already discussed that individuals perceive knowledge domains differently and, therefore, may not automatically generalize devaluation across unrelated domains. Overall, this may indicate that the devaluation implied by the SIE seems to be a topic-related phenomenon, at least at first encounter.

Moreover, despite the reported difficulty to initiate belief revision (e.g., Richardson, 1996; Lewandowsky et al., 2012), both Thomm et al. (2021a) and this replication found that participants' prior beliefs shifted toward the conclusion supported by the read evidence. Though unexpected, these findings can possibly be explained by literature on knowledge revision and refutational texts (Tippett, 2010; Kendeou et al., 2014; Butterfuss and Kendeou, 2021). We will elaborate more deeply on both issues in the general discussion.

In summary, despite the discrepant result regarding H1c, overall, we conclude that the present study successfully replicated the main findings reported in Thomm et al. (2021a).

Study 2

It remained an open question whether the devaluation found from Study 1 mainly related to doubt regarding the epistemic value of research, or whether it extends to the perceived pertinence of science to investigate a topic. These aspects of devaluation can occur independently or in combination. For example, individuals can question the potency of science to investigate a specific issue, while still considering science as generally pertinent to providing reliable and valid knowledge on it. As discussed above, devaluating the pertinence of science would constitute an even stronger case of science devaluation compared to discounting potency, solely. Thus, to better understand the nature of SIE-related devaluation, Study 2 examined whether belief-evidence conflicts affect both the perceived potency and pertinence of educational research to study educational topics. Regarding potency, we tested the same hypotheses as in Study 1. Concerning pertinence, we stated two research questions focused on the specific educational topic (i.e., GR effects) and on generalization to other relevant topics. We did not expect or test for generalization to other unrelated (i.e., medical and pseudo-scientific) topics, because we considered pertinence assessments as a strongly topic-related phenomenon (i.e., educational research is pertinent to answer educational research questions, but not for questions from other domains).

In addition to these substantive issues, in Study 2, we also sought to enhance the external validity of the experimental materials. To assure experimental control, Thomm et al. (2021a) and Study 1 had presented participants with the same texts across both evidence conditions, manipulating only the direction of the results and conclusion. In Study 2, we presented summaries of original published studies on grade retention effects as evidence. Consequently, participants read different studies in the respective condition.

Methods

All methods and procedures were identical to Study 1 unless indicated otherwise, below.

Design and procedure

Perception of the pertinence of educational research was added as a dependent variable to the design. This resulted in minor modifications of the procedure: After reading the introductory text and the scientific evidence, participants assessed the topic-specific potency and topic-specific pertinence of science, both displayed on one page of the online questionnaire. Next, participants rated the potency of science to study educational topics (domain-related) and other topics (domain-unrelated). Afterward, participants answered the pertinence questions on further educational topics. We deliberately placed potency before pertinence assessments for each respective topic, instead of balancing their order. This was done to ensure that potency measures remained unaffected by the respective pertinence measures and to maintain comparability to the effects on potency across Study 1 and Study 2. We judged this as being more important than ruling out potential position effects (see general discussion).

Participants

A total of N = 221 participants completed Study 2. We removed cases according to the preregistered exclusion criteria as follows: n = 10 participants with missing consent declarations; n = 2 participants not enrolled in a teacher education program; n = 9 for unreasonable response times (i.e., <5 or >120 mins); and n = 48 participants who spent less than 1 min on the page with the evidence (experimental manipulation); overall exclusion rate 31.2%. The final sample (N = 152) still yielded sufficient statistical power (86.7%) and allowed detecting effects as low as f^2 = 0.074 with 80% probability. Participants in the final sample were mostly female (84.1%), M = 22.23 years old (SD = 3.24 years) and mostly studying in a bachelor's degree program (49.3%) with a duration of study of M = 3.88 semesters (SD = 1.91). Of the participants, 33.6% were enrolled in a master's degree program (M = 8.61 semesters, SD = 2.74), and 17.1% in a state examination program (M = 5.23 semesters, SD = 2.69).

Materials

To create a new set of scientific abstracts, either confirming or disconfirming the effectiveness of GR, that were analogous in design to Study 1, we carried out an extensive literature research to identify appropriate original research articles addressing the effectiveness of GR. We then formed pairs of studies, one confirming and one disconfirming the effectiveness of GR, that were comparable in publication year, outcome variables, and/or the methods used. Due to the use of original studies as a basis, the abstracts were not equivalent across conditions, as they had been in Study 1. Specifically, the evidence presented to disconfirm the effectiveness of GR was more diverse than it was in the previous studies, reporting both null findings and negative effects of GR. We harmonized the abstracts in structure, length (M = 168.2, SD = 18.8 words), and complexity (e.g., statistical methods) to enhance comparability across conditions. Moreover, we added the original author names and the year of publication. The material was pretested and revised via cognitive interviews with N = 10 students from different subjects (i.e., psychology, education, and teaching) regarding comprehensibility, consistency, and methodological soundness. The abstracts are available in the respective OSF directory.

Measures

The measures of *prior belief* (both measurement points), *topic-specific potency*, and *potency for domain-related* and *domain-unrelated* topics were identical to Study 1. However, the reliabilities of the potency scales for related and unrelated topics were weaker than in Study 1 (education: α = 0.55, medicine: α = 0.76, and pseudo-science: α = 0.63). To maintain the consistency of the measures across studies, we decided to retain these scales as they were. The results for potency for further educational topics should be interpreted cautiously due to the low reliability of this scale.

We made minor adjustments in *source preference and choice* to cover a broader range of scientific sources (i.e., research results in scientific journals, scientific textbooks, opinion of an educational scientist, applied educational, or popular science journals) and non-scientific sources (i.e., the education section of the daily press, educational guidebooks, the experiences of a seasoned teacher, the experiences of other preservice teachers, the experiences of family or friends). Because of these changes, we inspected the factorial structure with exploratory factor analysis and built scales on this basis. Two items had to be deleted due to high cross-loadings. The resulting scales had acceptable reliabilities (preference for scientific source α = 0.70; preference for non-scientific sources α = 0.80).

Topic-specific doubt on the *pertinence of science* was measured with the statement "Science is not pertinent to answer the question of whether GR helps struggling students to compensate for their deficits in achievement." Pertinence devaluation items regarding further educational topics referred to the same topics as the potency assessments. Participants rated, for each topic, the statement "To what extent is science pertinent to investigate whether (e.g., computer-based learning supports students' knowledge acquisition)?" The average scores displayed acceptable reliability: α = 0.83.

Results

Table 2 provides an overview of the descriptive statistics. Scale means indicated that participants had favorable beliefs about the potency and pertinence of science and a high preference for scientific sources to inform themselves about the educational topic at stake. The ratings of potency and pertinence were positively correlated for both grade retention ($r = 0.46, p < 0.001$) and the further educational topics ($r = 0.45, p < 0.001$).

Devaluation of potency and pertinence of science, scientific source preferences, and source choice (hypotheses 1a, 1b, and 1c; Research question 1)

To inspect potential devaluation, we tested the impact of belief-evidence conflicts on the assessments of both the potency and pertinence of science. The overall regression model for topic-specific potency (log-transformed) failed statistical significance, $F(3, 148) = 2.19, p = 0.092, R^2 = 0.04$, even though the interaction effect entailed by H1a was significant, $b = 0.06$, $SE(b) = 0.02$, $t = 2.34$, $p = 0.02$. The regression model for topic-specific pertinence (log-transformed) on evidence condition, prior beliefs, and their interaction also failed statistical significance, $F(3, 148) = 1.84, p = 0.142, R^2 = 0.04$.

Also, the regression model for scientific source preference (H1b) was not statistically significant, $F(3, 148) = 0.07, p = 0.974, R^2 = 0.001$. The analogously performed regression for non-scientific source preferences (reflected and square-root-transformed) yielded a statistically significant overall model [$F(3, 148) = 2.94, p = 0.035, R^2 = 0.04$], but no significant individual effects. Finally, the logistic regression for source choice (H1c) also remained non-significant, $\chi^2(3) = 3.23, p = 0.308$.

Generalizing science devaluation to further educational and unrelated topics (hypotheses 2a and 2b; Research question 2)

Contrary to H2a, participants showed no systematic devaluation of the potency of science to investigate other educational topics, $F(3, 148) = 0.18, p = 0.905$,

TABLE 2 Study 2: Descriptive statistics by evidence condition.

	M (*SD*)	Skew (*SE*)	Kurtosis (*SE*)
Evidence confirming effectiveness of grade retention			
Belief about GR effectiveness (T1)	5.77 (1.74)	−0.42 (0.30)	−0.13 (0.60)
Belief about GR effectiveness (T2)	7.47 (1.18)	−1.09 (0.30)	1.86 (0.60)
Doubt on science's potency to study GR	2.37 (1.51)	1.35 (0.30)*	1.46 (0.60)
Doubt on science's pertinence to study GR	2.31 (1.43)	0.73 (0.30)	−0.64 (0.60)
Potency to study related educational topics	6.60 (1.12)	0.09 (0.30)	−0.49 (0.60)
Pertinence to study related educational topics	6.89 (1.44)	−0.58 (0.30)	0.17 (0.60)
Potency to study unrelated topics			
Medicine	6.06 (1.31)	−0.04 (0.30)	−0.82 (0.60)
Pseudo-science	2.61 (1.52)	1.17 (0.30)*	1.31 (0.60)
Source Preference			
Scientific sources GR	7.54 (1.25)	−0.97 (0.30)*	0.42 (0.60)
Non-scientific sources GR	6.01 (1.60)	−0.11 (0.30)	−0.39 (0.60)
Evidence disconfirming effectiveness of grade retention			
Belief about GR effectiveness (T1)	5.79 (1.88)	−0.45 (0.25)	−0.52 (0.50)
Belief about GR effectiveness (T2)	2.94 (1.54)	1.01 (0.25)	0.51 (0.50)
Doubt on science's potency to study GR	2.57 (1.62)	1.10 (0.25)*	0.64 (0.50)
Doubt on science's pertinence to study GR	2.09 (1.47)	1.58 (0.25)*	2.61 (0.50)
Potency to study related educational topics	6.67 (1.08)	−0.33 (0.25)	0.75 (0.50)
Pertinence to study related educational topics	7.08 (1.38)	−0.65 (0.25)	0.20 (0.50)
Potency to study unrelated topics			
Medicine	6.28 (1.42)	−0.24 (0.25)	0.13 (0.50)
Pseudo-science	3.12 (2.09)	1.10 (0.25)*	0.32 (0.50)
Source Preference			
Scientific sources GR	7.57 (1.08)	−0.72 (0.25)*	0.31 (0.50)
Non-scientific sources GR	5.92 (1.86)	−0.33 (0.25)	−0.74 (0.50)

Values represent untransformed data; GR = grade retention; *both P-P-plots and significant tests of skew ($z \geq |2.58|$, $p \leq 0.01$) indicated serious skewness.

R^2 = 0.004. Analogously, there were no effects on pertinence assessments (reflected and square-root-transformed) for these additional educational topics, $F(3, 148) = 1.22$, $p = 0.303$, $R^2 = 0.02$.

In line with H2b, regression models for potency regarding medical [$F(3, 148) = 0.58$, $p = 0.626$, $R^2 = 0.001$] and pseudo-scientific topics [log-transformed; $F(3, 148) = 1.05$, $p = 0.374$, $R^2 = 0.02$] were non-significant.

Belief change in the face of belief-contradictory evidence (hypothesis 3)

The mixed ANOVA for belief change yielded statistically significant main effects of evidence condition [$F(1, 149) = 115.77$, $p < 0.001$, $\eta^2 = 0.35$] and prior belief [$F(1, 149) = 47.63$, $p < 0.001$, $\eta^2 = 0.09$], as well as a statistically significant interaction [$F(1, 149) = 229.56$, $p < 0.001$, $\eta^2 = 0.33$] (**Figure 4**). Follow-up *t*-tests for dependent samples indicated that participants changed their beliefs in the direction of the evidence presented. Participants who had read confirming evidence increased their belief in the effectiveness of GR, $t(60) = 8.31$, $p < 0.001$, $d = 1.06$, and vice versa, $t(90) = 13.9$, $p < 0.001$, $d = 1.47$.

Conclusion

Study 2 provided mixed findings. In contrast to Study 1, it did not confirm the effect of topic-specific devaluation of science. Further, the findings from Study 2 did not indicate that participants tended to decrease their preferences for scientific sources and we did not observe any generalization to the study of other topics. The latter finding is in line with Study 1 and Thomm et al. (2021a). Regarding the perceived pertinence of science, a comparable pattern emerged: Despite being confronted with belief-challenging evidence, preservice teachers did not devalue the pertinence of science systematically. Finally, as in these prior studies, there was evidence of belief change in the direction of the presented evidence.

Overall, the results of Study 2 suggest that there was no systematic devaluation of science, thus raising the question of how stable and far-reaching the effects of belief-evidence

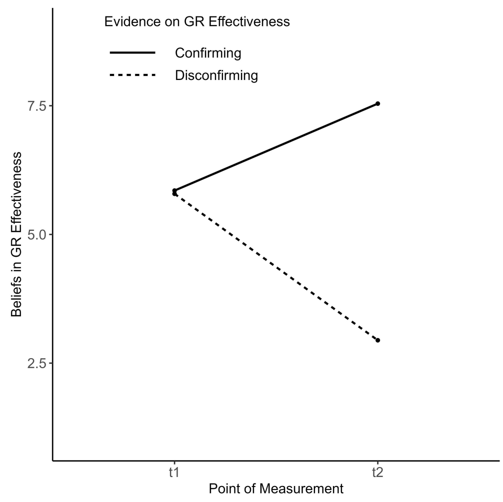


FIGURE 4
Belief change regarding the effectiveness of grade retention (Study 2).

conflicts on science devaluation are. However, as elaborated in more detail below, the discrepant pattern of results in Study 2 may be due to differences in the newly developed abstracts. The more heterogeneous evidence provided in the abstracts might have seemed less conclusive to the participants and left them with more opportunity to argue away the evidence.

General discussion

Discussion

Understanding the mechanisms of science devaluation is of great societal importance (Chinn and Brewer, 1998; Lewandowsky and Oberauer, 2016; Britt et al., 2019; Hornsey, 2020; Kienhues et al., 2020). The present study addressed such a mechanism, the SIE, that occurs when individuals encounter belief-threatening scientific evidence (Munro, 2010). As indicated above, this may frequently occur with educational topics (Asberger et al., 2021; Thomm et al., 2021a). Moreover, tendencies to devalue knowledge from educational research seem to be already prevalent in preservice teachers (van Schaik et al., 2018). To advance our understanding of the SIE and the related science devaluation, Study 1 replicated the preliminary findings from Thomm et al. (2021a). In Study 2, we aimed to draw a more detailed picture by distinguishing the effects on appraisals of the potency and the pertinence of science. To integrate the results, **Table 3** provides an overview of how the findings from the two present studies relate to prior research by Munro (2010) and Thomm et al. (2021a). The below discussion follows the order of the main hypotheses around the SIE, that is, the effects on the devaluation of science's potency and sources (H1a–H1c), the generalization of devaluation (H2a–H2b), and the resistance to belief change (H3). Subsequently, we address the findings on pertinence from Study 2.

Devaluation of science's potency and sources. The core hypothesis associated with the SIE is that people tend to devalue the potency of science to address a scientific issue when facing strong belief-threatening evidence. In addition to the direct assessment of the topic-related potency (H1a), source preferences (H1b), and choice (H1c) are indicators of such devaluation. The three studies that implemented the original paradigm (i.e., Munro, 2010; Thomm et al., 2021a; Study 1) delivered corroborating evidence on these hypotheses. One apparent discrepancy is that Study 1 identified only a descriptive tendency regarding the effects on source choice (i.e., non-significant overall model despite a significant interaction). This result needs to be contextualized in the mostly small effect sizes identified in the existing studies. Generally, the effects of encountering a single belief-evidence conflict on science devaluation seem to be small when judged by conventional rules of thumb. That being said, Thomm et al. (2021a) argue that these small effects should not be underestimated because, so far, we know little about the accumulative effects of repeated conflict experiences. In summary, despite the mentioned constraints, Study 1 indicates, in concert with prior research, that the SIE is a valid mechanism of science devaluation.

In contrast to these findings, Study 2 did not confirm the hypothesized devaluation tendencies. As hinted above, we are of the opinion that this discrepancy is most likely due to the changes in the materials applied in Study 2 (though other explanations are possible, such as chance, small/instable effects, or the interference of an additional pertinence assessment). Whereas, in the former studies, participants received methodologically strong and consistent evidence, the abstracts in Study 2 offered participants more opportunities to challenge the evidence itself. As Munro (2010) suggested, employment of the SIE may be more likely when there is very little opportunity to dismiss the validity of the evidence. Thus, the participants in Study 2 may have been able to reduce the cognitive conflict in other ways, rather than turning to the SIE. In this light, one might argue that changing the experimental materials in Study 2 was unfortunate because this may have diminished the likelihood to observe the effect of interest. We would respond, however, by saying that these materials more closely resembled the results teachers might encounter from a real literature search. The results from (educational) research rarely point unanimously toward a consistent answer. In contrast, though from this point of view the materials used in Study 1 and Thomm et al. (2021a) may seem somewhat artificial, they also represent the evidence teachers might encounter in real life. For example, journalistic media reports frequently present research findings as conclusive and consistent, without

TABLE 3 Comparison of findings on hypotheses about effects of the scientific impotence excuse.

Study	Increased doubt on potency of science for specific topic (H1a)	Decreased preference for scientific sources (H1b)	Decreased probability for choice of scientific sources (H1c)	Increased doubt on potency of science for other topics		Belief maintenance (H3)
				Related (H2a)	Unrelated (H2b)	
Munro (2010)	Yes	Yes	Yes	–	Yes	Yes
Thomm et al. (2021a)	Yes	Yes	Yes	No	No	No
Study 1 (replication)	Yes	Yes	No	No	No	No
Study 2 (follow-up)†	No*	No	No	No*	No	No

†Only study that provided participants with evidence presenting heterogeneous effects regarding the topic; *findings are identical for pertinence assessments.

communicating uncertainty (e.g., van der Bles et al., 2020). Hence, the different abstracts used in Studies 1 and 2 may pertain to different situations in which individuals encounter research. The findings from Study 2 inspire further research to evaluate more closely the *conditions* under which individuals are inclined to employ the SIE over other potential responses (Chinn and Brewer, 1998). Specifically, future studies might examine the effects of the presence and types of (a) methodological information contained in the abstracts and (b) information about the uncertainty associated with the evidence (e.g., different numerical or verbal formats for communicating uncertainty; van der Bles et al., 2020). Using such features to vary the strength and conclusiveness of the evidence would permit the creation of settings that should be differentially prone to elicit the SIE. Next to contributing to a better understanding of science devaluation, such studies could also make a valuable contribution to enhancing science communication (Kienhues et al., 2020).

Generalization to further related and unrelated topics. In contrast to Munro (2010), neither Thomm et al. (2021a) nor the present experiments delivered any evidence of generalization of devaluation to other related (i.e., educational) and unrelated (i.e., medical and pseudo-scientific) topics. The reasoning behind the respective hypotheses was that, if people discount the potency of science to deliver knowledge on a topic, this doubt would likely extend to similar topics, if not to all of (empirical) science. However, it appears that the generalization effects identified by Munro (2010) do not replicate, at least in applications with educational topics. Hence, we conclude that devaluation may not transfer easily, at least from single belief-evidence conflicts. This might be good news, at first glance. However, the experiences of belief-evidence conflicts may be aggravated by additional factors, such as tensions between the information and the individual's social identity. For example, Nauroth et al. (2014) found that computer gamers devaluated the scientific evidence on the negative effects of gaming that threatened their social identity. Similar effects may occur for teachers when the evidence contradicts not only their topic-related beliefs but also their professional identities or the values and practices of their community of practice. These issues should be addressed in further studies.

Resistance to belief change. Unlike Munro (2010), Thomm et al. (2021a) and both present studies found belief change in the direction of the evidence read with sizeable effects. This occurred even though beliefs are notoriously difficult to change (Richardson, 1996; Lewandowsky et al., 2012; Swire et al., 2017; Menz et al., 2021b). This belief change may seem surprising in light of the observed tendencies to devalue science. Two issues arise in this regard. First, the question regarding how the belief change might have been initiated can be answered by drawing upon the Knowledge Revision Components Framework–Multiple Documents (KReC-MD; Butterfuss and Kendeou, 2021). Though not explicitly designed that way, the studies' materials may have served the central principles of knowledge revision. For example, the introductory text on GR and the subsequent study abstracts may have simultaneously activated and contrasted prior and new knowledge and, thus, served the KReC-MD's principles of co-activation and competing activation. Second, one might wonder about the depth and stability of the belief change. The current data provide no evidence on this, unfortunately. While the result might reflect a true belief revision, it might as well be an instance of what Chinn and Brewer (1993, 1998) call *peripheral theory change*. That is, individuals might have provisionally changed their espoused beliefs while still preserving the original one. According to Thomm et al. (2021a), peripheral theory change is one explanation for the seeming contradiction between simultaneous science devaluation, as found in Study 1, and belief change. Being confronted with five pieces of fully (Study 1; Thomm et al., 2021a) or quite (Study 2) consistent evidence that covered various educational contexts and methodological approaches might have been strongly persuasive to participants. Hence, in the post assessment, they may have reported what they *should* believe according to science without actually believing it, as expressed in their devaluation of science's potency. Another explanation is that a true belief change may have occurred, but under epistemic vigilance (cf. Sperber et al., 2010). That is, participants might not have changed their beliefs blindly but added a doubt over science as a cognitive marker that the experienced epistemic conflict had not been solved satisfactorily (Thomm et al., 2021a). Despite these open questions, the finding that reading multiple science texts can initiate belief change may be promising. As a potential implication for teacher education, it may be helpful to present multiple sources of evidence to back up positions that may conflict with students' prior beliefs. Moreover, teacher educators might address rejection strategies like the SIE explicitly to make students aware of problematic reactions to belief-inconsistent information.

Devaluation of science's pertinence. Study 2 did not deliver any evidence of SIE-related effects on the participants' appraisal of the pertinence of educational science to investigate GR effectiveness or further educational topics. That is, the experiences of belief-evidence conflicts do not seem to raise preservice teachers' doubts regarding educational research as a relevant societal institution for contributing knowledge about educational issues. This result may be seen as reassuring. However, because pertinence was investigated only in Study 2, we do not know whether the effects on pertinence would have occurred had participants faced the more consistent evidence materials from Study 1. The limitations regarding the Study 2 materials discussed above may apply to pertinence, too. Moreover, as aforementioned, we cannot rule out that the potency and pertinence assessments interfered with each other. Hence, despite the demonstrated null effects, we suggest that the conditions of pertinence devaluation (as well as its relation to potency devaluation) are a worthwhile subject for further investigation.

Limitations

Beyond the issues discussed already, we acknowledge the following limitations. First, both studies evaluated science devaluation by examining the exemplary educational topic of GR effectiveness. Investigating the generalizability of the results to other educational topics would be warranted. Second, applying the preregistered exclusion criteria led to a substantial reduction in both studies' sample sizes. Though unfortunate, this is a common problem in online research. The resulting statistical power was still sufficiently high, but since the effects of the SIE seem to be small, future studies should use larger samples. Third, the minimum reading time criterion may have disadvantaged fast readers. In the same vein, though excluding participants without sufficient exposure to the treatment is reasonable, time criteria are always somewhat arbitrary. Latent class analysis of time on task might provide a more principled approach for this purpose (Bauer, 2022). Finally, the reliabilities of the source preference scales proved quite low in Study 2, as compared with Study 1 and Thomm et al. (2021a). Low reliabilities may have led to less precise and attenuated estimates of the regression coefficients.

Conclusion

Despite the discussed limitations, our contribution advances prior research to gain a deeper understanding of the mechanisms of science devaluation and, more generally, preservice teachers' attitudes toward and interactions with scientific evidence (van Schaik et al., 2018; Thomm et al., 2021b; Ferguson et al., 2022). Study 1 provided an overall successful direct replication of earlier findings regarding SIE-related science devaluation. Though Study 2 delivered no indication of science devaluation, it supported Munro's (2010) assumptions about the conditions under which the SIE occurs. Interestingly, in line with Thomm et al. (2021a), both studies hint that preservice teachers may have overall favorable attitudes toward science that, however, may be damaged under certain conditions. Moreover, we added evidence that belief change may be initiated (at least provisionally) by confronting preservice teachers with multiple pieces of scientific evidence, even when the texts do not fully satisfy the principles proposed in the refutation and knowledge revision literature (Tippett, 2010; Kendeou et al., 2014; Butterfuss and Kendeou, 2021).

Regarding practical implications, our findings raise the question of how teacher education can help to mitigate potential devaluation mechanisms. To this end, teacher educators need to explicitly address (preservice) teachers' prior beliefs about course-related topics that can shape how they interact with the research-based knowledge they are required to learn (Fives and Buehl, 2012; Asberger et al., 2021; Menz et al., 2021b). Making participants aware of the likelihood of conflicts between their prior beliefs and course contents, as well as of typical devaluative responses to such conflicts, may be effective preemptive measures in advance of learning. By informing participants about cognitive biases and potential devaluation mechanisms in advance, it could help them to expect and resolve the conflicts that sometimes arise, as well as to foster an open attitude toward conflicting information. Such techniques of inoculation, related refutation, and debunking have been found effective in other contexts and can be adopted by teacher educators (Kendeou et al., 2014; Cook et al., 2017; Lewandowsky and van der Linden, 2021; Pieschl et al., 2021). However, this requires equipping them with the knowledge of how to implement such methods in their lectures.

Data availability statement

The original contributions presented in this study are publicly available. This data can be found here: https://osf.io/xz8s2/.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the University of Erfurt (No. 20211123). The patients/participants provided their written informed consent to participate in this study.

Author contributions

HF contributed to the study conceptualization, methodology, project administration, data collection and curation, analysis and visualization, interpretation of results, and original draft of the manuscript. ET and JB contributed to the study conceptualization, methodology, supervision, interpretation of results, and reviewing and editing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was conducted using regular budget funds from the University of Erfurt allocated to JB's chair.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Asberger, J., Thomm, E., and Bauer, J. (2021). On predictors of misconceptions about educational topics: A case of topic specificity. *PLoS One* 16:e0259878. doi: 10.1371/journal.pone.0259878

Bauer, J. (2022). "A primer to latent profile and latent class analysis," in *Methods for Researching Professional Learning and Development: Challenges, Applications and Empirical illustrations*, eds M. Goller, E. Kyndt, S. Paloniemi, and C. Damsa (Cham: Springer International Publishing), 243–268. doi: 10.1007/978-3-031-08518-5_11

Bauer, J., and Prenzel, M. (2012). Science education. European teacher training reforms. *Science* 336, 1642–1643. doi: 10.1126/science.1218387

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educ. Res.* 31, 18–20. doi: 10.3102/0013189x031008018

Bråten, I., and Ferguson, L. E. (2015). Beliefs about sources of knowledge predict motivation for learning in teacher education. *Teach. Teach. Educ.* 50, 13–23. doi: 10.1016/j.tate.2015.04.003

Britt, M. A., Rouet, J. -F., Blaum, D., and Millis, K. (2019). A reasoned approach to dealing with fake news. *Policy Insights Behav. Brain Sci.* 6, 94–101. doi: 10.1177/2372732218814855

Bromme, R., and Gierth, L. (2021). "Rationality and the public understanding of science," in *The Handbook of Rationality*, eds M. Knauff and W. Spohn (Cambridge, MA: The MIT Press), 767–776.

Bromme, R., and Thomm, E. (2016). Knowing who knows: Laypersons' capabilities to judge experts' pertinence for science topics. *Cogn. Sci.* 40, 241–252. doi: 10.1111/cogs.12252

Butterfuss, R., and Kendeou, P. (2021). Krec-MD: Knowledge revision with multiple documents. *Educ. Psychol. Rev.* 33, 1475–1497. doi: 10.1007/s10648-021-09603-y

Cain, T. (2017). Denial, opposition, rejection or dissent: Why do teachers contest research evidence? *Res. Pap. Educ.* 32, 611–625. doi: 10.1080/02671522.2016.1225807

Calderhead, J., and Robson, M. (1991). Images of teaching: Student teachers' early conceptions of classroom practice. *Teach. Teach. Educ.* 7, 1–8. doi: 10.1016/0742-051X(91)90053-R

Chinn, C. A., and Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Rev. Educ. Res.* 63, 1–49. doi: 10.3102/00346543063001001

Chinn, C. A., and Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *J. Res. Sci. Teach.* 35, 623–654. doi: 10.1002/(SICI)1098-2736(199808)35:6<623:AID-TEA3>3.0.CO;2-O

Cook, J., Lewandowsky, S., and Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence. *PLoS One* 12:e0175799. doi: 10.1371/journal.pone.0175799

Dagenais, C., Lysenko, L., Abrami, P. C., Bernard, R. M., Ramde, J., and Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: A review of empirical research. *Evid. Policy* 8, 285–309. doi: 10.1332/174426412X654031

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., and McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educ. Res.* 47, 235–245. doi: 10.3102/0013189X18761042

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. -G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Ferguson, L. E., Bråten, I., Skibsted Jensen, M., and Andreassen, U. R. (2022). A longitudinal mixed methods study of Norwegian preservice teachers' beliefs about sources of teaching knowledge and motivation to learn from theory and practice. *J. Teach. Educ.* doi: 10.1177/00224871221105813 [Epub ahead of print].

Festinger, L. (1957). *A Theory of Cognitive Dissonance (vol. 2)*. Redwood City, CA: Stanford University Press

Field, A., Miles, J., and Field, Z. (2013). *Discovering Statistics Using R*. Thousand Oaks, CA: SAGE.

Fives, H., and Buehl, M. M. (2012). "Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us?," in *APA Educational Psychology Handbook, Vol 2: Individual Differences and Cultural and Contextual Factors*, eds K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, and M. Zeidner (Washington, DC: American Psychological Association), 471–499.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*, 3rd Edn. Los Angeles, CA: SAGE.

Hammersley, M. (2013). *The Myth of Research-Based Policy and Practice*. London: SAGE.

Hayes, A. F. (2022). "Introduction to mediation, moderation, and conditional process analysis: A regression-based approach," in *Methodology in the Social Sciences*, 3rd Edn, (New York, NY: The Guilford Press).

Hornsey, M. J. (2020). Why facts are not enough: Understanding and managing the motivated rejection of science. *Curr. Dir. Psychol. Sci.* 29, 583–591. doi: 10.1177/0963721420969364

Kendeou, P., Walsh, E. K., Smith, E. R., and O'Brien, E. J. (2014). Knowledge revision processes in refutation texts. *Discourse Process.* 51, 374–397. doi: 10.1080/0163853X.2014.913961

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift Für Pädagogische Psychologie* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kienhues, D., Jucks, R., and Bromme, R. (2020). Sealing the gateways for post-truthism: Reestablishing the epistemic authority of science. *Educ. Psychol.* 55, 144–154. doi: 10.1080/00461520.2020.1784012

Kitcher, P. (2011). Public knowledge and its discontents. *Theory Res. Educ.* 9, 103–124. doi: 10.1177/1477878511409618

Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.* 108, 480–498. doi: 10.1037/0033-2909.108.3.480

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Public Interest* 13, 106–131. doi: 10.1177/1529100612451018

Lewandowsky, S., and Oberauer, K. (2016). Motivated rejection of science. *Curr. Dir. Psychol. Sci.* 25, 217–222. doi: 10.1177/0963721416654436

Lewandowsky, S., and van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *Eur. Rev. Soc. Psychol.* 32, 348–384. doi: 10.1080/10463283.2021.1876983

Lysenko, L., Abrami, P. C., Bernard, R. M., Dagenais, C., and Janosz, M. (2014). Educational research in educational practice: Predictors of use. *Can. J. Educ.* 37, 1–26.

McIntyre, D. (2005). Bridging the gap between research and practice. *Camb. J. Educ.* 35, 357–382. doi: 10.1080/03057640500319065

Menz, C., Spinath, B., and Seifried, E. (2021a). Where do pre-service teachers' educational psychological misconceptions come from? *Zeitschrift Für Pädagogische Psychologie* 35, 143–156. doi: 10.1024/1010-0652/a000299

Menz, C., Spinath, B., and Seifried, E. (2021b). Misconceptions die hard: Prevalence and reduction of wrong beliefs in topics from educational psychology among preservice teachers. *Eur. J. Psychol. Educ.* 36, 477–494. doi: 10.1007/s10212-020-00474-5

Munro, G. D. (2010). The scientific impotence excuse: Discounting belief-threatening scientific abstracts. *J. Appl. Soc. Sci.* 40, 579–600. doi: 10.1111/j.1559-1816.2010.00588.x

Nauroth, P., Gollwitzer, M., Bender, J., and Rothmund, T. (2014). Gamers against science: The case of the violent video games debate. *Eur. J. Soc. Psychol.* 44, 104–116. doi: 10.1002/ejsp.1998

Nisbet, E. C., Cooper, K. E., and Garrett, R. K. (2015). The partisan brain. *Ann. Am. Acad. Pol. Soc. Sci.* 658, 36–66. doi: 10.1177/0002716214555474

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Rev. Educ. Res.* 62, 307–332. doi: 10.3102/00346543062003307

Pieschl, S., Budd, J., Thomm, E., and Archer, J. (2021). Effects of raising student teachers' metacognitive awareness of their educational psychological misconceptions. *Psychol. Learn. Teach.* 20, 214–235. doi: 10.1177/1475725721996223

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Richardson, V. (1996). "The role of attitudes and beliefs in learning to teach," in *Handbook of Research on Teacher Education*, 2nd Edn, eds J. Sikula, T. J. Buttery, and E. Guyton (New York, NY: Macmillan), 102–119.

Rousseau, D. M., and Gunia, B. C. (2016). Evidence-based practice: The psychology of EBP implementation. *Annu. Rev. Psychol.* 67, 667–692. doi: 10.1146/annurev-psych-122414-033336

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* 13, 90–100. doi: 10.1037/a0015108

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., et al. (2010). Epistemic vigilance. *Mind Lang.* 25, 359–393. doi: 10.1111/j.1468-0017.2010.01394.x

Swire, B., Ecker, U. K. H., and Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *J. Exp. Psychol. Learn. Mem. Cogn.* 43, 1948–1961. doi: 10.1037/xlm0000422

Tabachnick, B. G., and Fidell, L. S. (2014). *Using Multivariate Statistics (6th ed., Pearson New International Edition)*. Harlow: Pearson.

Thomm, E., Gold, B., Betsch, T., and Bauer, J. (2021a). When preservice teachers' prior beliefs contradict evidence from educational research. *Br. J. Educ. Psychol.* 91, 1055–1072. doi: 10.1111/bjep.12407

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021b). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Zeitschrift Für Pädagogische Psychologie* 35, 173–184. doi: 10.1024/1010-0652/a000301

Tippett, C. D. (2010). Refutation text in science education: A review of two decades of research. *Int. J. Sci. Math. Educ.* 8, 951–970. doi: 10.1007/s10763-010-9203-x

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., and Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7672–7683. doi: 10.1073/pnas.1913678117

van Schaik, P., Volman, M., Admiraal, W., and Schenke, W. (2018). Barriers and conditions for teachers' utilisation of academic knowledge. *Int. J. Educ. Res.* 90, 50–63. doi: 10.1016/j.ijer.2018.05.003

Winch, C., Oancea, A., and Orchard, J. (2015). The contribution of educational research to teachers' professional learning: Philosophical understandings. *Oxf. Rev. Educ.* 41, 202–216. doi: 10.1080/03054985.2015.1017406

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 26 October 2022
DOI 10.3389/feduc.2022.975105

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University, Germany

REVIEWED BY
Friederike Hendriks,
Technische Universitat Braunschweig, Germany
Eric Klopp,
Saarland University, Germany
Tina Seidel,
Technical University of Munich, Germany

*CORRESPONDENCE
Leila E. Ferguson
leila.ferguson@kristiania.no

SPECIALTY SECTION
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

RECEIVED 21 June 2022
ACCEPTED 05 October 2022
PUBLISHED 26 October 2022

CITATION
Ferguson LE and Bråten I (2022) Unpacking pre-service teachers' beliefs and reasoning about student ability, sources of teaching knowledge, and teacher-efficacy: A scenario-based approach. *Front. Educ.* 7:975105. doi: 10.3389/feduc.2022.975105

Unpacking pre-service teachers' beliefs and reasoning about student ability, sources of teaching knowledge, and teacher-efficacy: A scenario-based approach

Leila E. Ferguson [1]* and Ivar Bråten [2]

[1]School of Health Sciences, Kristiania University College, Oslo, Norway, [2]Department of Education, University of Oslo, Oslo, Norway

The beliefs teachers hold may provide information about their more or less evidence-informed reasoning about educational issues. However, gaining a clear picture of teachers' beliefs has proven difficult. A promising line of inquiry uses scenario-based approaches to assess teachers' enacted beliefs. Accordingly, we assessed 75 Norwegian pre-service teachers' beliefs about student ability, sources of teaching knowledge, and teacher efficacy by analyzing their written responses to authentic classroom scenarios, with these responses also providing information about participants' reasoning about the scenarios. While participants' responses seemed to be evidence-informed in many ways, there were also indications of the opposite, such as limited consideration of educational research in pedagogical decision-making. The results contribute uniquely to an understanding of pre-service teachers' beliefs and reasoning about educational issues. As such, they may help researchers and teacher educators to better understand the beliefs pre-service teachers hold, as well as to facilitate further development of these beliefs. Implications for future research and teacher education are discussed.

KEYWORDS

teacher beliefs, teacher education, pre-service teachers, scenario-based assessment, evidence-informed reasoning

Introduction

On which information pre- and in-service teachers base their decisions and actions is an important question for teacher education, as well as for the society at large. Exemplary teacher education programs are theoretically-rather than craft-oriented and grounded in the complexities of modern teaching (Kitchen and Petrarca, 2016). At the same time, pre-service teachers in the information society have more access to scientific, research-based knowledge than their predecessors. Still, empirical studies have clearly documented pre- as well as in-service teachers' preference for informal or experience-based over research- or evidence-informed sources of teaching knowledge (Thomm et al., 2021c; Ferguson et al., 2022). Possible reasons for this dilemma include lacking competency or efficacy in using and reasoning about research-based knowledge, or even unwillingness to do so (Thomm et al., 2021b). Teacher educators meet students who have vast experiences as students in classrooms. Based on those experiences, pre-service teachers in initial teacher education programs may hold well-developed, potentially powerful belief systems about education that are incompatible with established bodies of knowledge from relevant fields, such as the educational and learning sciences (Menz et al., 2021). Further, these beliefs are likely to influence their interactions with educational theories and research (Ferguson et al., 2022).

Reproduction of teaching practices primarily grounded in own experiences may be undesirable, detrimental even, for student learning and development (Csandi et al., 2021), as well as being inconsistent with current national and international policies that recommend evidence-informed practice where reasonable (European Commission, 2007). Thus, there have been several investigations into teachers' evidence-informed reasoning (or the lack of it) about educational problems (Csandi et al., 2021; Zimmermann and Mayweg-Paus, 2021), with evidence-informed reasoning referring to taking relevant theory and research into consideration when making decisions and taking action to solve educational problems (Ferguson, 2021). In this study, we wanted to explore this problem space from a teacher beliefs perspective. Inspired by a broad framework of teacher beliefs (Fives and Buehl, 2008; Fives et al., 2019) and state-of-the-art methods for investigating beliefs (Bullough, 2015; Sabatini et al., 2018; Lunn Brownlee et al., 2021), we designed and tested an innovative scenario-based problem-solving approach in addressing how pre-service teachers enacted their beliefs about student ability, sources of teaching knowledge, and teacher efficacy, also looking for signs of evidence-informed reasoning in their written responses to those scenarios. Of note is that our focus on these three types of beliefs (i.e., beliefs about student ability, sources of teaching knowledge, and teacher efficacy) was based on the empirically grounded framework of teacher beliefs discussed by Fives and Buehl (2008, 2016), Fives et al. (2019). As parts of a belief system (Fives and Buehl, 2008), such beliefs may be more or less congruent, with beliefs in the malleability of student ability, formalized sources of teaching knowledge (i.e., educational theory and research), and their own ability to support learning for all students potentially indicating a congruent system of teacher beliefs. This study did not aim to investigate relations between beliefs about student ability, sources of teaching knowledge, and teacher efficacy, however. And, while we also did not aim to investigate relations between such teacher beliefs and evidence-informed reasoning, the scenario-based approach to studying beliefs about educational issues that we used may act as a window to exploring both beliefs and evidence-informed reasoning. In what follows, we present a conceptualization of teacher beliefs and relevant work on beliefs about student ability, sources of teaching knowledge, and teacher-efficacy. Before we present our scenario-based study and discuss the implications of our findings for teacher educators and educational researchers, we briefly discuss a scenario-based approach to studying teacher beliefs.

Teacher beliefs

Teacher beliefs are individual interpretations and experience-based propositions held by teachers that influence their teaching (Fives et al., 2019; Ferguson and Lunn Brownlee, 2021). Given teachers' long apprenticeship of observation (Lortie, 1975) and, not least, participation in educational settings, their beliefs are likely to be rather deeply entrenched, and may act as a barrier to acquisition of new knowledge in the form of theories and research in teacher education (Guilfoyle et al., 2020; Menz et al., 2021). Following Fives and Buehl (2008), we view teacher beliefs as parts of a belief system that gives meaning to their interactions, intentions, and actions (Buehl and Beck, 2015; Buehl and Fives, 2016; Dweck and Molden, 2017). Broadly speaking, teacher belief researchers have focused on how pre-service teachers and teachers frame and comprehend their experiences, for example, how they interpret students' behavior and academic performance, plan and adapt their teaching, and perceive themselves as teachers. Presumably, such beliefs will also influence the sources of teaching knowledge they choose to engage with, whether in terms of educational research literature or respected practice teachers and colleagues (Tschannen-Moran and Hoy, 2001; Fives and Buehl, 2008). Some beliefs that teachers hold can be termed misconceptions or misinformation beliefs because they indicate reliance on incorrect information or lack of evidence, whereas other teacher beliefs can be termed accurate because they indicate reliance on correct information or evidence (Ecker et al., 2022).

Teachers' beliefs about student ability

A well-known example of the potential role of teachers' beliefs about student ability is the original Pygmalion effect study by Rosenthal and Jacobsen (1968). In that study, teachers were led to believe that a proportion of the young students in their classrooms were so-called "late-bloomers," as identified by a fictitious diagnostic test, and could therefore be expected to experience higher cognitive growth and resulting gains in IQ scores in the future, a prediction that was confirmed and upheld in follow-up studies. The results of this well-known, yet oft miscited study, have been used to highlight the importance of teachers' views of their students and the students' chances of future success (Jussim and Harber, 2005). Proposed explanations for the actual, resulting differences in performance by alleged late-bloomers and the control group included differences in amounts and types of feedback, emotional support, and availability in terms of time spent with the different students, as well as the provision of suitably challenging opportunities for growth for those who were identified as "late-bloomers" (Jussim, 1986).

Such effects also have resonance in the work of Dweck et al. (e.g., Dweck and Leggett, 1988; Dweck, 2000; Yeager and Dweck, 2012; Dweck and Molden, 2017). In essence, Dweck's meaning system theory holds that beliefs about the malleability or stability of human attributes, such as ability or intelligence, give meaning to situations in which those attributes are involved. More specifically, viewing ability as malleable is considered more adaptive in learning and achievement settings because it can lead to mastery goals, strategic effort, and persistence and ingenuity in the face of challenge or setback (Dweck, 2000; Dweck and Molden, 2017). For teachers, their views of students' ability are also likely to influence their perceptions and behaviors in terms of the goals and ambitions they hold for the students and how they interpret student behavior in the classroom (Hattie, 2012). That is, teachers viewing student ability as malleable may be more likely support learning for all students and attribute students' successes and failures to their own teaching, among other factors, rather than to stable attributes of the students. On the other hand, teachers viewing student ability as stable, may be more likely to attribute different levels of academic performance to different levels of "smartness," and thus harbor higher ambitions for the smarter students. Accordingly, empirical studies focusing on teachers' beliefs about student performance (Jonsson et al., 2012; Patterson et al., 2016) have suggested that teachers who attribute student performance to underlying, stable ability are more likely to hold stereotypical views of students (Jonsson and Beach, 2012), for example, as "smart," "bright," "lazy," or "ungifted," which, in turn, influence the teaching practices these teachers tend to engage with their students (Patterson et al., 2016).

Patterson et al. (2016), who examined 53 American pre- and in-service teachers' beliefs about factors influencing student academic performance using a quantitative, survey-based approach, identified distinct factors related to the school (e.g., the school culture), the family (e.g., parents' income and education), and the students themselves (e.g., their intelligence). These authors argued that the relative weight of teachers' beliefs about the importance of the different factors would contribute to differences in their approaches to students, for example, in terms of teacher effort, instructional methods, and interactions.

Teacher beliefs about sources of teaching knowledge

Beliefs about sources of knowledge and, in turn, how individuals select, evaluate, and use such sources are essential epistemic questions with particular relevance to teacher beliefs and reasoning, given the perennial debates on the nature of teaching knowledge and valid sources of teaching knowledge (Shulman, 1987; Buehl and Fives, 2009; Thomm et al., 2021c). Teachers' beliefs about sources of teaching knowledge may be considered a domain-specific form of source of knowledge beliefs (Buehl and Alexander, 2001; Guilfoyle et al., 2020), which will likely influence the information teachers chose to engage with and use.

While beliefs about sources of knowledge traditionally have been conceptualized as falling on a continuum from reliance on external authority to personal construction of knowledge (Schommer, 1990; Hofer and Pintrich, 1997), more recent conceptualizations have highlighted the importance of testimony from external sources in advanced, scientific reasoning (Chinn et al., 2011). In terms of teaching, valuable testimony may come from both collegial and scientific sources. Thus, teachers may come to rely on a mix of craft, experience-based and theoretical, research-based knowledge sources, which makes their resulting teaching knowledge somewhat personalized in the sense that it is influenced by teachers' life experiences in addition to recognized bodies of knowledge from different fields (Shulman, 1987; European Commission, 2007).

Buehl and Fives (2009) used a mixed-methods approach to study beliefs about sources of teaching knowledge among 110 pre- and in-service teachers who responded to open-ended questions such as "where does knowledge of how to teach come from?" The authors identified the following six main themes related to sources of teaching knowledge in participants' responses: formal education in terms of pre-service education and professional development; formalized bodies of knowledge, including research articles and the Internet; observational (or vicarious) learning from other teachers; collaboration and shared meaning making; enactive experiences from personal, professional, and other experiences; and self-reflection and synthesis of information and experiences. In accordance with this study, pre-service teachers' and teachers' emphasis on personal, enactive, and experienced-based knowledge has been confirmed in later work (e.g., Bråten and Ferguson, 2015; Kiemer and Kollar, 2021; Thomm et al., 2021c; Ferguson et al., 2022). Still, the mechanisms explaining this emphasis have only recently become the object of more systematic investigation. For example, Hendriks et al. (2021) found an interaction between pre-service teachers' epistemic aims (gaining insights into educational research vs. receiving practically applicable knowledge), which were experimentally manipulated, and how they judged the expertise of researchers vs. teachers. Specifically, when the aim was theoretical explanations of education, the pre-service teachers ascribed more expertise to researchers than to teachers, but when the aim was to gain more practical knowledge, they ascribed more expertise to experienced teachers than to researchers. Further, pre-service teachers who perceived educational research to be more useful were also more likely to ascribe higher expertise to researchers.

Perceived irrelevance of educational research among pre-service teachers and teachers may highlight the need for teacher educators to make their own evidence-based practice more explicit (Ferguson, 2021) and try to foster beliefs about educational research as ways of considering one's own practice in a different light (Guilfoyle et al., 2020). That existing beliefs about sources of knowledge may act as obstacles to engaging with educational evidence is also consistent with findings reported by Thomm et al. (2021a). These authors introduced the idea of motivated reasoning about educational research that contradicted pre-service teachers' existing beliefs, showing that pre-service teachers may be more critical to the usefulness of educational research in reasoning about an educational issue when the evidence is at odds with their own prior beliefs. However, Thomm et al. (2021a) also found that pre-service teachers reported positive views of scientific (i.e., educational sciences) sources.

Teacher-efficacy beliefs

Teacher-efficacy beliefs refer to the beliefs teachers hold about themselves and their ability to perform given tasks, especially supporting learning, engagement, and performance in students (Tschannen-Moran et al., 1998; Tschannen-Moran and Hoy, 2001). Originating in Bandura's (1997) construct of perceived self-efficacy, teacher-efficacy applies to teachers' beliefs about personal or collective efficacy, or a combination of the two (Tschannen-Moran et al., 1998). Further, teacher-efficacy may influence how opportunities and challenges in the teaching environment are perceived, as well as the choice of learning activities, effort, and perseverance (Skaalvik and Skaalvik, 2007), and its influence may even extend to the levels of lesson planning and organization (Tschannen-Moran and Hoy, 2001). Teacher-efficacy is a multi-dimensional and context-specific construct that is based on teachers' prior experiences and their interpretations and attributions (Skaalvik and Skaalvik, 2007). More specifically, its dimensionality includes efficacy for adapting instruction to student needs, motivating and engaging students, managing classrooms and maintaining discipline, cooperating with colleagues and parents, coping with changes and challenges, and influencing student outcomes (Tschannen-Moran and Hoy, 2001; Skaalvik and Skaalvik, 2007, 2014).

Fives and Buehl (2008) drew parallels between teacher-efficacy and lay theories of intelligence (see Section "The present study"), extending views of malleable or stable human attributes to the area of teaching. That is, teachers may hold the belief that teaching ability is a stable attribute they are born with (or not), as opposed to viewing teaching as skills and knowledge that can be learned and further developed, with the latter view possibly being more conducive to engaging with and making use of educational research and theories. Thus, teachers who believe they can make an influence may not only be more likely to invest effort in their teaching, but also to draw on more, and varied sources of teaching knowledge in doing so. Patterson et al. (2016) further suggested that high-efficacy teachers tend to focus on the aspects of teaching situations they can control, which, in turn, may contribute to feelings of efficacy as well as satisfaction.

A scenario-based approach to measuring teacher beliefs

Gaining an understanding of pre-service teachers' beliefs and associated reasoning is important if teacher educators and researchers are to capitalize on the funds of knowledge, or correct the misconceptions or non-availing beliefs, that these fledgling teachers possess. However, measuring the "messy construct" (Pajares, 1992) of teacher beliefs is a tricky business (Tschannen-Moran and Hoy, 2001; Schraw and Olafson, 2015). Earlier studies have employed manifold approaches, with "questionnaires, verbal reports, performance observations, self-reflective writing, tests and exams, vignettes, scales, portfolios, visual representations, and instructional and classroom artifacts" dominated existing assessment strategies in the literature (Schraw and Olafson, 2015, p. 90). A major assessment issue that is particularly troublesome for teacher belief researchers is the unclear relation between reported beliefs and beliefs enacted in practice (Buehl and Beck, 2015). For example, there is some evidence that teachers' espoused beliefs are not present in their enacted practices and that teachers engage in practices they indicate that they do not support (Buehl and Beck, 2015). Further, beliefs may be enacted in different ways depending on context, but such context-specificity may be difficult to capture by quantitative methods such as questionnaires that ask teachers to think about their past or future practice in abstract and general terms (Patterson et al., 2016). Similarly, the provision of pre-conceived answers in multiple-choice measures may provide appealing alternatives that draw attention to desired or ideal responses that are not commensurate with the intricate realties of practice.

On the other hand, qualitative approaches have several advantages in trying to capture messy constructs, including the discovery (rather than testing) of variables and allowance for deeper understanding of complex constructs and relations (Olafson et al., 2015). Well-designed, transparent qualitative methods may therefore offer new insights into pre-service teachers' beliefs. One such approach, which also has been used in other areas of educational psychology research (Hartmann et al., 2021; Lunn Brownlee et al., 2021; Wang et al., 2021), is scenario-based. Basically, a scenario-based approach presents an imagined or hypothesized event or context that establishes a credible purpose for the individual's decisions and actions. As such, it not only provides a framework for the assessment but also represents a step toward more ecological validity (Sabatini et al., 2018). In this study, we aimed to create a series of domain-specific scenarios that were instructionally relevant to the participants and covered the constructs that we targeted. These scenarios introduced hypothetical problematic events related to teaching and learning that required decisions and actions on part of the participants (see Section "Materials" for further description of the content of these scenarios).

Compared to self-reports or questionnaires, the scenario-based approach that we implemented allows for exploration of beliefs at a level closer to enactment, and it may thus be more useful in exploring the context-dependent nature of beliefs (Bullough, 2015). For example, Kiemer and Kollar (2021) recently explored a scenario-based approach to assessment in their investigation of teacher beliefs, asking pre-service teachers to advise a colleague about their actions in an imagined teaching scenario. Accordingly, in this study, we opted for a scenario-based approach as a potentially valid method for exploring pre-service teachers' beliefs about student ability, sources of teaching knowledge, and teacher-efficacy. In doing this, we were inspired by similar attempts to use a scenario-based approach in related fields, such as epistemic beliefs and multiple-text comprehension (Barzilai and Weinstock, 2015; Sabatini et al., 2018; Lunn Brownlee et al., 2021; Wang et al., 2021).

The present study[1]

Building on prior theoretical and empirical work focusing on teacher beliefs, as well as the assumption that practical, problem-based contexts (i.e., scenarios) may provide entry points for considered reflection (Hartmann et al., 2021), we implemented a scenario-based approach in studying pre-service teachers' beliefs about student ability, sources of teaching knowledge, and teacher-efficacy as they reasoned about authentic pedagogical problems. Specifically, we addressed the following research questions:

(1) What beliefs do pre-service teachers hold about student ability, sources of teaching knowledge, and teacher-efficacy, as revealed by a scenario-based approach?
(2) What do participants' responses to the scenarios reveal about their reliance on evidence from educational research to inform their decisions and actions?

1 This study is part of a large longitudinal mixed methods project (Ferguson et al., 2022). However, the research questions, the materials, the analyses, and the results are all unique to this study and not reported elsewhere.

Materials and methods

Participants

Participants were 75 second-year pre-service teachers enrolled in 4-year teacher education programs at a public university college in southeast Norway (49 female, 26 male; M_{age} = 21.10, SD = 4.22). Eighty-nine percent of the participants had Norwegian as their first language and the rest were proficient in Norwegian. Participation was rewarded by entry into a prize draw for one of two gift cards (approx. USD 85) for shopping centers.

The teacher education programs followed national guidelines for teacher education (Norwegian Ministry of Education and Research, 2011) and consisted of 240 European Credit Transfer and Accumulation System (ECTS) credits. Sixty ECTS credits were allocated to Pedagogy and Pupil-Related Skills (PEL) class (often referred to as Education Studies in international literature; e.g., Guilfoyle et al., 2020). The remaining 180 ECTS credits focused on subject-specific (e.g., mathematics, Norwegian language, English as a foreign language) knowledge, skills, and general competences. For more detailed information about these teaching education programs, see Afdal and Spernes (2018).

Materials

Participants were presented with and responded to three scenario-based problem-solving tasks in writing (see **Appendix A** for the exact wording of each scenario). Scenario one described a sixth-grade classroom context in which students' performance on a natural science test varied greatly, and the participants were asked to discuss possible reasons for observed differences among students, with this discussion presumably reflecting their beliefs about students' ability, as well as other factors that might influence performance.

The second, two-part scenario was designed to capture participants' beliefs about sources of teaching knowledge and their reasoning about such sources. Based on the large variation in student performance noted in scenario one, participants were asked (a) to describe sources of knowledge they would use to design a new teaching sequence taking the large differences among students into consideration and (b) to justify their decisions.

Scenario three asked participants to imagine a future situation where they have become the science teacher of a challenging class. Participants were asked to discuss their possibilities to ensure a satisfactory learning outcome for all students in this class, based on their own strengths and weaknesses as a teacher. This task was designed to capture participants' beliefs about teacher-efficacy.

All scenarios were group-administered on paper and participants responded in writing. There were no time limit for reading or responding to the scenarios.

Data analysis

The thematic analysis of the scenario responses was both grounded in the data and informed by the authors' knowledge and interpretation of prior theoretical and empirical work on teacher beliefs, including work on beliefs about student ability, sources of teaching knowledge, and teacher efficacy (e.g., Dweck, 2000; Skaalvik and Skaalvik, 2014; Bråten and Ferguson, 2015; Guilfoyle et al., 2020). Please see Sections "Teachers' beliefs about student ability," "Teacher beliefs about sources of teaching knowledge," "Teacher-efficacy beliefs," for further discussion of the work that informed the authors' thematic analysis of the scenario responses. We applied a three-stage coding process in interacting with the data and relevant literature to explore and elaborate the emergent themes.

First, we studied anonymized participant responses for each scenario. Responses were segmented into units of analysis representing distinct and coherent ideas of varying extent. This means that the idea units could vary from a single word (e.g., the name of a specific source of teaching knowledge, such as "Internet") to a sentence or string of sentences that reflected participants' beliefs or reasoning.

We initially identified 176 idea units in response to the first scenario. Fifty-three (30%) of these idea units were hand coded by both authors, and the first author coded the remaining 123 idea units.

For the second scenario, 133 ideas were identified for the first part of the scenario, targeting participants' beliefs about sources of teaching knowledge. Forty-six (35%) of these were coded by both authors. Regarding the second part of the second scenario, asking participants to justify their choices of sources, 47 idea units representing reasoned justifications were identified, of which 12 (26%) were coded by both authors. Of note is that 30 participants failed to provide justifications for choices of sources but rather described their own teaching plans. While these ideas may reflect participants' beliefs and reasoning, their failure to provide sources for their thinking impaired our ability to make claims about these participants' reasoning and its sources (see limitations in the "Discussion" section).

For the third scenario, we identified 183 idea units referring to ways of teaching the challenging class, 46 idea units referring to strengths that might increase their chances of successfully teaching the new class, and 15 idea units referring to weaknesses that might hinder their success in this regard. Both authors coded 52 (28%), 22 (48%), and five (36%) idea units relating to ways of teaching, strengths, and weaknesses, respectively, and the remaining idea units were coded by the first author, consulting the second author whenever uncertainties were encountered. Please note that while the units of analysis remained intact throughout the data analysis, some of the numbers were altered because categories were merged in the third step of the analysis. Whenever a percentage of the idea units was coded by both authors, the authors collaboratively read, segmented, and categorized those idea units, resolving any disagreements through discussion.

The first step of the data analysis, described above, involved multiple readings of each participant's responses and led to the creation of a set of precursory codes. As such, our preliminary analysis focused on fundamental, yet rather specific themes that emerged from the data, for example, specific ways of interacting with students (e.g., "use a strict tone," "be a clear leader," "show interest in pupils."). In the second step of the analysis, which also was collaborative, we therefore focused on identifying broader themes based on the preliminary analysis (for example, "classroom management," "relational approaches," "variation in teaching methods.").

In the third step of the data analysis, participants' responses were transcribed and imported into NVivo. In NVivo, each of the emergent themes in step 2 became a parent node while the preliminary codes were represented by initial child nodes. Our use of NVivo increased the transparency of the data and allowed for further insight into the contents of each emerging theme, as well as comparison of parent and child nodes across the whole data set. This allowed us to re-examine the emerging themes and how they related to each other in terms of prevalence, as well as to gain more insight into the depths and variations within each theme and possible overlaps. To avoid redundancy, we also merged previously identified themes that were similar in content and formed meaningful sub-themes, for example with respect to types of motivation (e.g., "make the learning materials exciting" and "make teaching interesting" were both coded as "motivating approaches").

Given the nature of the questions, the three focal themes were: (1) teacher beliefs about students' ability, (2) beliefs about sources of teaching knowledge, and (3) beliefs about teacher-efficacy. However, our scenario-based approach and the rich nature of the resulting data also allowed for other emerging themes to be identified and provide information about participants' beliefs and reasoning.

Results

Participants enacted a range of educationally relevant beliefs in response to the three scenarios, including, but not limited to student ability, sources of teaching knowledge, and teacher-efficacy. We were also able to identify signs of (more-or-less) evidence-informed reasoning in their responses. The wording of the scenario-based problems also elicited justifications for participants' choices of sources of teaching knowledge (Scenario 2) and suggestions for a teaching approach based on perceived strengths and weaknesses (Scenario 3). **Tables 1–5** include an overview of the emerging themes and sub-themes for each scenario, as well as illustrative idea units within those themes. In the following, we present an overview of the different emerging themes and also highlight variation within each theme.

Scenario 1

The first scenario focused on possible reasons for variation in student performance on a natural science test, designed to capture participants' beliefs about student ability (Dweck, 2000; Patterson et al., 2016). As can be seen in **Table 1**, we identified five emergent themes that concerned reasons for differences in student performance, which we labeled teaching, individual differences, motivational differences, sociocultural context, and test context. Moreover, each theme consisted of several sub-themes, which we also describe in this section. The most prevalent idea units (n = 70) reflected beliefs about the role of the teacher and aspects of their teaching practice in creating differences in student performance, encompassing the sub-themes of adapted teaching, management of teaching, and variation in teaching methods. Adapted teaching (n = 44) refers to the principle of adapted teaching, an approach which has a central position within the Norwegian egalitarian education system (Norwegian Ministry of Education and Research, 1999). For example, "that I, as a teacher, haven't adapted my teaching well enough," "In this case I, as a teacher, haven't done a good enough job with evaluation, to pinpoint where all my students are in the knowledge acquisition process," and "that the way I taught wasn't directed toward the whole class, but rather toward a smaller group that had more knowledge about the subject. The activities could have been set up on a more individual basis" were ideas falling within the sub-theme of adapted teaching. Management of teaching (n = 15) included ideas about the central role of the teacher in the students' learning process (e.g., "Perhaps one has failed to go through the topic in a thorough manner" and "It might be that the teacher hasn't explained the subject well enough"). Variation in methods (n = 11) refers to engaging in different teaching methods in class, for example, "if one has just been teaching from the blackboard then it may be smart to vary (teaching methods) more" and "Maybe the skewed results are caused by the fact that one has varied teaching methods too little. Maybe one needs to use other teaching methods to reach the whole class."

Beliefs about individual differences were reflected in 53 of the idea units concerning differences in student performance. There were seven sub-themes referring to ability [n = 15; e.g., "there will be a variety of students with different talents and abilities in every class. Therefore, there is no 'unnatural' distribution (of results) here"], social and academic learning difficulties (n = 15; e.g., "there may be trouble at the social level, personal problems, illness etc." and "There may be reading and writing difficulties in this group of students"), concentration (n = 7, e.g., "the differences may be because not everyone has managed to concentrate on the teaching"), time needed to learn (n = 7, e.g., "some people take longer time before they understand the curriculum"), learning styles [n = 5; e.g., "Some students learn best with an auditive learning style, some (are) more visual and some more tactile"], learning strategies [n = 2; e.g., "Many (students) do not know which working methods suit them best and study for a test by reading from the start to the end of the chapter"], and maturity (n = 2; e.g., "the students are at different levels of maturity. They don't develop at the same tempo") as reasons for differences in student performance on the science test.

Beliefs about motivational differences were reflected in 42 idea units. Within this theme, there were six sub-themes: Individual interest, effort, situational interest, general motivation, willingness to learn, and goal-orientation. Individual interest (n = 15) focused on students' levels of intrinsic interest as a reason for differences in performance, for example, "It may also be caused by interest, the five students who have been (i.e., scored) extremely poorly are not so into the topic, while the five who have performed well think the topic is interesting." Effort (n = 13) concerned investment of personal resources in the time preceding the test, for example, "effort before the test" and "It may also be that some of the students didn't "bother" to read - that this is the reason for the poor results." Situational interest (n = 7) is regarded as more contextualized and transitional than is individual interest (Hidi, 2001). From participants' responses, it seems that more emphasis was placed on the teacher's effort in sparking situational interest (i.e., in comparison to individual interest), for example, "whether one (the teacher) has made it interesting enough to engage everyone" and "(the teacher) hasn't awakened enough "nosiness" in their students." General motivation (n = 4) ideas were broad references to "motivation" without any further specification, such as "motivation. . . probably has a lot to say." Willingness to learn (n = 2) were specific statements referring to a will to learn, such as "another factor can also be the students' own will to learn." Finally, goal-orientation (n = 1) concerned (the lack of) students' engagement with learning goals specified by the teacher: "It may also be the case that the students just don't care about achieving the specified goals."

The 21 idea units reflecting beliefs about the sociocultural context could be categorized into the sub-themes of home situation, resources and equipment, and classroom culture. Home situation (n = 14) referred to support, interest, and pressure from the students' families, for example, "Help and being followed up at home also play an important role" and "whether they have parents who can push their children to read extra." Resources and equipment (n = 4) concerned both teaching resources and students' physical placement in the classroom, for example, "This is mainly caused by a lack of time and resources for the individual students" and "equipment, etc.

TABLE 1 Emergent themes, sub-themes, and illustrative idea units based on responses to Scenario 1.

Emergent theme	Sub-themes	Illustrative idea units
Teaching (n = 70)	Adapted teaching (n = 44), management of teaching (n = 15), variation in teaching methods (n = 11)	Other students have perhaps not gotten the adapted teaching that they need, and have therefore not benefited from teaching that has been the same across the board for the whole class (S104); It might be that the teacher hasn't explained the subject well enough (S27); Had the teacher maybe tried to vary the teaching (S71).
Individual differences (n = 53)	Ability (n = 15), learning difficulties (n = 15), concentration (n = 7), time to learn (n = 7), learning styles (n = 5), learning strategies (n = 2), maturity (n = 2)	In every class there are a variety of students with different abilities and starting points (S7); The differences may be because not everyone has managed to concentrate in class (S16); For example, specific reading or writing difficulties (S15); Some may have diagnoses, language difficulties, etc. that stop them from learning so much (S48).
Motivational differences (n = 42)	Individual interest (n = 15), effort (n = 13), situational interest (n = 7), general motivation (n = 4), willingness to learn (n = 2), goal orientation (n = 1)	Motivation is also a big factor here (S43); Students who did badly may have problems with motivation and interest for the topic or subject (S91).
Sociocultural context (n = 21)	Home situation (n = 14), resources and equipment (n = 4), classroom culture (n = 3)	A culture for doing one's best has not been established, and there is no culture for creating a sense of wonder (S9); Help and support at home also play an important role (S34); I think students' placement in the classroom can have a lot to say. Some students perhaps need to sit closer to the board to see. . . (S100).
Test context (n = 15)	Test preparation (n = 8), test difficulty (n = 7)	It's possible that the test wasn't targeted to different levels (of ability) (S32); That the questions on the test were formulated in a difficult way (S16); As teacher I should give notice 1 week in advance so that students can prepare (S21).

Numbers refer to idea units within each emergent theme and sub-theme. S followed by a number refers to a particular participant.

TABLE 2 Emergent themes, sub-themes, and illustrative idea units based on responses to Scenario 2.

Emergent theme	Sub-themes	Illustrative idea units
Informal (n = 58)	Colleagues (n = 31), students (n = 14), own resources (n = 12), family and friends (n = 1)	I would discuss with my colleagues. . . Maybe I can get some colleagues to observe my lessons to see what I can do better and what I have to change (S5); Initially I would ask students how they thought they learned best (S20).
Formal (n = 32)	Textbooks and educational literature (n = 29), teacher education (n = 2), research (n = 1)	I would have checked books from teacher education curriculum (S78); Old textbooks (S53).
Digital media resources (n = 28)	–	Relevant internet pages such as: smartskole.no (https://www.smartskole.no/) (SmartSchool.no) (S25); nrk.no (national broadcasting company) (S28); nettartikler (internet articles) (S42); films on YouTube (S49).

Numbers refer to idea units within each emergent theme and sub-theme. S followed by a number refers to a particular participant.

TABLE 3 Emergent themes, sub-themes, and illustrative idea units based on responses to the second part of Scenario 2.

Emergent theme	Sub-themes	Illustrative idea units
Gaining new ideas and inspiration for own consideration (n = 13)	–	In order to gain more perspectives on how to awaken interest in students (S3); I would use the internet to get ideas (S29).
Others' experiences (n = 10)	–	As a rule, they will find themselves in similar situations and have tried out different tasks in their own classes (S29); they talk from experience (S45).
Specific answers and pre-prepared exercises (n = 9)	–	You can find good to go exercises for smartboard use (S25), If I didn't have any answers, I would check the internet or ask friends/family (S53).
Research-based or professional knowledge (n = 2)	–	To gather some professional material about the subject (S42).
Other pedagogical justifications (n = 13)	Student participation (n = 5), academic adaption (n = 3), concretization (n = 2), variation (n = 2), accessibility (n = 1)	To connect (the subject and learning materials) to news and facts that are relevant for students and their everyday life (S51).

Numbers refer to idea units within each emergent theme and sub-theme. S followed by a number refers to a particular participant.

TABLE 4 Emergent themes, sub-themes, and illustrative idea units based on responses to Scenario 3: Approaches to teaching the challenging class.

Emergent theme	Sub-themes	Illustrative idea units
Adapting instruction (n = 114)	Formal and informal evaluation (n = 31), academic adaption (n = 27), organization of instruction (n = 23), variation (n = 19), motivating approaches (n = 11), flexibility of homework (n = 3)	First I would give a test to check the academic level. Converse with the class about what kind of teaching works for them (S15).
Classroom management (n = 43)	Classroom environment (n = 24), leadership (n = 19)	It is important to set clear rules for what is allowed in class, and make clear demands of students (then everyone feels that they have been seen) (S5); Create a good and secure classroom environment (S90).
Social interactions with students (n = 31)	Relational approaches (n = 13), teacher's way of being (n = 9), student participation (n = 6), formal and informal evaluation of social competence (n = 3)	It would be important for me to get to know all of the pupils well, so that I would know how they work and "have more strings to play on" (S4).
Cooperation with parents and colleagues (n = 2)	Collaborate with parents, collaborate with colleagues	Try to collaborate with parents to make sure that the majority of students are supported at home (S20); Ask for a classroom assistant (S40).

Numbers refer to idea units within each emergent theme and sub-theme. S followed by a number refers to a particular participant.

TABLE 5 Emergent themes, sub-themes, and illustrative idea units based on responses to Scenario 3: Strengths and weaknesses as a teacher.

Emergent theme	Sub-themes	Illustrative idea units
Positive personal characteristics (n = 29)	Humanistic views (n = 8), patience and calmness (n = 6), warmth, kindness, and openness (n = 5), creativity (n = 4), communication skills (n = 4), all-rounder (n = 2)	". . . I appreciate all children and believe there is good in everyone. . ." (S4); I am patient, which can be useful (S49).
Mastery of tasks (n = 9)	Pedagogical competence (n = 8), mastery of subject matter (n = 1)	My strengths as a teacher are that I can assess their academic level and guide them on their way (S73); I would have a lesson plan and the knowledge to answer nearly everything they wonder about (S23).
Personal weaknesses (n = 4)	Lack of experience or efficacy	My weakness is that I panic when I don't know how to handle a situation, especially when the situation is new for me (S83); I feel this could be difficult, as I am still young and relatively uncertain (S100).
Task-related weaknesses (n = 11)	Lack of structure (n = 8), lack of knowledge (n = 3)	I don't know the class, that could be a weakness (S37), I am not very good at being an authoritarian (teacher) (S55).

Numbers refer to idea units within each emergent theme and sub-theme. S followed by a number refers to a particular participant.

has a meaning." Also, classroom culture (n = 3) was highlighted as a possible reason for differences in student performance, for example, "There is no culture for doing one's best in the classroom."

Finally, test context (n = 15) was believed to be a reason for the differences in student performance. The ideas that fell into this theme concerned aspects of the test and the testing. The sub-themes were test preparation and test difficulty. Test preparation (n = 8) reflected the view that the teacher had given too short notice of the test (i.e., 3 days), for example, "I, the teacher, should give notice 1 week in advance so the students can get prepared." Test difficulty (n = 7) concerned the wording of the test questions and the academic level of the test, for example, "that the questions in the test were formulated in a difficult manner" and "Maybe the test was too demanding for the students."

Scenario 2

The second scenario was designed to target participants' beliefs and reasoning about sources of teaching knowledge by asking them to consider where they would gather ideas for a teaching plan in a natural science class and how they would justify their decisions. Participants expressed a range of ideas about sources of knowledge that were captured by three main themes (see **Table 2**). Notably, an overweight of the ideas (n = 58) reflected a reliance on informal and craft-based knowledge sources, while 32 ideas reflected a reliance on more formal, evidence-informed sources of teaching knowledge. There were also 28 idea units that indicated reliance on digital media resources located on the Internet.

Informal and craft-based sources of teaching knowledge contained the sub-themes of colleagues, students, own internal resources, and family and friends. Colleagues (n = 31) emerged as a particularly prevalent knowledge source for the participants, and it was evidenced in idea units such as "I would discuss with my colleagues. . . Maybe I can get some colleagues to observe my lessons to see what I can do better and what I should change. . .It might be an idea to observe other teachers in their teaching for inspiration" and "As a new teacher, I can talk to more experienced teachers." Students (n = 14) were also considered a source of knowledge for our participants, for example, "I could have talked to the students to find out which methods they would learn most from" and "ask the students how they want the lessons to be." Participants' own experience, knowledge, thinking, reasoning, and creativity were categorized together to form a sub-theme (n = 12) indicating reliance on their own, internal resources, exemplified by "earlier experiences" and "I would use elements of the earlier lesson that had proven to be good and discard the parts that didn't work." Discussion with family and friends whilst being observant of student confidentiality, was also a suggested source (n = 1).

Within more formal sources of teaching knowledge, textbooks and educational literature (n = 29), teacher education (n = 2), and research (n = 1) were identified as sub-themes. Textbooks and educational literature included subject specific and education textbooks, teacher guides, and the national curriculum, for example, "I would look to relevant literature on learning strategies," "I would have used the book from education studies," and "I would also look at the goals from the core curriculum in the upcoming topic." Teacher education was referenced infrequently (e.g., "and from teacher education"), and research only featured in one of the participants' responses ("If someone has researched this, then it is interesting to see what results they attained").

Finally, digital media resources located on the Internet can be exemplified by "maybe I could have found some suggestions on the internet" and "find good and complete teaching plans for smartboard." There were also references to specific webpages, for example https://www.nrk.no/ (the national broadcasting company; see **Table 2**).

Of note is that a number of the participants (n = 30) failed to answer the question of where they would gather ideas for the teaching plan, simply presenting their own suggestions for such a plan. Since the participants were explicitly asked to provide the sources of their ideas, these responses were considered invalid in the context of this study and will, therefore, not be presented.

In the second part of Scenario 2, participants were asked to give reasons for their choices. Forty-seven idea units were identified, with eight students responding to the first part of the scenario failing to respond to the second part (i.e., not justifying their suggested sources of knowledge). The responses referred to four knowledge-related justifications: gaining new ideas and inspiration in order to consider these and use them as they wished (n = 13), testimony/others' experience, mainly referring to colleagues (n = 10), looking for specific answers and pre-prepared exercises (n = 9), and using research-based or professional knowledge as evidence (n = 2). There was also one category of responses (other pedagogical justifications, n = 13) that referred to pedagogical principles such as academic adaption and variations as justifications.

The most common justification, gaining ideas and inspiration, focused on getting suggestions for classroom practice, with references to "gathering more perspectives" that could be adapted in the way participants wished (e.g., "in this way I could see who learns from what... and arrive at something that works"). References to similar situations and experience-based knowledge were, not surprisingly, common justifications, with experienced colleagues given particular importance [e.g., "if they've gone through the curriculum (on this topic) before"]. Participants were also interested in finding solutions and pre-prepared, perhaps tried and tested, exercises from the internet and textbooks, perhaps suggesting a lack of motivation to use evidence to make reasoned decisions about their teaching (e.g., "there are lots of good exercises on the internet"). There were sparse justifications referring to the need for professional (i.e., evidence-informed) knowledge (gathering professional materials on the topic). Finally, participants also failed to provide epistemic justifications for their choice of sources of teaching knowledge, rather referring to pedagogical reasons for their choices. Sub-themes in the emerging theme of pedagogical justifications referred to student participation, academic adaption, variation, and accessibility, for example "it is important to vary teaching," "because the students have easy access to these (sources)."

Scenario 3

The third scenario was designed to capture participants' beliefs about their perceived efficacy by asking them to consider their chances for ensuring a satisfactory learning outcome for all students in a challenging class, taking their own strengths and weaknesses as a teacher as a point of departure. For this scenario, a large proportion of the participants (n = 69) detailed specific approaches they would use. We identified the four main themes of adapting instruction (n = 114), classroom management (n = 43), social interactions with students (n = 31), and co-operation with parents and colleagues (n = 2). These main themes and their respective sub-themes are presented and exemplified in **Table 4**.

Approaches that were coded as adapting instruction consisted of six sub-themes: Formal and informal evaluation of student knowledge and learning preferences (n = 31), academic adaption (n = 27), organization of instruction (n = 23), variation (n = 19), motivating approaches (n = 11), and flexibility and homework (n = 3). Formal and informal evaluation of student knowledge and learning preferences was mainly concerned with students' academic strengths and weaknesses, but also with preferences in terms of learning strategies and teaching methods, for example, "What is it that the students think is challenging? What do students think about this subject? I would start by asking this kind of question to find out more about them" and "What do they find difficult and which methods do they think work for them?" Building on such processes of evaluation, many participants highlighted the importance of adapted education, for example, "Make the topic a little more concrete and directed toward everyday life" and "I would have gone through the materials thoroughly with the class, and thereafter maybe let them work with tasks that are adapted to each individual." Organization of instruction referred to specific ways of working with the challenging class, for example, "I would have taken the students on a lot of excursions" and "Here I would have to try out different exercises, forms of teaching, and so on. When I eventually learn what works and not, I would go for more of those methods." Variation for variation's sake was also reflected in quite a few idea units, such as "I would also vary the teaching, so that the students could get to see several sides of me." Motivating approaches were directed toward fun, interest, and mastery, for example, "I would therefore use creativity to make the teaching more fun and exciting." Finally, a few idea units referred to flexibility and homework as approaches (e.g., ". . .something they have prepared at home. In my time at school, I have also seen that giving homework is usual and can help on tests."

Beliefs about classroom management consisted of idea units focusing on class environment (n = 24) and class leadership (n = 19). Class environment concerned the culture within the classroom, social competence, and attitudes, as well as security, for example, "Show the students that each and every one is seen every day and be a secure (adult)" and "tried to have a good dialog with the students continually, so that I know what the students' expectations of me are." Class leadership was more related to a strict tone in the classroom, being a clear leader, and establishment of clear rules. The aspects of leadership seemed to be tightly linked to one another, for example, "when I come into this class, it will be important for me to show who I want to be as an adult and leader for them and what I expect from them" and "Keep quite a strict tone."

The emergent theme termed social interactions with students focused primarily on emotional and relational aspects of teacher-student interactions, and the sub-themes concerned relational approaches (n = 13, e.g., "and form good relations with the students" and "let them get to know me"), teacher's way of being (n = 9, e.g., "The absolute first thing I would do is to go into this class without too many prejudices. Everyone deserves a chance with me as a new teacher"), student participation (n = 6, e.g., "It is also important to get input from the students and take these into consideration"), and formal and informal evaluation of social competence (n = 3, e.g., "In this class I would use a lot of time getting to know them. Find out what they like, also out of school").

Finally, two idea units reflected beliefs about co-operation with parents ("tried to collaborate well with the parents, so that as many of the students as possible were being attended to at home") and colleagues ("perhaps ask for an assistant at the start").

With respect to participants' perceived strengths and weaknesses in relation to teaching the challenging class, we identified two emerging themes representing beliefs about strengths (viz., positive personal characteristics and mastery of tasks) and two emerging themes representing beliefs about weaknesses (viz., personal weaknesses and task-related weaknesses). These main themes and their respective sub-themes are presented and exemplified in **Table 5**.

Idea units concerning positive personal characteristics (n = 29) could be categorized into six sub-themes: holding humanistic views and appreciating personal differences (n = 8, e.g., "That I appreciate all children and think that everyone has some good in them is an advantage"); patience and calmness [n = 6, e.g., "my strength in the classroom is that I am calm and this can be "infectious" (i.e., spread widely) in the class"]; warmth, kindness, and openness (n = 5, e.g., "I think I would get to know the class quickly, I am good at being open"); creativity (n = 4, e.g., "I am a creative person"); communication skills (n = 4, e.g., "That I am a clear leader also helps"); and being "all-rounders" (n = 2, e.g., "My strength is all-roundedness"). Idea units concerning mastery of tasks (n = 9) referred to pedagogical competence (n = 8, e.g., "I am good at adapting my teaching, so I know that everyone has a good learning outcome from my lessons" and "My strengths as a teacher are that I can evaluate their level and support them on the way") and mastery of subject matter (n = 1, "I will have the knowledge to answer almost anything they might wonder about").

Regarding weaknesses participants believed could hinder their ability to help all students learn (n = 15), personal weaknesses (n = 4) focused on lack of experience or efficacy, for example, "My weakness is that I panic when I don't know how to handle a situation, especially if the situation is new for me"). Finally, task-related weaknesses included lack of structure (n = 8, e.g., "From time to time I am easy to get to digress" and "My biggest weakness is that I might get carried away with myself if I think something is more interesting than the students do") and lack of knowledge about the subject or the students (n = 3, e.g., "I don't know the class, which might be a weakness").

Discussion

In this study, we introduced problem-based contexts to gain insight into second-year pre-service teachers' beliefs about student ability, sources of teaching knowledge, and teacher-efficacy. We also aimed to investigate what participants' responses revealed about their reliance on educational research as a means of informing their pedagogical decisions and actions, given that researchers and teacher educators have highlighted a lack of evidence-informed decisions and actions among teachers and student teachers (Bråten and Ferguson, 2015; Guilfoyle et al., 2020; Kiemer and Kollar, 2021).

Our findings align with prior research (Fives and Buehl, 2008; Bråten and Ferguson, 2015; Ferguson et al., 2022) but also provide new insight by merit of participants' responses to scenarios and their reasoning about the described problems. In particular, the pre-service teachers' beliefs about student performance actually placed most emphasis on the role of the teacher, their actions, and the ways they interact with students and adapt and vary their teaching methods and exercises. While taking the importance of their role to heart is, indeed, important, it is also relevant that the future teachers appreciated that students present themselves with individual differences in terms of ability, learning difficulties, and focus on learning. However, it is somewhat discouraging that beliefs in learning styles also feature prominently in the minds of some of our participants and are used to explain individual differences among students.

In terms of student motivation, our participants' beliefs seemed evidence-informed with respect to multiple motivational variables, such as interest, effort, and goal-orientation. While previous studies have found that inexperienced teachers tend to think of motivation in terms of a unidimensional construct that is, or is not, present in students (Patrick and Pintrich, 2001), participants in this study seems to hold more nuanced beliefs (see also, Ferguson and Bråten, 2018). Moreover, participants considered contextual factors such as the sociocultural background and context of students' learning and aspects of the test that featured in the scenario. Thus, our findings aligned with previous research by Patterson et al. (2016), which identified school, family, and student factors relating to teachers' beliefs about student performance. However, participants in this study seemed more focused on their own responsibility, given the prevalence of the category of teaching in the responses to Scenario 1. While it is neither uncommon for new teachers to have more focus on their own behavior than that of their students, nor unwarranted, given supporting evidence (Hattie, 2012), it may also be important for teacher educators to help pre-service teachers focus on other (important) factors that influence student performance, to help avoid teacher burnout in the long run (Skaalvik and Skaalvik, 2007).

Concerning the findings relating to the second scenario, about sources of teaching knowledge and participants' justifications for their choices, the categories of responses may be somewhat unsurprising, disappointing even, since most participants opted to rely on informal knowledge sources such as experienced colleagues, and since participants hardly referred to educational research at all, neither as sources of teaching knowledge nor when justifying their choice of sources. In general, participants' beliefs about sources of teaching knowledge therefore could not be considered consistent with or conducive to evidence-informed reasoning about sources of teaching knowledge and their justification. However, the nature of our data and the scenario-based approach allowed for added insights such as the ways in which the participants intended to use sources of knowledge. Although they were intent on finding out what colleagues might do in similar situations, they also signaled considerable independence in these situations. That is, participants intended to gain colleagues' perspectives as one of several views, sometimes also including student perspectives or knowledge from educational literature as other sources of teaching knowledge. While some of our participants were interested in finding readymade teaching exercises, it is difficult to draw further conclusions as to how the pre-service teachers intended to use these sources, since this was not elaborated in participants' responses.

Regarding participants' responses to the third scenario, designed to assess teacher efficacy, they detailed specific approaches to ensuring satisfactory learning outcomes in a challenging class before discussing their chances of success in light of their own strengths and weaknesses as a teacher. The four main themes of adapting instruction, classroom management, social interactions with students, and co-operation with parents and colleagues may be mapped on to measures of teacher-efficacy that are designed to reflect the multi-dimensional nature of teachers' work (Tschannen-Moran and Hoy, 2001), as well as the aims of the Norwegian national curriculum and school reforms (Skaalvik and Skaalvik, 2007).

Teacher-efficacy may be influenced by the factors that teachers view as important in influencing student performance. Thus, the pre-service teachers in this study who regarded factors relating to the teacher as being important for student performance may have had higher teacher-efficacy than those who focused more on internal student factors (individual differences) and socio-cultural differences relating to, for example, home circumstances (Skaalvik and Skaalvik, 2007). This is because factors relating to the teacher may support their experience of being able to exert effort and behaviors that influence student outcomes, compared to factors that are more out of their control. Further, the personal characteristics and task-related strengths and weaknesses that were highlighted by participants seemed more-or-less evidence-informed. As such, they adequately referred to aspects of teaching knowledge and experience, although a few of the responses concerning personal characteristics also suggested that views of certain characteristics of teachers as innate, rather than learned, still exist. Such responses were sparse, however, and perhaps somewhat ambiguous as they did not reveal the underlying mechanism of growth vs. fixed views (Dweck, 2000).

In sum, the results from this study uniquely contribute to the literature by showing the nuances of teacher beliefs about student performance, sources of teaching knowledge, and teacher-efficacy, and they provide further insight into pre-service teachers' limited consideration of research as evidence in pedagogical decision-making. We believe that our study has both methodological and theoretical implications in addition to its importance for practice.

Our study used a methodological approach that provides a more contextualized understanding of teachers' beliefs and, thus, can be assumed to generate more valid responses than those typically generated by asking participants to rate their beliefs on a questionnaire (Schraw and Olafson, 2015). However, our study also highlighted that pre-service teachers, despite the scenario-based approach, may find it somewhat difficult to explain and justify the sources of teaching knowledge that they draw on. More generally, teacher beliefs may often be tacit and difficult to articulate, and more research is needed on the relation between tacit beliefs and teaching practice, as well as on how future teachers can be helped to articulate their beliefs. Our results may also raise issues concerning the nature of evidence in teacher knowledge and educational research, and the extent to which empirical, experimental data may have to be supplemented with more ecologically valid studies that probe teacher thinking. For example, such studies may ask teachers with more and less experience to think aloud to determine how knowledge sources are actually used and how teachers try to integrate theory and practice.

Hopefully, the picture of teacher beliefs that we have painted in this study may help teacher educators understand pre-service teachers' beliefs and help them develop availing beliefs in teacher education (Sugrue, 1997; Schraw and Olafson, 2015; Lunn Brownlee et al., 2016; Mor-Hagani and Barzilai, 2022). Further, it may help teacher educators correct misconceptions among pre-service teachers that are detrimental to their own and their future students' learning (Menz et al., 2021), as well as to nurture their tendency to engage in evidence-informed practice (Buehl and Beck, 2015; Csandi et al., 2021; Hendriks et al., 2021). Working with pre-service teachers' beliefs in teacher education may also help future teachers think about concrete situations in terms of more abstract, theoretical, and evidence-informed sources of teaching knowledge, rather than acting on gut-feeling or (unexamined) habits (Csandi et al., 2021; Kiemer and Kollar, 2021; Spernes and Bjordal, 2022). In particular, our study may inform teacher educators about the nuances of beliefs they may encounter in teacher education programs, and how those beliefs may be consistent with what they teach in the programs but also take the form of some stubborn misconceptions (Menz et al., 2021) that need special attention. Changing such misconceptions may require extensive modeling and scaffolding by teacher educators who open up their own teaching beliefs and practices to demonstrate the evidence-base they employ in their teaching, and how it aligns with their educational beliefs (Ferguson, 2021).

Limitations and future research

A limitation of the present study is the nature of participants' succinct answers to the scenarios, with think-aloud data presumably allowing for firmer conclusions regarding participants' reasoning and use of evidence, in particular. Presumably, more ecological validity could also have been achieved by having pre-service teacher complete learning logs or observing them in action. However, our methodological approach was less invasive for our participants (and their students) and had no potential negative real-life consequences. In future research, a scenario-based approach may be extended by use of video cases or scenarios that can be discussed collectively and with the added advantage of time to reflect. Also, real-life examples may be presented as a starting point for discussions of theoretical constructs, characteristics, and teacher moves (Csandi et al., 2021; Spernes and Bjordal, 2022).

Of note is also that the themes emerging from the data and interpreted by the researchers, of course, do not contain exhaustive lists or possibilities when it comes to (pre-service) teachers' beliefs or reasoning. This is related to the content of the scenarios that we created. Although these three scenarios were designed to capture beliefs and reasoning about student ability, sources of teaching knowledge, and teacher efficacy, respectively, the hypothetical problem contexts they represented also might have been more or less likely to elicit evidence-informed thinking drawing on relevant educational research. For example, whereas the second scenario, in particular, may have provided valuable information about participants' evidence-informed reasoning (or the lack of it), the third scenario might have been better suited to activate participants' self-perceptions or self-evaluations than their reasoning about the usefulness or relevance of educational research. In future studies, it therefore seems important to both broaden the scope of the scenarios and ensure that they are equally well suited to reveal evidence-informed reasoning (or the lack of it) on the part of the participants. In this way, it may also be possible to maintain a clearer, more distinct focus on evidence-informed reasoning in analyzing and reporting scenario-based data, rather than focusing on beliefs and exploring to what extent these beliefs might or might not align with evidence-informed reasoning about educational issues, as we did in the current study. We acknowledge that it could be regarded as a limitation of this study that the data we collected did not lend themselves to a clear differentiation between beliefs and reasoning in participants' responses to the scenarios, with the signs of evidence-informed reasoning (or the lack of it) that we were able to identify being interwoven with participants' beliefs about student ability, sources of teaching knowledge, and teacher-efficacy.

Further, our data collection was a one-shot event in two particular programs of teacher education. What we were able to gain is therefore a snapshot in time of pre-service teachers in two teacher education programs in Norway that could be extended by conducting data collection at multiple time points taken throughout the teacher education period across different programs. As such, much more research is needed to obtain deeper insights into teacher beliefs and reasoning over time and in different contexts. Finally, we would like to highlight the need for more qualitative investigations into the mechanisms that connect teachers' beliefs and their (evidence-based) reasoning.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study was exempt from ethical approval procedures since no sensitive or identifying data was collected and the study was anonymous. The patients/participants provided their written informed consent to participate in this study.

Author contributions

Both authors designed the study, analyzed the data, wrote the manuscript, and had worked on the final version of the manuscript. LF completed the data collection and prepared the first draft of the manuscript. IB revised the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Afdal, H. W., and Spernes, K. (2018). Designing and redesigning research-based teacher education. *Teach. Teach. Educ.* 74, 215–228. doi: 10.1016/j.tate.2018.05.011

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.

Barzilai, S., and Weinstock, M. (2015). Measuring epistemic thinking within and across topics: A scenario-based approach. *Contemp. Educ. Psychol.* 42, 141–158. doi: 10.1016/j.cedpsych.2015.06.006

Bråten, I., and Ferguson, L. E. (2015). Beliefs about sources of knowledge predict motivation for learning in teacher education. *Teach. Teach. Educ.* 50, 13–23. doi: 10.1016/j.tate.2015.04.003

Buehl, M. M., and Alexander, P. A. (2001). Beliefs about academic knowledge. *Educ. Psychol. Rev.* 13, 385–418. doi: 10.1023/A:1011917914756

Buehl, M. M., and Beck, J. (2015). "The relationship between teachers' beliefs and practices," in *International handbook of research on teachers' beliefs*, eds H. Fives and M. Gregoire Gill (Milton Park: Routledge), 66–84.

Buehl, M. M., and Fives, H. (2009). Exploring teachers' beliefs about teaching knowledge: Where does it come from? Does it change? *J. Exp. Educ.* 77, 367–407. doi: 10.3200/JEXE.77.4.367-408

Buehl, M. M., and Fives, H. (2016). "The role of epistemic cognition in teacher learning and praxis," in *Handbook of epistemic cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (Milton Park: Routledge), 247–264.

Bullough, R. V. (2015). "Methods for studying beliefs: Teacher writing, scenarios, and metaphor analysis," in *International handbook of research on teachers' beliefs*, eds H. Fives and M. Gregorie Gill (Milton Park: Routledge), 150–169.

Chinn, C. A., Buckland, L. A., and Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. *Educ. Psychol.* 46, 141–167. doi: 10.1080/00461520.2011.587722

Csandi, A., Kollar, I., and Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: Better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. London: Psychology Press.

Dweck, C. S., and Leggett, E. (1988). A social-cognitive approach to motivation and personality. *Pyschol. Rev.* 95, 256–273. doi: 10.1037/0033-295X.95.2.256

Dweck, C. S., and Molden, D. C. (2017). "Mindsets: Their impact on competence motivation and acquisition," in *Handbook of competence and motivation*, 2nd Edn, eds A. J. Elliot, C. S. Dweck, and D. S. Yeager (New York, NY: The Guilford), 135–154.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., et al. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* 1, 13–29. doi: 10.1038/s44159-021-00006-y

European Commission (2007). *Towards More Knowledge-based Policy and Practice in Education and Training*. Available online at: https://op.europa.eu/en/publication-detail/-/publication/962e3b89-c546-4680-ac84-777f8f10c590/language-en

Ferguson, L. E. (2021). Evidence-informed teaching and practice-informed research. *Z. Pädagogische Psychol.* 35, 199–208. doi: 10.1024/1010-0652/a000310

Ferguson, L. E., and Bråten, I. (2018). Student teachers' beliefs about learning, teaching, and teaching knowledge: A qualitative study. *Teach. Educ. Pract.* 31, 348–365.

Ferguson, L. E., and Lunn Brownlee, J. (2021). "Teacher beliefs and epistemologies," in *Oxford bibliographies of education*, ed. A. Hynds (Oxford: Oxford University Press), doi: 10.1093/obo/9780199756810-0276

Ferguson, L. E., Bråten, I., Skibsted Jensen, M., and Andreassen, R. (2022). A longitudinal mixed methods study of Norwegian preservice teachers' beliefs about sources of teaching knowledge and motivation to learn from theory and practice. *J. Teach. Educ.* Advance online publication. doi: 10.1177/00224871221105813

Fives, H., and Buehl, M. M. (2008). What do teachers believe? Developing a framework for examining beliefs about teachers' knowledge and ability. *Contemp. Educ. Psychol.* 33, 134–176. doi: 10.1016/j.cedpsych.2008.01.001

Fives, H., and Buehl, M. M. (2016). Teachers' beliefs, in the context of policy reform. *Policy Insights Behav. Brain Sci.* 3, 114–121. doi: 10.1177/2372732215623554

Fives, H., Barnes, N., Chiavola, C., Saiz de La Mora, K., Oliveros, E., and Mabrouk-Hattab, S. (2019). "Reviews of teachers' beliefs," in *Oxford research encyclopedias: Education*, ed. J. Lampert (Oxford: Oxford University Press), doi: 10.1093/acrefore/9780190264093.013.781

Guilfoyle, L., McCormack, O., and Erduran, S. (2020). The "tipping point" for educational research: The role of pre-service science teachers' epistemic beliefs in evaluating the professional utility of educational research. *Teach. Teach. Educ.* 90, 1–15. doi: 10.1016/j.tate.2020.103033

Hartmann, U., Kindlinger, M., and Trempler, K. (2021). Integrating information from multiple texts relates to pre-service teachers' epistemic products for reflective practice. *Teach. Teach. Educ.* 87, 1–12. doi: 10.1016/j.tate.2020.103205

Hattie, J. (2012). *Visible learning for teachers*. Milton Park: Routledge. doi: 10.4324/9780203181522

Hendriks, F., Seifried, E., and Menz, C. (2021). Unraveling the "smart but evil" stereotype: Pre-service teachers' evaluations of educational psychology researchers versus teachers as sources of information. *Z. Pädagogische Psychol.* 35, 1–15. doi: 10.1024/1010-0652/a000300

Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educ. Psychol. Rev.* 13, 191–209. doi: 10.1023/A:1016667621114

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Jonsson, A. C., and Beach, D. (2012). Predicting the use of praise among pre-service teachers: The influence of implicit theories of intelligence, social comparison, and stereotype acceptance. *Educ. Inq.* 3, 259–281. doi: 10.3402/edui.v3i2.22033

Jonsson, A. C., Beach, D., Korp, H., and Erlandson, P. (2012). Teachers' implicit theories of intelligence: Influences from different disciplines and scientific theories. *Eur. J. Teach. Educ.* 35, 387–400. doi: 10.1080/02619768.2012.662636

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychol. Rev.* 93, 429–445. doi: 10.1037/0033-295X.93.4.429

Jussim, L., and Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Pers. Soc. Psychol. Rev.* 9, 131–155. doi: 10.1207/s15327957pspr0902_3

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. Do pre-service teachers' beliefs regarding the utility of (non-) scientific information sources matter? *Z. Pädagogische Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kitchen, J., and Petrarca, D. (2016). "Approaches to teacher education," in *International handbook of teacher education*, Vol. 1, eds J. Lughran and M. L. Hamilton (Berlin: Springer), 137–186. doi: 10.1007/978-981-10-0366-0_4

Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago, IL: University of Chicago Press.

Lunn Brownlee, J., Ferguson, L. E., Scholes, L., McDonald, S., Stahl, G., Comber, B., et al. (2021). Middle school students' science epistemic beliefs: Implications for measurement. *Int. J. Educ. Res.* 105, 1–11. doi: 10.1016/j.ijer.2020.101719

Lunn Brownlee, J., Schraw, G., Walker, S., and Ryan, M. (2016). "Changes in preservice teachers personal epistemologies," in *Handbook of epistemic cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (Routledge), 300–317.

Menz, C., Spinath, B., and Seifried, E. (2021). Where do pre-service teachers' educational psychological misconceptions come from? The roles of anecdotal versus scientific evidence. *Z. Pädagogische Psychol.* 35, 143–156. doi: 10.1024/1010-0652/a000299

Mor-Hagani, S., and Barzilai, S. (2022). The multifaceted nature of teachers' epistemic growth: Exploring teachers' perspectives on growth in epistemic performance. *Teach. Teach. Educ.* 115, 1–13. doi: 10.1016/j.tate.2022.103714

Norwegian Ministry of Education and Research, (1999). *Act relating to primary and secondary education and training (the Education Act)*. Available online at: https://lovdata.no/dokument/NLE/lov/1998-07-17-61 (accessed May 23, 2022).

Norwegian Ministry of Education and Research (2011). *National guidelines for differentiated primary and lower secondary teacher education programmes for years 1-7 and 5-10.* Available online at http://www.regjeringen.no/en/dokumenter/national-guidelines/2010/national-guidelines-for-differentiated-t/id640249/ (accessed May 23, 2022).

Olafson, L., Grandy, C. S., and Owens, M. C. (2015). "Qualitative approaches to studying teachers' beliefs," in *International handbook on teachers' beliefs*, eds H. Fives and M. G. Gill (Milton Park: Routledge), 128–149.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Rev. Educ. Res.* 62, 307–332. doi: 10.3102/00346543062003307

Patrick, H., and Pintrich, P. R. (2001). "Conceptual change in teachers' intuitive conceptions of learning, motivation, and instruction: The role of motivational and epistemological beliefs," in *Understanding and teaching the intuitive mind: Student and teacher learning*, eds B. Torff and R. J. Sternberg (Mahwah, NJ: Erlbaum), 117–143.

Patterson, M. M., Kravchenko, N., Chen-Bouck, L., and Kelley, J. A. (2016). General and domain-specific beliefs about intelligence, ability, and effort among preservice and practicing teachers. *Teach. Teach. Educ.* 59, 180–190. doi: 10.1016/j.tate.2016.06.004

Rosenthal, R., and Jacobsen, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York, NY: Holt, Rinehart and Winston.

Sabatini, J., O'Reilly, T., Wang, Z., and Dreier, K. (2018). "Scenario-based assessment of multiple source use," in *Handbook of multiple source use*, eds J. L. G. Braasch I, Bråten, and M. T. McCrudden (Milton Park: Routledge), 447–465. doi: 10.4324/9781315627496-25

Schommer, M. (1990). The effects of beliefs about the nature of knowledge on comprehension. *J. Educ. Psychol.* 82, 498–504. doi: 10.1037/0022-0663.82.3.498

Schraw, G., and Olafson, L. (2015). "Assessing teachers' beliefs: Challenges and solutions," in *International handbook of research on teachers' beliefs*, eds H. Fives and M. G. Gill (Milton Park: Routledge), 87–105.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harv. Educ. Rev.* 57, 1–21. doi: 10.17763/haer.57.1.j463w79r56455411

Skaalvik, E. M., and Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *J. Educ. Psychol.* 99, 611–625. doi: 10.1037/0022-0663.99.3.611

Skaalvik, E. M., and Skaalvik, S. (2014). Teacher self-efficacy and perceived autonomy: Relations with teacher engagement, job satisfaction, and emotional exhaustion. *Psychol. Rep.* 114, 68–77. doi: 10.2466/14.02.PR0.114k14w0

Spernes, K., and Bjordal, F. R. (2022). "Lærerstudenters erfaringer fra egen grunnskoletid som grunnlag for refleksjon om klasseledelse [Student teachers' experiences from own schooling as a foundation for reflection on classroom managment]," in *Proceedings of the forskning på egen undervisning og pedagogisk praksis [Conference on researching own teaching and educational practice]* (Fredrikstad: Høgskolen I Østfold).

Sugrue, C. (1997). Student teachers' lay theories and teaching identities: Their implications for professional development. *Eur. J. Teach. Educ.* 20, 213–225. doi: 10.1080/0261976970200302

Thomm, E., Gold, B., Betsch, T., and Bauer, J. (2021a). When preservice teachers' prior beliefs contradict evidence from educational research. *Br. J. Educ. Psychol.* 91, 1055–1072. doi: 10.1111/bjep.12407

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021b). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Z. Pädagogische Psychol.* 35, 173–184. doi: 10.1024/1010-0652/a000301

Thomm, E., Seifried, E., and Bauer, J. (2021c). Informing professional practice: (Future) teachers' choice, use, and evaluation of (non-)scientific sources of educational topics. *Z. Pädagogische Psychol.* 35, 121–126. doi: 10.1024/1010-0652/a000309

Tschannen-Moran, M., and Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teach. Teach. Educ.* 17, 783–805. doi: 10.1016/S0742-051X(01)00036-1

Tschannen-Moran, M., Hoy, A. W., and Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Rev. Educ. Res.* 68, 202–248. doi: 10.3102/00346543068002202

Wang, Z., O'Reilly, T., Sabatini, J., McCarthy, K., and McNamara, D. (2021). A tale of two tests: The role of topic and general academic knowledge in traditional versus contemporary scenario-based reading. *Learn. Instr.* 72, 1–12. doi: 10.1016/j.learninstruc.2021.101462

Yeager, D. S., and Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educ. Psychol.* 47, 302–314. doi: 10.1080/00461520.2012.722805

Zimmermann, M., and Mayweg-Paus, E. (2021). The role of collaborative argumentation in future teachers' selection of online information. *Z. Pädagogische Psychol.* 35, 185–198. doi: 10.1024/1010-0652/a000307

Appendix A

The teaching scenarios

You are about to answer three questions. Read each task carefully and take the time you need to answer thoroughly.

Question A: You are teaching a 6th grade natural science class. The class consists of 22 students, 10 boys and 12 girls. After working with the topic "the human body," you set a class test on that chapter. The class received 3 days notice about the test. The test results show that five students completed all tasks correctly, while five students performed very poorly on the test. The test performance was average for the rest of the class. Discuss reasons for these great differences in students' learning outcomes. Answer the question as fully as possible. Use the time you need to reflect when you respond.

Question B: You are now going to create a teaching plan for the next topic in natural science class that takes into consideration the great differences in student learning outcomes that were apparent on the last test. Where will you gather ideas for this teaching plan? Give reasons for your choices.

Question C: Imagine a year has passed. You are now teaching a new 6th grade class in natural science. This class is known for being the school's most challenging class to teach. You have taken over as form teacher for this class since their previous form teacher has retired. Discuss your chances of being able to ensure a good learning outcome in natural science for all the students in this class–taking your own strengths and weaknesses as a teacher as a point of departure.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/feduc.2022.977631

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University,
Germany

REVIEWED BY
Timothy Fukawa-Connelly,
Temple University,
United States
Tom Rosman,
Leibniz-Institute
for Psychology (ZPID), Germany
Hendrik Lohse-Bossenz,
University of Education Heidelberg,
Germany

*CORRESPONDENCE
Elisabeth Bauer
elisabeth.bauer@psy.lmu.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 24 June 2022
ACCEPTED 05 October 2022
PUBLISHED 03 November 2022

CITATION
Bauer E, Sailer M, Kiesewetter J, Fischer MR and Fischer F (2022) Diagnostic argumentation in teacher education: Making the case for justification, disconfirmation, and transparency. *Front. Educ.* 7:977631. doi: 10.3389/feduc.2022.977631

Diagnostic argumentation in teacher education: Making the case for justification, disconfirmation, and transparency

Elisabeth Bauer[1]*, Michael Sailer[1], Jan Kiesewetter[2], Martin R. Fischer[2] and Frank Fischer[1]

[1]Education and Educational Psychology, Department of Psychology, Ludwig-Maximilians-Universität in Munich, Munich, Germany, [2]Institute for Medical Education, University Hospital, Ludwig-Maximilians-Universität in Munich, Munich, Germany

Research on diagnosing in teacher education has primarily emphasized the accuracy of diagnostic judgments and has explained it in terms of factors such as diagnostic knowledge. However, approaches to scientific argumentation and information processing suggest differentiating between *diagnostic judgment* and *diagnostic argumentation*: When making accurate diagnostic judgments, the underlying reasoning can remain intuitive, whereas diagnostic argumentation requires controlled and explicable reasoning about a diagnostic problem to explain the reasoning in a comprehensible and persuasive manner. We suggest three facets of argumentation for conceptualizing diagnostic argumentation, which are yet to be addressed in teacher education research: *justification* of a diagnosis with evidence, *disconfirmation* of differential diagnoses, and *transparency* regarding the processes of evidence generation. Therefore, we explored whether preservice teachers' diagnostic argumentation and diagnostic judgment might represent different diagnostic skills. We also explored whether justification, disconfirmation, and transparency should be considered distinct subskills of preservice teachers' diagnostic argumentation. We reanalyzed data of 118 preservice teachers who learned about students' learning difficulties with simulated cases. For each student case, the preservice teachers had to indicate a diagnostic judgment and provide a diagnostic argumentation. We found that preservice teachers' diagnostic argumentation seldom involved all three facets, suggesting a need for more specific training. Moreover, the correlational results suggested that making accurate diagnostic judgments and formulating diagnostic argumentation may represent different diagnostic skills and that justification, disconfirmation, and transparency may be considered distinct subskills of diagnostic argumentation. The introduced concepts of justification, disconfirmation, and transparency may provide a starting point for developing standards in diagnostic argumentation in teacher education.

KEYWORDS

teacher education, diagnostic argumentation, scientific argumentation, diagnostic accuracy, diagnostic judgment, diagnostic knowledge

Introduction

Diagnostic skills are relevant in many fields, one of which is teacher education (Heitzmann et al., 2019). Teachers' diagnosing is a prototypical practice scenario for evidence-oriented practice, and as such, it is crucial for teachers' professionalism (Fischer, 2021). Previous research on teachers' diagnosing has primarily investigated diagnostic accuracy—i.e., the correctness of diagnostic judgments—because inaccurate judgments can easily disadvantage students by, for example, leading to unsuitable or insufficient educational interventions (e.g., Loibl et al., 2020; Urhahne and Wijnia, 2020; Kramer et al., 2021a). Besides making accurate diagnostic judgments, communicating diagnostic considerations is another vital aspect of diagnostic skills, for example, for purposes such as reporting diagnostic findings (Bauer et al., 2020) or collaborative diagnosing (Kiesewetter et al., 2017). However, thus far, there is no clear conceptualization of diagnostic argumentation, which we define as explaining a diagnostic judgment and the underlying reasoning comprehensibly and persuasively (see Walton, 1990; Berland and Reiser, 2009). It is also unclear whether professionals (e.g., teachers) who can make accurate diagnostic judgments are capable of offering sufficient diagnostic argumentation. This raises the question of whether accurate diagnostic judgment and diagnostic argumentation are fully based on the same knowledge—reflecting one overarching diagnostic skill—or whether they need to be considered different subskills of diagnosing. This differentiation might have implications for teaching diagnostic skills, such as the definition of learning objectives and the design and implementation of learning environments (see Grossman et al., 2009).

To our knowledge, no systematic research has differentiated between the concepts of diagnostic argumentation and diagnostic judgment. Therefore, we propose a conceptualization of diagnostic argumentation that consists of three facets: *justification* of a diagnosis with evidence, *disconfirmation* of differential diagnoses, and *transparency* regarding the processes of evidence generation. We explore diagnostic argumentation in terms of these three facets and investigate whether they indicate one joint underlying skill or different aspects of diagnostic skills by analyzing their interrelations with one another and with a potentially joint knowledge base. We also explore how justification, disconfirmation, and transparency in diagnostic argumentation are related to the accuracy of diagnostic judgments in the context of teacher education.

Diagnosing in teacher education

Teacher education is one of the fields in which learning diagnostic skills is an important matter of professionalization (Grossman, 2021). In particular, teachers have to diagnose students' performance, progress, and learning prerequisites (e.g., Praetorius et al., 2013; Südkamp et al., 2018). However, these aspects also include the initial identification of clinical problems, such as learning difficulties (e.g., dyslexia) and behavioral disorders (e.g., attention deficit hyperactivity disorder, i.e., ADHD; e.g., Poznanski et al., 2021). In all these contexts, we broadly define *diagnosing* as a "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty in order to make […] educational decisions" (Heitzmann et al., 2019, p. 4). Other associated terms are used for diagnosing in teacher education as well, such as *assessment* (e.g., Herppich et al., 2018). As part of teachers' professional activities, diagnosing is crucially related to the discussion around teachers' evidence-oriented practice (Stark, 2017) and is possibly a prototypical practice scenario (Fischer, 2021). Teachers are expected to use knowledge on theories, methods, procedures, and findings from educational research (e.g., Kiemer and Kollar, 2021) to reflect their experiences, possibly overcome dysfunctional intuitive approaches and—at least partially—guide their diagnostic activities and interventions. Teacher education programs are increasingly acknowledging the relevance of facilitating diagnostic skills, and research in teacher education has also addressed the issue of how diagnostic skills are learned (e.g., Chernikova et al., 2020; Loibl et al., 2020; Sailer et al., 2022).

Teachers' diagnostic judgments

Previous research on teachers' diagnosing has focused on how teachers make diagnostic judgments (e.g., Loibl et al., 2020; Urhahne and Wijnia, 2020; Kramer et al., 2021a). Loibl et al. (2020) suggested distinguishing between the processes and products of teachers' diagnostic judgments In terms of product indicators, research on teachers' diagnostic judgments has focused on diagnostic accuracy—i.e., the correctness of diagnostic judgments—because inaccurate judgments can lead to unsuitable or insufficient educational interventions that easily disadvantage students (e.g., Urhahne and Wijnia, 2020). There is also an increasing amount of research investigating teachers' judgment processes, for example, in terms of diagnostic activities such as generating hypotheses, generating and evaluating evidence, and drawing conclusions (e.g., Wildgans-Lang et al., 2020; Codreanu et al., 2021; Kramer et al., 2021a). In addition, research has begun to focus more on the role of information processing in teachers' judgment processes (e.g., Loibl et al., 2020). Teachers' diagnostic judgment processes can involve intuitive information processing—i.e., fast recognition of patterns of information—which facilitates flexible and adaptive acting in the classroom; teachers can also engage in controlled information processing when spending time and effort on consciously evaluating evidence and its causal relations (Kahneman, 2003; Evans, 2008). Teachers' information processing in making diagnostic judgments depends on situational characteristics (Loibl et al., 2020), such as the available time for making a judgment (Rieu et al., 2022), the consistency and conclusiveness of the available evidence, and teachers' perceptions of their situational accountability (Pit-ten

Cate et al., 2020). In classrooms with multiple students, teachers often need to make intuitive judgments, prioritize tasks, and decide where to invest their time and cognitive resources (Feldon, 2007; Vanlommel et al., 2017). With respect to achieving diagnostic accuracy, research suggests regarding judgment processes (e.g., in terms of information processing) as processes that interact with teachers' characteristics, especially their diagnostic knowledge (e.g., Loibl et al., 2020; Kramer et al., 2021a).

The role of diagnostic knowledge

Diagnostic knowledge is generally considered an important basis of diagnostic skills (Heitzmann et al., 2019). Having a sufficient base of specific diagnostic knowledge seems to be a necessary condition for achieving accurate diagnostic judgments (Kolovou et al., 2021). In addition, advanced diagnosticians' well-organized knowledge structures enable them to recognize patterns of critical case information correctly, without necessarily conducting a controlled analysis of the underlying causal relations (see Kahneman, 2003; Evans, 2008; Boshuizen et al., 2020). Research has suggested that performing complex cognitive tasks requires not only knowledge about relevant concepts but also knowledge about *how* to systematically approach the task (e.g., Van Gog et al., 2004). In the context of teacher education, Shulman (1986) suggested that, besides domain-specific content, distinguishing between different types of knowledge—such as conceptual and strategic knowledge—is relevant to capturing different functionalities of knowledge, such as acting adaptively in response to various problems and situations. In the course of developing strategic knowledge, basic aspects of conceptual knowledge are abstracted and integrated with episodic knowledge into cognitive scripts about approaching certain problems or situations (e.g., Shulman, 1986; Schmidmaier et al., 2013; Boshuizen et al., 2020). This means that conceptual and strategic knowledge about the same specific content are likely related but address different aspects of solving a task. Conceptual and strategic knowledge have been adapted and empirically investigated in the context of diagnosing in medical education (e.g., Stark et al., 2011; Schmidmaier et al., 2013): *conceptual diagnostic knowledge* (CDK) consists of concepts, such as diagnoses and their relations with each other and with evidence, whereas *strategic diagnostic knowledge* (SDK) refers to how to proceed in diagnosing a specific problem (i.e., how to reject or confirm differential diagnoses and which informational sources provide critical evidence for doing so). Researchers addressing diagnosing in teacher education have also suggested distinguishing between CDK and SDK (e.g., Förtsch et al., 2018). Therefore, CDK and SDK seem crucial for correctly processing relevant case information and making accurate diagnostic judgments.

Diagnostic argumentation

Beyond making accurate diagnostic judgments, there are instances in which teachers or other diagnosticians need to explain their reasoning and the resulting diagnostic judgment in a comprehensible and persuasive manner, which we suggest to designate as *diagnostic argumentation* (see Walton, 1990; Berland and Reiser, 2009). Diagnostic argumentation is required in situations in which explanations are directed toward a recipient, such as a collaborating teacher or school psychologist (e.g., Kiesewetter et al., 2017; Csanadi et al., 2020; Radkowitsch et al., 2021). The context of identifying students' clinical problems is one example in which diagnostic argumentation is particularly relevant for teachers, as in many educational systems, final judgments about clinical diagnoses are made by clinical professionals (e.g., school psychologists), with whom teachers might need to collaborate (Albritton et al., 2021). However, also in other contexts, diagnostic argumentation facilitates a collaborative process of considering and reconciling competing explanations and thus, if necessary, can help improve the diagnosing (see Berland and Reiser, 2009; Csanadi et al., 2020). There are also nonimmediate dialogical situations (see Walton, 1990), such as writing a report about diagnostic findings (Bauer et al., 2020), in which information may need to be comprehensible and persuasive to potential recipients at a later point in time.

Especially when engaging in a face-to-face critical exchange of arguments in collaborative or otherwise dialogical diagnosing, teachers might involve in argumentation processes and a controlled analysis of the available evidence and potential explanations before making a diagnostic judgment. Collaborative generation and evaluation of evidence and a critical evaluation of others' arguments can improve the quality of argumentative outcomes (Mercier and Sperber, 2017; Csanadi et al., 2020). In other contexts, teachers might make intuitive judgments without a controlled analysis of all the available evidence and causal relations. If the information processing for a diagnostic judgment mainly involves intuitive pattern recognition, parts of the reasoning can remain implicit (Evans, 2008). However, comprehensively explaining a judgment and its underlying reasoning initially requires that the reasoning be explicable or at least constructible in retrospect. In terms of nondialogical situations, such as writing reports, initial evidence suggests that compared to medical education, there seems to be a lower standardization in teacher education (Bauer et al., 2020), which could facilitate constructing persuasive explanations in retrospect. For these reasons, it might not necessarily be a given that teachers who make accurate judgments in nondialogical diagnostic situations are capable of subsequently providing comprehensible and persuasive explanations of their reasoning. This open question has yet to be explored by research.

Justification, disconfirmation, and transparency in diagnostic argumentation

To explore how diagnostic judgment and diagnostic argumentation are related, it is first necessary to define what kind

10.3389/feduc.2022.977631

of information is expected to be provided in the context of diagnostic argumentation. We argue that besides providing comprehensible explanations, diagnostic argumentation also aims to persuade potential recipients of the presented reasoning (Berland and Reiser, 2009) and, thus, requires providing information that enables a recipient's understanding and evaluation of the efforts made during diagnosing (see Chinn and Duncan, 2018). Therefore, to further define the concept of diagnostic argumentation, we suggest three facets that might facilitate recipients' understanding of the presented reasoning: justification, disconfirmation, and transparency. We propose that these three facets of diagnostic argumentation resemble approaches in scientific argumentation (see Sampson and Clark, 2008; Mercier and Heintz, 2014), namely justifying one's reasoning with evidence (e.g., Toulmin, 1958), considering and disconfirming alternative explanations (e.g., Lawson, 2003), and emphasizing the credibility of informational sources with methodological transparency (e.g., Chinn et al., 2014). In what follows, we explain the three facets in further detail.

Justification denotes the provision of evidence in support of a claim (e.g., Toulmin, 1958; Hitchcock, 2005), which allows recipients to raise potential issues about the reasoning that was presented. In the context of diagnostic argumentation, diagnostic judgments are claims that need to be justified by providing evidence derived from the case information. Therefore, justifications evaluate relevant case information as evidence from which to draw conclusions concerning a judgment (see Fischer et al., 2014).

Disconfirmation emphasizes discussing differential diagnoses that may have been hypothesized when diagnosing a given case. As a process of uncertainty reduction (Heitzmann et al., 2019), diagnosing involves generating and evaluating different hypotheses (Klahr and Dunbar, 1988; Fischer et al., 2014) that resemble competing claims in argumentation. Similar to the scientific approach of disconfirmation (e.g., Gorman et al., 1984), a rebuttal of competing claims supports the persuasiveness of the final claim (e.g., Toulmin, 1958; Lawson, 2003). In diagnostic argumentation, differential diagnoses are competing claims that should be explicated and discussed to facilitate the persuasiveness of the final judgment by demonstrating that alternative explanations have been considered. Recipients can build on this information to evaluate and criticize whether relevant differential diagnoses have been missed or mistakenly rejected.

Transparency regarding the processes of evidence generation provides information about the reliability of the methodology for generating evidence from informational sources (Chinn et al., 2014; Fischer et al., 2014). In diagnostic argumentation, transparency is achieved by describing the processes underlying evidence generation, thus allowing recipients to evaluate the presented evidence and diagnostic conclusions. Explicating how evidence was generated facilitates a recipient's understanding and ability to criticize the quality of the evidence and, ultimately, the validity of the conclusions (Vazire, 2017).

Analogously to approaches involved in scientific argumentation (see Sampson and Clark, 2008; Mercier and Heintz, 2014), we suggest that justification, disconfirmation, and transparency in diagnostic argumentation facilitate a recipient's understanding and evaluation of the efforts made during diagnosing. We are unaware of any research in teacher education that has conceptualized or investigated a skill similar to what we have defined as diagnostic argumentation, including the facets of justification, disconfirmation, and transparency. Therefore, in this study, we aimed to explore the interrelations between justification, disconfirmation, and transparency in diagnostic argumentation, as well as their relations with making accurate diagnostic judgments, and the explanatory roles of CDK and SDK (see Figure 1).

Research questions

We propose that justification, disconfirmation, and transparency are three relevant facets of diagnostic argumentation and that diagnostic argumentation and diagnostic judgment might represent two distinct diagnostic skills that may, however, both be partially explained by CDK and SDK. Understanding the interrelations between these skills and knowledge might provide relevant information for teacher educators and the field of teacher education.

In investigating the proposed concept of diagnostic argumentation, it is also important to explore whether justification, disconfirmation, and transparency might represent distinct subskills or indicators of one joint underlying diagnostic skill (RQ1). To approach this question, we investigated how the individual facets (1a) and different combinations of the facets (1b) occur within preservice teachers' diagnostic argumentation and analyzed the facets' relations (1c) in preservice teachers' diagnostic argumentation. We assumed that finding close relationships would indicate a joint basis of knowledge and skills; by contrast, small relationships or a lack thereof would indicate that the three facets represent different subskills of diagnostic argumentation.

In terms of distinguishing between the three facets as different subskills, a related question is to what extent justification, disconfirmation, and transparency are based on conceptual diagnostic knowledge and strategic diagnostic knowledge (RQ2). Because CDK and SDK are thought to be a major basis for the reasoning presented in diagnostic argumentation (Heitzmann et al., 2019), we assumed that they also partially explain justification, disconfirmation, and transparency; that is, CDK and SDK may be needed to generate evidence from informational sources (explicated in transparency) and to make a warranted connection between the evidence and a diagnosis (explicated in justification) or several differential diagnoses (explicated in disconfirmation). Exploring the degree to which CDK and SDK explain justification, disconfirmation, and transparency in diagnostic argumentation can provide an initial basis for future research on teachers' prerequisites for diagnostic argumentation.

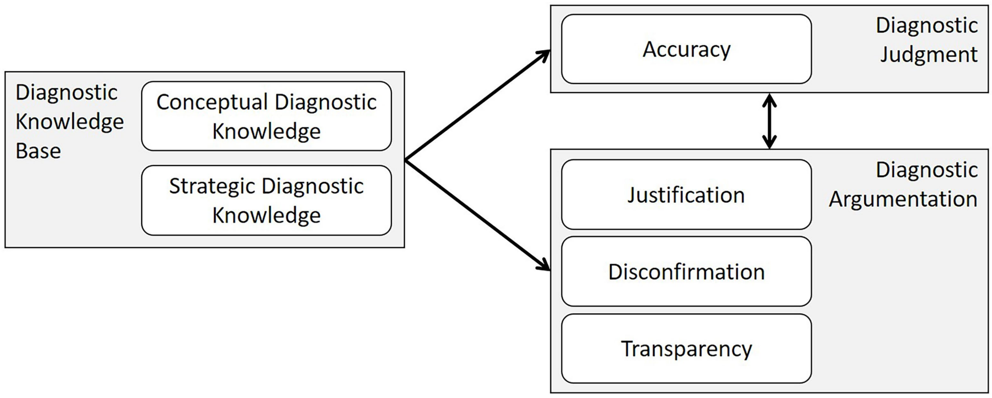


FIGURE 1
Theoretical model of the potential relationships between diagnostic knowledge and the skills of diagnostic judgment and diagnostic argumentation.

Given that diagnostic argumentation additionally aims to be persuasive instead of solely verbalizing the reasoning made while processing information, further knowledge and skills beyond CDK and SDK may contribute to justification, disconfirmation, and transparency in diagnostic argumentation.

For the same reason, we assumed that, despite a presumably joint basis of CDK and SDK, diagnostic accuracy might not necessarily be related to justification, disconfirmation, and transparency. Therefore, we explored whether diagnostic judgment (indicated by diagnostic accuracy) and diagnostic argumentation (indicated by justification, disconfirmation, and transparency) might represent different diagnostic skills (RQ3). In doing so, we assumed that identifying close relationships would indicate a joint underlying diagnostic skill; by contrast, small relationships or a lack thereof would indicate that diagnostic argumentation and diagnostic judgment might represent different diagnostic skills.

Materials and methods

Participants

In this study, we reanalyzed data that were originally collected to train an AI-based adaptive feedback component for a simulation-based learning environment (see Pfeiffer et al., 2019). A total of 118 preservice teachers participated in the data collection and processed simulated cases pertaining to students' clinical problems. Participants were $M = 22.96$ years old ($SD = 4.10$), the majority were women (102 women, 15 men, and 1 nonbinary), and they were in their first to 13th semester ($M = 4.62$, $SD = 3.40$) of a teacher education program. We recruited preservice teachers in all semesters because relevant courses about students' clinical problems were not compulsory or bound to a specific semester but could be taken in any semester. Participants subjectively rated their prior knowledge of students' clinical problems prior to receiving any instruction about the content of the study. On average, they indicated a medium rating of their own prior knowledge (on a rating scale ranging from 1 to 5 points: prior knowledge about ADHD, $M = 2.78$, $SD = 0.81$; prior knowledge about dyslexia, $M = 2.47$, $SD = 0.76$). We assumed that this sample mirrors the diverse population of preservice teachers.

Research design

We chose a quantitative and correlational research design to determine the relationships between the following variables: justification, disconfirmation, and transparency in diagnostic argumentation; CDK and SDK; and the accuracy of diagnostic judgment.

Simulation and tasks

We asked participants to take on the role of a teacher and process eight cases of primary and secondary students with performance-related or behavioral problems that might or might not indicate a clinical diagnosis in the range of ADHD or dyslexia. Two independent domain experts, one school psychologist and one psychotherapist for children and adolescents, validated the case materials before they were implemented in CASUS, a case-based online learning environment.[1] Participants solved the cases consecutively. The cases included several informational sources, such as samples of the students' written exercises and school certificates, reports of observations from inside and outside the classroom, and conversations with the respective students, their parents, and other teachers (the German-language case materials can be accessed at https://osf.io/hn7wm/). Participants could freely

1 http://www.casus.net/

choose how many and which informational sources to consult and in which order they wanted to do so (see Figure 2).

One example is the case of a secondary school student named Anna, who is showing symptoms of attention-deficit disorder (ADD). An initial problem statement describes Anna as a fifth-grade student, 11 years old, who constantly needs to be pushed to finish her tasks and who has poor grades in many subjects, especially the core subjects, such as math and the language subjects. The learners could examine written observations of Anna's in-class and out-of-class behavior, read recordings of conversations with Anna or with her parents and several teachers, or look at Anna's last annual report and an example of a written exercise. Her behavior is described as very calm and distracted. She reads very slowly, and it is difficult for her to answer questions about a text that she has just read. She often fails to follow the exact instructions for tasks or fails to complete them fully. Moreover, she often does not bring all the required school supplies or arrives late in the morning. At a parent–teacher conference, Anna's mother backs up the impression of disorganized and slow learning behavior when talking about Anna's homework. Anna's last annual report and the conversations with the other teachers show that her grades are also affected by her inattentiveness, except artistic subjects and gym class. She mostly interacts with one friend and tends to remain distant from the other students. Anna herself points out that it is hard for her to concentrate because she feels easily distracted. However, at home, where there are fewer ambient noises, she can focus on and enjoy reading, drawing, and painting. Overall, the case information is designed in such a way that the diagnosis of ADD is the most likely diagnosis, despite the fact that several differential diagnoses may be relevant. The other cases included the same kinds of informational sources as Anna's case.

To complete a case and move on to the next case, participants had to complete two tasks. First, they had to make a diagnostic judgment, answering the question of whether the simulated student has issues that warrant further diagnosing of a clinical problem and, if so, which diagnosis may apply. Second, we asked participants to write an argumentation text about their conclusions and their reasoning about the case. For the purpose of this study, participants received no further guidance or support regarding how to write their diagnostic argumentation.

Procedure

The data were collected on computers in a laboratory setting, with three to 20 participants simultaneously joining the study. They worked individually at separate desks and were not permitted to speak to each other. We introduced the participants to the aims and procedure of the study and familiarized them with the learning environment. After giving informed consent to participate in the study, participants received randomly assigned codes to log on to the CASUS learning environment to anonymize the data. When entering the online learning environment, participants first received a 25 min theoretical input concerning the topic of diagnosing in general and the diagnosing of ADHD and dyslexia in particular to activate existing knowledge and ensure the minimum amount of knowledge required for solving the cases. Participants were not

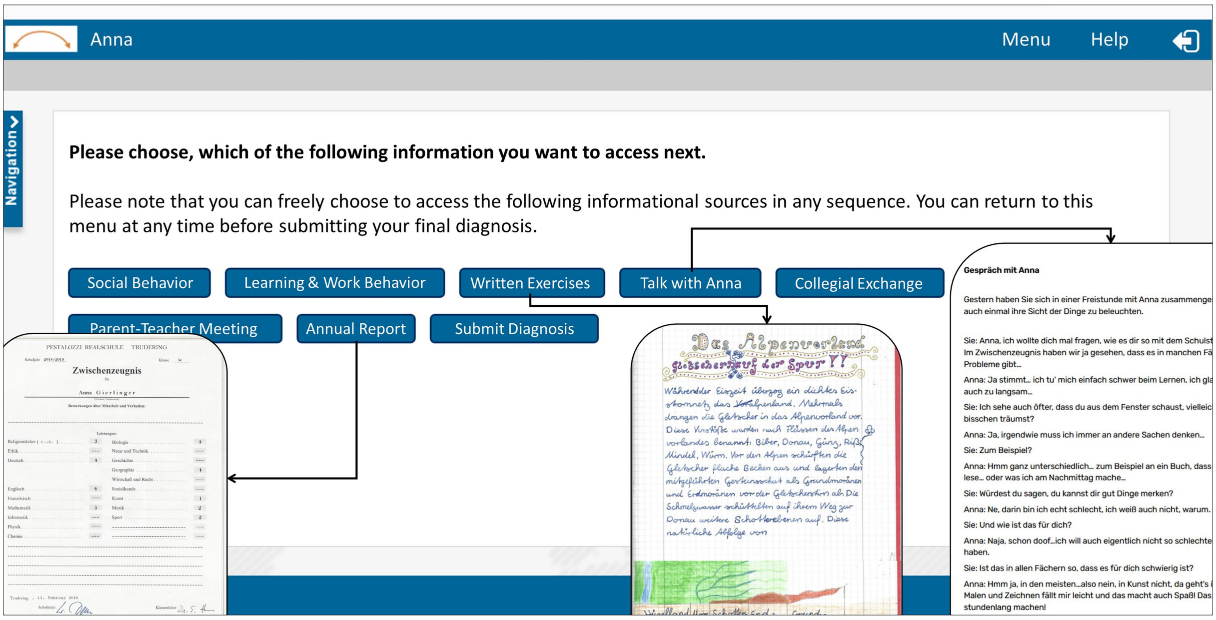


FIGURE 2
Example of the case materials in the simulation-based learning environment.

allowed to take any notes or go back to the input part at a later point to avoid biases in subsequent testing and learning. Following the theoretical input, participants spent around 25 min on a pretest that assessed their CDK and SDK. Subsequently, participants entered the learning phase consisting of the eight simulated cases, with a break of 10 min after four cases. They had to finish one case at a time to gain access to the next case. All participants received the cases in the same sequence. The time on task for all cases was around 1 h. Subsequently, participants spent around 25 min on a posttest. Generally, participants were allowed to work at their own pace. Overall, participants spent around 3 h from login to logout. During the study, researchers were available to help with technical issues or questions about navigation but did not answer any content-related questions. Participants received monetary compensation of 35 euros.

Data sources and measurements

The data sources used for the presented analyses are the CDK and SDK scores from the pretest as well as the written diagnostic judgments and diagnostic argumentation texts from six of the eight cases. We decided to exclude two cases from the analysis because their case information turned out to be more ambiguous and inconclusive compared to the other cases.

Diagnostic knowledge

Conceptual diagnostic knowledge

CDK was assessed in the pretest after participants received the theoretical input. We used 14 single-choice items about diagnosing ADHD and dyslexia with four answer options each (one correct answer and three distractors). The CDK questionnaire was developed prior to the study to assess participants' CDK, which was considered relevant for processing the simulated cases. Two independent domain experts, one school psychologist and one psychotherapist for children and adolescents, validated the CDK questionnaire. One example item is "Which of the following is *not* one of the cardinal symptoms of ADHD?" with the answer options (a) Inattentiveness, (b) Hyperactivity, (c) Impulsivity, and (d) Impatience. Participants received one point per correct answer. The points were aggregated into a total score, ranging from 0 to 14 points, for CDK.

As suggested by Stadler et al. (2021; see also Diamantopoulos and Siguaw, 2006; Taber, 2018), we calculated variance inflation factors (VIFs) for all items to avoid having redundant items representing the formative knowledge construct. The maximum VIF was VIF_{max} = 1.30, which is well below the recommended cut-off of 3.3.

Strategic diagnostic knowledge

Subsequent to assessing CDK, we measured SDK using four key-feature cases (two key-feature cases about ADHD and two about dyslexia) with two multiple-choice questions each (see Page et al., 1995). Key-feature cases present a brief description consisting of a few sentences before asking about the strategic approaches used to diagnose the case. The key-feature cases were developed prior to the study to assess participants' SDK, which was considered relevant for processing the simulated cases. Two independent domain experts, one school psychologist and one psychotherapist for children and adolescents, validated the key-feature cases. One example key-feature case introduced the fourth grader Luis, who has always been a rather poor reader but has begun to fall farther behind his classmates over the last few months and just recently again received the lowest grade in the class on a reading test. He cannot summarize the contents of a short text even immediately after reading it and can only read aloud very slowly. Apart from his performance issues, he has a chronic disease due to which he cannot regularly attend school for stretches of several weeks. After reading this brief case description, two multiple-choice questions were asked.

The first of the two multiple-choice questions per key-feature case asked participants to choose all relevant differential diagnoses out of a list of clinical as well as non-clinical differential diagnoses (one to three correct options out of seven to nine answer options). Participants received points for correctly choosing relevant options and not choosing irrelevant options. We calculated one mean score across all options per key-feature case, resulting in a diagnosis score of 0 to 1 for the first question for each key-feature case.

The second of the two multiple-choice questions per key-feature case asked participants to choose from a list of further approaches and resources relevant to confirm or disconfirm a given set of differential diagnoses (three to six correct options out of seven to 10 answer options). Participants received points for correctly choosing relevant options and not choosing irrelevant options. We calculated one mean score across all options per key-feature case, resulting in a resource score of 0 to 1 for the second question for each key-feature case.

The four diagnosis scores and four resources scores were accumulated into a total score of 0 to 8 points for SDK on the pretest. There were no redundant items (VIF_{max} = 1.09).

Accuracy of diagnostic judgment

To measure diagnostic accuracy, we coded all the written diagnoses as accurate (1 point), partially accurate (0.5 points), or inaccurate (0 points). We coded written diagnoses as accurate if indicating a diagnosis that was considered the correct solution when designing the cases (e.g., ADD for the case Anna). The written diagnoses were coded as partially accurate if correctly indicating the higher-level class of diagnoses for the accurate diagnosis (e.g., if the correct diagnosis was ADD and the participants indicated ADHD). A total of 12.5% of the diagnoses were double-coded, resulting in an interrater reliability (IRR) of Cohen's κ = 0.80 (Cohen, 1960). The internal consistency across the six cases was

McDonald's ω = 0.37 (McDonald, 1999). For further analyses, we calculated a total score from the points achieved for diagnostic accuracy with a possible range of 0 to 6 points.

Justification, disconfirmation, and transparency in diagnostic argumentation

We operationalized justification, disconfirmation, and transparency based on a coding of the six cases' diagnostic argumentation texts.

Justification

We operationalized the presence or absence of justification in diagnostic argumentation as *evaluating evidence* co-occurring with *drawing conclusions* within the temporal context of two sentences, resulting in 1 or 0 points per diagnostic argumentation. In this study, we reanalyzed data that were originally used to train an AI-based adaptive feedback algorithm for a simulation-based learning environment (see Pfeiffer et al., 2019). Four expert raters coded the diagnostic argumentation texts segmented by sentences regarding the categories *evaluating evidence* and *drawing conclusions*. They initially read the complete diagnostic argumentation before coding *evaluating evidence* and *drawing conclusions* for the individual sentences. *Evaluating evidence* was defined as explicitly presenting or interpreting case information (e.g., "Markus behaves aggressively and gets offended very easily"). *Drawing conclusions* was defined as explicitly accepting or rejecting at least one diagnosis (e.g., "I think most likely the diagnosis is ADHD"). The raters simultaneously coded 15% of the data before dividing the rest of the data because of substantial agreement (IRRs: Fleiss' κ = 0.71 for *drawing conclusions*; Fleiss' κ = 0.75 for *evaluating evidence*; Fleiss, 1971; Landis and Koch, 1977). The internal consistency across six cases was sufficient (McDonald's ω = 0.60; McDonald, 1999). We calculated a total justification score for each participant, with a possible range of 0 to 6 points.

Disconfirmation

We operationalized disconfirmation as present if two or more *differential diagnoses* were addressed, resulting in 1 or 0 points per diagnostic argumentation. This round of coding was done separately from the coding of justification and transparency for the purpose of our reanalysis. Two expert raters coded the diagnostic argumentation texts of six cases regarding a set of *differential diagnoses*. The coding scheme consisted of 27 differential diagnoses, which included non-clinical (e.g., insufficient schooling, emotional stress, and problematic home environment) and clinical differential diagnoses (e.g., ADHD, ADD, dyslexia, and autism). The raters considered the facet of disconfirmation as being included in the diagnostic argumentation if two or more of these differential diagnoses were discussed in one diagnostic argumentation, independent of which diagnosis the participant indicated as the final diagnosis. The raters simultaneously coded 15% of the data before dividing the rest of the data (overall IRR: Cohen's κ = 0.92; Cohen, 1960). The internal consistency was sufficient (McDonald's ω = 0.60; McDonald, 1999). We calculated a total disconfirmation score for each participant, with a possible range of 0 to 6 points.

Transparency

We operationalized transparency in diagnostic argumentation as at least one explication of *generating evidence*, resulting in 1 or 0 points per diagnostic argumentation. The coding for transparency was done in in the same round as the coding for justification. Four expert raters coded the diagnostic argumentation texts regarding *generating evidence*, which was defined as an explicit description of accessing informational sources (i.e., tests or observations; e.g., "I observed Anna's school-related behavior and achievement"). The raters simultaneously coded 15% of the data before dividing the rest of the data because of substantial agreement (IRR: Fleiss' κ = 0.70; Landis and Koch, 1977). The internal consistency was sufficient (McDonald's ω = 0.71; McDonald, 1999). We calculated a total transparency score for each participant, with a possible range of 0 to 6 points.

Statistical analyses

For RQ1, we explored the descriptive statistics of justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation texts in terms of both individual facets (1a) and facet combinations (1b). We considered facet combinations as types of argumentation texts and depicted them in relation to the individual facets using Epistemic Network Analysis (ENA; Shaffer, 2017). The ENA algorithm analyzes and accumulates co-occurrences of elements in coded data, such as the three facets of argumentation within individual argumentation texts, to create a multidimensional network model, which is depicted as a dynamic network graph. To determine the types of argumentation texts, we grouped the argumentation texts according to the presence or absence of each argumentation facet in each argumentation text. The ENA algorithm then accumulated co-occurrences of the three facets across the argumentation texts to create a network model. We depicted this model as a two-dimensional network graph that showed the relative location of the argumentation types within the resulting two-dimensional space. We used the ENA online tool to create the network graphs.[2] In addition to the descriptive analyses, we calculated Pearson correlations with participants' overall justification, disconfirmation, and transparency scores (1c). To investigate RQ2, we calculated a multivariate multiple linear regression with the predictors CDK and SDK and the dependent variables justification, disconfirmation, and transparency. For RQ3, we first created two separate ENA networks by grouping the diagnostic argumentation texts that addressed either accurate or inaccurate

2 https://www.epistemicnetwork.org/

TABLE 1 Prevalence of the individual facets justification, disconfirmation, and transparency in the 709 diagnostic argumentation texts.

	Number of argumentation texts including the facet	Number of argumentation texts missing the facet
Justification	468 (66%)	241 (34%)
Disconfirmation	183 (26%)	526 (74%)
Transparency	327 (46%)	382 (54%)

TABLE 2 Prevalence of the argumentation types, indicated by combinations of the facets Justification (J), Disconfirmation (D), and Transparency (T), in the 709 argumentation texts.

Argumentation types, indicated by combinations of the three facets	Number of facets included	Number of argumentation texts	Percent of argumentation texts
JDT	3	83	11.7%
JDO	2	90	12.7%
JOT	2	129	18.2%
ODT	2	7	1.0%
JOO	1	166	23.4%
ODO	1	3	0.4%
OOT	1	108	15.2%
OOO	0	123	17.3%

diagnostic judgments; we tested the difference between the group means' locations in the network space using a *t*-test. To facilitate the statistical testing of the groups' network differences, we used the option of means rotation, which aligns the two group means on the X-axis of the network, thus, depicting systematic variance in only one dimension in the two-dimensional space (Shaffer, 2017). Moreover, we again calculated Pearson correlations, including the participants' overall scores for diagnostic accuracy, justification, disconfirmation, and transparency. We also explored partial correlations, controlling for CDK and SDK. For RQ1c and RQ3, including multiple comparisons (three Pearson correlations each), the significance level was Bonferroni-adjusted to $\alpha = 0.0167$ ($\alpha = 0.05/3$). For the other analyses, the significance level was set to $\alpha = 0.05$.

Results

RQ1: Justification, disconfirmation, and transparency

To investigate whether justification, disconfirmation, and transparency represent distinct subskills or one joint underlying diagnostic skill (RQ1), we analyzed the prevalence of the individual facets (1a) and the combinations of the facets (1b) in preservice teachers' individual argumentation texts. Moreover, we analyzed the relationships between justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation (1c). We considered findings of close relations to indicate a joint basis of knowledge and skills, and small or no relations to indicate that the three facets represent different aspects of diagnostic skills.

RQ1a: Prevalence of the facets in preservice teachers' argumentation texts

Analyzing the descriptive statistics of the prevalence of justification, disconfirmation, and transparency in preservice teachers' individual argumentation texts, we found that *justification* was the most common of the three facets in all diagnostic argumentation texts (see Table 1): Participants explicitly stated conclusions and justified them by evaluating evidence alongside the conclusion in 66% ($M = 0.66$; $SD = 0.47$) of all argumentation texts. *Disconfirmation* was found in 26% ($M = 0.26$; $SD = 0.44$) of all diagnostic argumentation texts, indicating that the majority of diagnostic argumentation texts did not involve differential diagnoses but tended to focus on one final diagnosis. Moreover, we found *transparency* concerning the processes of evidence generation in 46% ($M = 0.46$; $SD = 0.50$) of all argumentation texts, indicating that approximately half of the diagnostic argumentation texts explained the processes of evidence generation.

RQ1b: Combinations of the facets in preservice teachers' argumentation texts

Descriptive statistics of the combinations of justification, disconfirmation, and transparency are outlined in Table 2. The combinations of the three facets can be considered different types of diagnostic argumentation texts, which we distinguished using the following abbreviations: *J* indicates the presence of justification, *D* indicates the presence of disconfirmation, *T* indicates the presence of transparency, and *O* indicates the absence of a facet (e.g., *JOT* indicates justification and transparency without disconfirmation; see Table 2 for all argumentation types and their prevalence). A

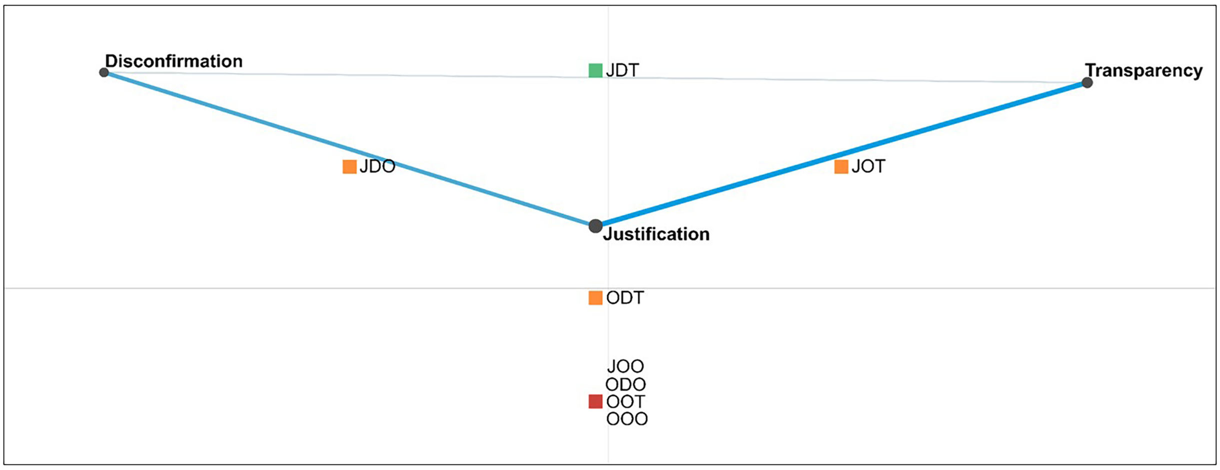


FIGURE 3
Argumentation types (indicated by colored squares) plotted in a two-dimensional space to indicate the relationship between the argumentation types and the individual facets (indicated by gray nodes) and the co-occurrences of the facets (indicated by blue lines). argumentation types are characterized by: J, Justification; D, Disconfirmation; T, Transparency; O, Absence of a Facet. For example: JOT, Justification and transparency without disconfirmation.

notable pattern was that argumentation texts addressing more than one diagnosis usually discussed the different diagnoses by evaluating evidence to make and justify conclusions (*JDT* and *JDO*), whereas hardly any argumentation texts addressed differential diagnoses without making and justifying related conclusions (*ODT* and *ODO*). However, diagnostic argumentation texts frequently presented a confirmatory justification of a single diagnosis without discussing alternative explanations (*JOT* and *JOO*). Consequently, including disconfirmation in diagnostic argumentation was dependent on including justification, but justification in diagnostic argumentation was not dependent on including the facet of disconfirmation, suggesting a relationship of unidirectional dependency.

To illustrate the types of argumentation texts and their relationships with the individual facets, we used ENA to plot both the argumentation types (indicated by colored squares) and the individual facets (indicated by gray nodes) in a two-dimensional space (see Figure 3). The two-dimensional space was built based on the co-occurrences of two argumentation facets each, which are indicated by the blue lines. The thickness of the blue lines represents the relative frequency of the co-occurrences (e.g., the thick line between justification and transparency relates to the 212 co-occurrences of justification and transparency in *JDT* and *JOT*). The positioning of argumentation types (indicated by the colored squares) along the X-axis is relative to the facets' co-occurrences, which is why *JOT* is located toward the right-sided node of transparency and *JDO* is located toward the left-sided node of disconfirmation. The central positioning of justification is due to its high overall prevalence (see Table 1). The positioning of argumentation types along the Y-axis indicates the argumentation texts' comprehensiveness regarding the three facets, with the extremes of *JDT* (all facets are present) and *OOO* (all facets are missing).

Overall, the findings indicate that preservice teachers tend to primarily provide justification in their diagnostic argumentation as an antecedent to including disconfirmation, transparency, or both. Moreover, the results suggest that there may be a relationship of unidirectional dependency of disconfirmation on justification.

RQ1c: Relations of justification, disconfirmation, and transparency

Beyond exploring the three facets in the individual argumentation texts, we also analyzed the descriptive statistics and correlations of preservice teachers' justification, disconfirmation, and transparency across the cases. The descriptive results of the facets' total scores (see Table 3) were consistent with the pattern found in the individual argumentation texts (see Table 1). Participants mostly focused on *justification* (M = 3.83, SD = 1.58), rarely used *disconfirmation* (M = 1.53, SD = 1.41), and put a medium emphasis on *transparency* (M = 2.67, SD = 1.81). The correlational analysis (see Table 3) indicated that justification and disconfirmation were significantly correlated, with a large effect ($r = 0.568$, $p < 0.001$). By contrast, transparency was not significantly correlated with justification ($r = 0.055$, $p = 0.554$) or disconfirmation ($r = 0.025$, $p = 0.787$). Considering the unidirectional dependency of disconfirmation on justification (see the results of RQ1b), we interpreted the overall result pattern as suggesting that justification, disconfirmation, and transparency are distinct facets of diagnostic argumentation rather than indicators of a uniform skill.

TABLE 3 Descriptive results and Pearson correlations of preservice teachers' scores for the three argumentation facets justification, disconfirmation, and transparency, as well as conceptual and strategic diagnostic knowledge and diagnostic accuracy.

		M	SD	1.	2.	3.	4.	5.
1.	Justification	3.83	1.58					
2.	Disconfirmation	1.53	1.41	$r = 0.568$, $p = 0.000$				
3.	Transparency	2.67	1.81	$r = 0.055$, $p = 0.554$	$r = 0.025$, $p = 0.787$			
4.	Conceptual diagnostic knowledge	8.86	1.66	$r = 0.265$, $p = 0.004$	$r = 0.234$, $p = 0.011$	$r = 0.041$, $p = 0.659$		
5.	Strategic diagnostic knowledge	6.70	0.39	$r = 0.252$, $p = 0.006$	$r = 0.042$, $p = 0.652$	$r = 0.194$, $p = 0.035$	$r = 0.130$, $p = 0.161$	
6.	Diagnostic accuracy	4.42	0.94	$r = 0.284$, $p = 0.002$	$r = 0.105$, $p = 0.259$	$r = 0.059$, $p = 0.526$	$r = 0.185$, $p = 0.045$	$r = 0.222$, $p = 0.016$

RQ2: Relations of conceptual and strategic diagnostic knowledge with justification, disconfirmation, and transparency

To explore the extent to which CDK and SDK predicted the dependent variables of justification, disconfirmation, and transparency, we calculated a multivariate multiple linear regression. Participants achieved $M = 8.86$ points ($SD = 1.66$) out of a maximum of 14 points on the CDK test and $M = 6.70$ points ($SD = 0.39$) out of a maximum of eight points on the SDK test (see Table 3). The Pearson correlations of the three argumentation facets with the variables CDK and SDK are reported in Table 3. The overall regression model with the predictors CDK and SDK significantly predicted justification—$F(2, 115) = 7.725$, $p = 0.001$—and explained 11.8% of the variance. Both CDK ($\beta = 0.236$, $p = 0.009$) and SDK ($\beta = 0.222$, $p = 0.013$) contributed significantly to the model. Similarly, disconfirmation was significantly predicted by the overall regression model, with the predictors CDK and SDK—$F(2, 115) = 3.331$, $p = 0.039$—explaining 5.5% of the variance. Whereas CDK ($\beta = 0.232$, $p = 0.012$) contributed significantly to the model, SDK did not ($\beta = 0.012$, $p = 0.898$). By contrast, transparency was not significantly predicted by the overall regression model, including both predictors, CDK and SDK—$F(2, 115) = 2.264$, $p = 0.109$—which explained 3.8% of the variance. CDK ($\beta = 0.016$, $p = 0.861$) was not a significant predictor of transparency; however, SDK ($\beta = 0.192$, $p = 0.040$) was a significant predictor of transparency in diagnostic argumentation.

Overall, justification, disconfirmation, and transparency were each partially explained by CDK, SDK, or both, with small effect sizes. Across the three facets, there were considerable differences in the amounts of variance explained by CDK and SDK. Moreover, the pattern in which CDK and SDK predicted justification, disconfirmation, and transparency differed considerably.

RQ3: Relationship between diagnostic judgment and diagnostic argumentation

To explore whether diagnostic judgment and diagnostic argumentation represent different diagnostic skills, we started again by plotting argumentation texts in ENA. First, we grouped argumentation texts according to diagnostic accuracy to compare argumentation concerning inaccurate versus accurate judgments. Second, we explored preservice teachers' total scores to investigate whether diagnostic accuracy correlated with justification, disconfirmation, and transparency in diagnostic argumentation.

To explore whether the argumentation texts differed if concerning an accurate vs. an inaccurate judgment, we grouped the individual argumentation texts by diagnostic accuracy and created one overall ENA network per group. We descriptively compared the networks of the groups of argumentation texts concerning accurate judgments (see Figure 4A) and inaccurate judgments (see Figure 4C), which we found to be highly similar (see also the comparison plot in Figure 4B, which shows the other two networks' differences). To determine whether the two groups of argumentation texts differed significantly, we centered the networks, resulting in the two group means (indicated by colored squares, with confidence intervals indicated by colored dashed boxes) depicted in Figure 4B.

All networks in Figure 4 were rotated to align both group means to the X-axis, which enabled statistical testing of group differences in a single dimension (Shaffer, 2017). The positioning of the group mean of argumentation texts concerning inaccurate judgments ($M = -0.01$, $SD = 0.38$, $n = 100$) was not statistically significantly different from the positioning of the group mean of argumentation texts concerning accurate judgments ($M = 0.01$, $SD = 0.41$, $n = 457$; $t(153.53) = 0.56$, $p = 0.58$, Cohen's $d = 0.06$). The analysis suggests that, overall, argumentation texts did not differ if addressing an accurate versus an inaccurate judgment.

We proceeded with a correlational analysis of preservice teachers' total scores to investigate whether their overall diagnostic accuracy was correlated with justification, disconfirmation, and transparency (see Table 3). On average, participants achieved a diagnostic accuracy of $M = 4.42$ points ($SD = 0.94$) out of a maximum of six achievable points. We found that participants' diagnostic accuracy and justification were significantly correlated, with a small effect ($r = 0.284$, $p = 0.002$). By contrast, diagnostic accuracy was not significantly correlated with either disconfirmation ($r = 0.105$, $p = 0.259$) or transparency ($r = 0.059$, $p = 0.526$).

To determine the role of CDK and SDK in explaining the relationship between diagnostic accuracy and justification, we calculated a partial correlation between diagnostic accuracy and

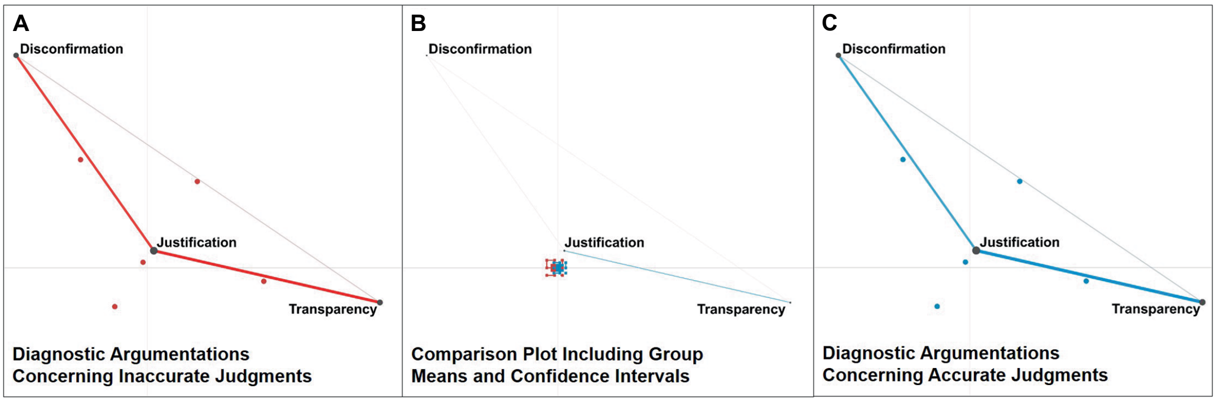


FIGURE 4
Networks across diagnostic argumentation texts, grouped by diagnostic accuracy and rotated by group means: **(A)** shows the network of diagnostic argumentation texts concerning inaccurate judgments. **(C)** shows the network of diagnostic argumentation texts concerning accurate judgments. **(B)** shows the comparison plot, which depicts the differences between the other two networks, as well as the group means (indicated by colored squares) and confidence intervals (indicated by colored dashed boxes) of the other two networks.

justification, statistically controlling for CDK and SDK (see Table 3 for the Pearson correlations of CDK and SDK with the argumentation facets and diagnostic accuracy). We found that the resulting partial correlation between diagnostic accuracy and justification in diagnostic argumentation remained significant, with a small effect ($r = 0.211$, $p = 0.023$). Thus, controlling for CDK and SDK hardly decreased the effect size of the correlation between diagnostic accuracy and justification. Consequently, our results suggest that CDK and SDK are not the variables that primarily explain the relationship between diagnostic accuracy and justification.

Overall, the results only indicate a weak relationship between the accuracy of preservice teachers' diagnostic judgments on the one hand, and justification, disconfirmation, and transparency in their diagnostic argumentation on the other. CDK and SDK did not explain the small correlation between diagnostic accuracy and justification. Moreover, groups of argumentation texts concerning inaccurate versus accurate judgments did not show a statistically significant difference. These findings suggest that diagnostic judgment and diagnostic argumentation can be considered different diagnostic skills.

Discussion

In exploring whether justification, disconfirmation, and transparency represent distinct subskills or one joint underlying diagnostic skill (RQ1), we found that preservice teachers primarily provide justification in their diagnostic argumentation as an antecedent to including disconfirmation or transparency in their diagnostic argumentation. Furthermore, we found a unidirectional dependency of disconfirmation on justification; diagnostic argumentation texts presenting more than one diagnosis usually discussed the differential diagnoses by evaluating evidence to make conclusions; however, preservice teachers often only argued for their final diagnosis without discussing competing explanations. Concerning the interrelations between justification, disconfirmation, and transparency, we found that they were distinguishable facets of diagnostic argumentation. Determining the extent to which justification, disconfirmation, and transparency were explained by CDK and SDK (RQ2), we found that justification was predicted by CDK about diagnoses and evidence as well as SDK about diagnostic approaches and activities. Disconfirmation of different diagnoses was only predicted by CDK of diagnoses. By contrast, transparency about the diagnostic approaches for generating evidence was only predicted by SDK of diagnostic proceedings for generating evidence. However, the variance explained by CDK and SDK was low. Furthermore, the accuracy of diagnostic judgments and justification, disconfirmation, and transparency in diagnostic argumentation did not necessarily seem to be related (RQ3). Overall, groups of argumentation texts addressing either accurate or inaccurate diagnostic judgments did not show a statistically significant difference. However, in contrast to disconfirmation and transparency, we found that justification in diagnostic argumentation was significantly correlated with the accuracy of diagnostic judgments. Despite statistically controlling for CDK and SDK, the relationship between the accuracy of diagnostic judgments and justification in diagnostic argumentation remained significant, suggesting that other variables may be important in explaining the relationship.

Overall, we interpreted the results as suggesting that diagnostic judgment and diagnostic argumentation might be different diagnostic skills. Finding a relationship between the accuracy of diagnostic judgments and justification in diagnostic argumentation supports the relevance and validity of the construct of diagnostic argumentation. Yet, the argumentation facets seemed to be sufficiently distinguishable from one another and from diagnostic accuracy. Finding differences regarding the predictive patterns of CDK and SDK (see Förtsch et al., 2018) supports the notion that justification, disconfirmation, and transparency are distinct subskills of diagnostic argumentation. Justification involves explicitly evaluating evidence as the basis for concluding a diagnosis (see Fischer et al., 2014; Heitzmann et al., 2019). Therefore, justification requires CDK about relevant concepts (e.g., diagnoses, evidence, and their interrelations; see Förtsch et al., 2018). Moreover, justification requires making warranted connections between evidence and diagnoses (e.g., Toulmin, 1958) to conclude or reject diagnoses, which seems to be facilitated by SDK (see Förtsch et al., 2018). Disconfirmation involves addressing differential diagnoses to demonstrate that alternative explanations have been considered (e.g., Toulmin, 1958; Lawson, 2003), which seems to primarily require CDK about differential diagnoses. By contrast, transparency, which involves describing the processes behind evidence generation (see Chinn et al., 2014; Vazire, 2017), seems to rely on SDK when it comes to the process of diagnosing a specific problem (e.g., which informational sources can deliver critical evidence).

Large amounts of variance in justification, disconfirmation, and transparency remained unexplained by CDK and SDK. Our findings raise the question of which additional kinds of knowledge and skills may be used when formulating justified, disconfirming, and transparent diagnostic argumentation. Beyond CDK and SDK, we propose two additional variables that might play a role in explaining justification, disconfirmation, and transparency within diagnostic argumentation: (1) knowledge about standards in diagnosing and diagnostic argumentation (see Chinn et al., 2014; Bauer et al., 2020) and (2) argumentation skills that are transferrable across domains (Hetmanek et al., 2018). In teacher education, there seems to be limited agreement about standards in diagnostic practices compared with other fields, such as medical education (Bauer et al., 2020). Teacher education programs do not yet systematically teach agreed-upon standards for communicating in situations that require what we defined as diagnostic argumentation. Consequently, preservice teachers likely do not have much knowledge about standards in diagnostic argumentation. There might also be differences between teacher and medical education in what are considered suitable standards for diagnostic argumentation (Bauer et al., 2020). Moreover, teachers and teacher educators might vary in their views regarding the role of scientific standards in diagnostic argumentation. Therefore, it is important to continue to discuss such standards in teacher education. We suggest using justification, disconfirmation, and transparency as a starting point from which to further discuss, systematize, and teach standards for diagnostic argumentation in teacher education.

The performance differences and higher prevalence of justification observed in the current study may be explained by argumentation skills that are transferrable across domains. It has been suggested that cross-domain transferable skills can, to some extent, compensate for a lack of more specifically relevant knowledge (e.g., knowledge about standards in diagnostic argumentation; Hetmanek et al., 2018). Accordingly, knowledge about standards in diagnostic argumentation, as well as cross-domain transferable argumentation skills, may be relevant for explaining justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation beyond their CDK and SDK. Other possible sources of variance are additional kinds of knowledge used in diagnosing that were not considered in this study, such as scientific knowledge that is not pertinent to the context (e.g., Hetmanek et al., 2015) or subjective theories, beliefs, and epistemic goals (Stark, 2017).

CDK and SDK also did not explain the relationship found between the accuracy of diagnostic judgments and justification in diagnostic argumentation. Beyond a joint knowledge base, another variable that could potentially explain the relationship between accuracy and justification may be the different types of information processing that occur during the judgment process (see Loibl et al., 2020). The literature on dual-process theories (see Kahneman, 2003; Evans, 2008) suggests that controlled information processing results in more conscious and explicable reasoning compared to intuitive information processing (e.g., pattern recognition; see Evans, 2008). Thus, a controlled analysis of evidence during the judgment process could affect the accuracy of diagnostic judgments (see Coderre et al., 2010; Norman et al., 2017) and at the same time facilitate justification in diagnostic argumentation.

Limitations and future research

One methodological limitation that needs to be discussed is the low internal consistency of diagnostic accuracy across diagnostic judgments, which may hide further correlations that were not observed in the results. Low internal consistency values are a common issue in measurement instruments with small numbers of items (e.g., Monteiro et al., 2020). However, we did not assume that low internal consistency was a major issue for our interpretations because we still found the theoretically expected relations of diagnostic accuracy with the variables CDK and SDK.

The operationalization of the judgment process in the simulation-based learning environment might be considered to limit generalizability to real-life practice situations, in which teachers' judgment processes might take place over several days or weeks and involve higher degrees of complexity and ambiguity compared to our simulated cases. However, in our simulation, preservice teachers could decide by themselves how much evidence they wanted to collect, and in which order they would access which informational sources (e.g., conversation protocols). Therefore, we argue that, for the purpose of our research goals, the simulation provided a sufficient representation of a real-world diagnostic situation.

Descriptive results of the participants' performance in all three argumentation facets across the measurement points of the different cases suggest that participants' performance generally decreased throughout the data collection. The long duration of the study might have exhausted the participants or decreased their motivation. In addition, some participants might have concluded from the order of the tasks in the simulated cases that they would not need to include their initially indicated diagnostic judgments as a conclusion in their subsequently written diagnostic argumentation texts. Given that the operationalization of justification required participants not only to evaluate evidence but also to explicate conclusions in their argumentation texts, their argumentation skills in terms of justification might have been underestimated in our study. Therefore, generalizing to teachers in authentic classroom situations based on our participants' performance should be done with caution.

There are areas other than students' clinical problems in which teachers' diagnosing is relevant (e.g., assessing a student's level of skill). Our choice of topic might limit the generalizability of the findings to other areas of diagnosing in teacher education. However, we consider the conceptualization of diagnostic argumentation (i.e., justification, disconfirmation, and transparency) presented in this article nonspecific to the content area of clinical problems. Thus, we expect the result pattern to be replicable in other areas of teachers' diagnosing, which could be investigated in further research.

To explore the research questions addressed in this paper, we reanalyzed the data collected in a prior cross-sectional study. The sample was too small to employ structural equation modeling, which would have been preferable to analyzing the data with correlation and regression analyses. Although our results provide initial evidence of the potential relationships between the investigated constructs, they must be replicated in future research using larger samples and advanced methods.

Future research is necessary to further validate the findings that diagnostic argumentation is a diagnostic skill that is distinct from diagnostic judgment. For this purpose, we recommend the approach to investigate preservice teachers' performance based on both qualitative and quantitative data as illustrated in our study. In particular, possible joint predictors of accurate diagnostic judgments and justified diagnostic argumentation, such as controlled information processing during the judgment process, require further clarification because CDK and SDK did not seem to explain the relation between accuracy and justification. Additionally, further research in teacher education should investigate the knowledge and skills that underlie justification, disconfirmation, and transparency beyond CDK and SDK, such as knowledge about standards in diagnosing, cross-domain transferrable argumentation skills, as well as subjective theories, beliefs, and experiential knowledge regarding evidence-oriented practice.

In our study, we did not specify a particular recipient to whom preservice teachers should direct their diagnostic argumentation. However, diagnostic argumentation might vary considerably depending on the recipient (e.g., a teacher colleague, a school psychologist, or a parent) and the argumentative situation (during a collaborative judgment process or subsequent to making a judgment). For example, prior research in collaborative diagnosing has emphasized the potential role of meta-knowledge about the collaborating professional's role and responsibilities (Radkowitsch et al., 2021). Therefore, future studies might systematically investigate the role of different recipients in teachers' diagnostic argumentation.

Research may also validate whether professionals in teacher education perceive justification, disconfirmation, and transparency as facilitating comprehensibility and persuasiveness in diagnostic argumentation or whether our suggested conception of argumentation facets needs to be further specified for the area of teacher education. One interesting and potentially relevant direction in which to further develop our conception might be found in the literature on professional vision, which distinguishes between describing and interpreting evidence as two different forms of how evidence is reported and evaluated in the context of teachers' diagnosing (Seidel and Stürmer, 2014; Kramer et al., 2021b). Moreover, researchers could explore the potential of different learning opportunities and support measures for fostering preservice teachers' diagnostic argumentation. Similarly, researchers could investigate whether diagnostic judgment and diagnostic argumentation have similar or different developmental trajectories and might benefit from similar or different forms of instruction.

Conclusion

In this article, we presented evidence suggesting that diagnostic judgment and diagnostic argumentation might represent different diagnostic skills. Preservice teachers do not necessarily seem to be equally capable of making accurate diagnostic judgments on the one hand, and formulating justified, disconfirming, and transparent diagnostic argumentation on the other. We suggest that justification, disconfirmation, and transparency can be considered relevant facets and distinct subskills of diagnostic argumentation, as our results appear to indicate differences in the underlying knowledge bases. Despite the fact that CDK and SDK explain some variance in justification, disconfirmation, and transparency, the portion of variance they explain might be rather small. Thus, additional variables may be relevant predictors of justification, disconfirmation, and transparency in diagnostic argumentation, such as knowledge of diagnostic standards or cross-domain transferable argumentation skills. Including these additional constructs in further investigations would be a promising direction for future research on diagnostic argumentation. In addition, it seems particularly important that researchers and educators in the field of teacher education, as well as in-service teachers as practitioners in the field, further reflect on standards in diagnosing and diagnostic argumentation. Justification, disconfirmation, and transparency may serve as a productive set of constructs for establishing standards for teachers' diagnostic argumentation in the future.

Data availability statement

The data presented in this article will be made available by the authors upon request. Requests to access the data should be directed to elisabeth.bauer@psy.lmu.de.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Medical Faculty of LMU Munich (no. 17-249). The participants provided their written informed consent to participate in this study.

Author contributions

EB, MS, JK, MF, and FF developed the study concept and contributed to the study design. EB performed the data analysis. EB, MS, and FF interpreted the data. EB drafted the manuscript. MS, JK, MF, and FF provided critical revisions. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the German Federal Ministry of Research and Education (FAMULUS-Project 16DHL1040) and the Elite Network of Bavaria (K-GS-2012-209).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Albritton, K., Chen, C.-I., Bauer, S. G., Johnson, A., and Mathews, R. E. (2021). Collaborating with school psychologists: moving beyond traditional assessment practices. *Young Except. Children* 24, 28–38. doi: 10.1177/1096250619871951

Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., et al. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education: an interdisciplinary comparison. *Front. Psychol.* 11, 1–9. doi: 10.3389/fpsyg.2020.562665

Berland, L. K., and Reiser, B. J. (2009). Making sense of argumentation and explanation. *Sci. Educ.* 93, 26–55. doi: 10.1002/sce.20286

Boshuizen, H. P., Gruber, H., and Strasser, J. (2020). Knowledge restructuring through case processing: the key to generalise expertise development theory across domains? *Educ. Res. Rev.* 29:100310. doi: 10.1016/j.edurev.2020.100310

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., and Fischer, F. (2020). Facilitating diagnostic competences in higher education - a meta-analysis in medical and teacher education. *Educ. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Chinn, C. A., and Duncan, R. G. (2018). "What is the value of general knowledge of scientific reasoning? Scientific reasoning and argumentation: the roles of domain-specific and domain-general knowledge," in *Scientific reasoning and argumentation*. eds. F. Fischer, A. C. Clark, K. Engelmann and J. Osborne (New York: Routledge), 77–101.

Chinn, C. A., Rinehart, R. W., and Buckland, L. A. (2014). "Epistemic cognition and evaluating information: applying the AIR model of epistemic cognition," in *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. eds. D. Rapp and J. Braasch (Cambridge, MA: MIT Press), 425–453.

Coderre, S., Wright, B., and McLaughlin, K. (2010). To think is good: querying an initial hypothesis reduces diagnostic error in medical students. *Acad. Med.* 85, 1125–1129. doi: 10.1097/ACM.0b013e3181e1b229

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., and Seidel, T. (2021). Exploring the process of preservice teachers' diagnostic activities in a video-based simulation. *Front. Educ.* 6:626666. doi: 10.3389/feduc.2021.626666

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Csanadi, A., Kollar, I., and Fischer, F. (2020). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Diamantopoulos, A., and Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. *Br. J. Manag.* 17, 263–282. doi: 10.1111/j.1467-8551.2006.00500.x

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Feldon, D. F. (2007). Cognitive load and classroom teaching: the double-edged sword of automaticity. *Educ. Psychol.* 42, 123–137. doi: 10.1080/00461520701416173

Fischer, F. (2021). Some reasons why evidence from educational research is not particularly popular among (pre-service) teachers: a discussion. *Zeitschrift für Pädagogische Psychologie* 35, 209–214. doi: 10.1024/1010-0652/a000311

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i2.96

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76:378, –382. doi: 10.1037/h0031619

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., et al. (2018). Systematizing professional knowledge of medical doctors and teachers: development of an interdisciplinary framework in the context of diagnostic competences. *Educ. Sci.* 8:207. doi: 10.3390/educsci8040207

Gorman, M. E., Gorman, M. E., Latta, R. M., and Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *Br. J. Psychol.* 75, 65–79. doi: 10.1111/j.2044-8295.1984.tb02790.x

Grossman, P. (2021). *Teaching core practices in teacher education*, Cambridge: Harvard Education Press.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., and Williamson, P. (2009). Teaching practice: a cross-professional perspective. *Teach. Coll. Rec.* 111, 2055–2100. doi: 10.1177/016146810911100905

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., et al. (2019). Facilitating diagnostic competences in simulations: a conceptual framework and a research agenda for medical and teacher education. *Frontline Learn. Res.* 7, 1–24. doi: 10.14786/flr.v7i4.384

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teach. Teach. Educ.* 76, 181–193. doi: 10.1016/j.tate.2017.12.001

Hetmanek, A., Engelmann, K., Opitz, A., and Fischer, F. (2018). "Beyond intelligence and domain knowledge: scientific reasoning and argumentation as a set of cross-domain skills. Scientific reasoning and argumentation: the roles of domain-specific and domain-general knowledge," in *Scientific reasoning and argumentation*. eds. F. Fischer, A. C. Clark, K. Engelmann and J. Osborne (New York: Routledge), 203–226. doi: 10.4324/9780203731826

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M. R., Gräsel, C., et al. (2015). Wozu nutzen Lehrkräfte welche Ressourcen? Eine Interviewstudie zur Schnittstelle zwischen bildungswissenschaftlicher Forschung und professionellem Handeln im Bildungsbereich [for which purposes do teachers use which resources? An interview study on the relation between educational research and professional educational practice]. *Unterrichtswissenschaft* 43, 194–210.

Hitchcock, D. (2005). Good reasoning on the Toulmin model. *Argumentation* 19, 373–391. doi: 10.1007/s10503-005-4422-y

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift für Pädagogische Psychologie* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kiesewetter, J., Fischer, F., and Fischer, M. R. (2017). Collaborative clinical reasoning—a systematic review of empirical studies. *J. Contin. Educ. Health Prof.* 37, 123–128. doi: 10.1097/CEH.0000000000000158

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognit. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Kolovou, D., Naumann, A., Hochweber, J., and Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teach. Teach. Educ.* 100:103298. doi: 10.1016/j. tate.2021.103298

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T., and Neuhaus, B. J. (2021a). Investigating pre-service biology teachers' diagnostic competences: relationships between professional knowledge, diagnostic activities, and diagnostic accuracy. *Educ. Sci.* 11:89. doi: 10.3390/educsci11030089

Kramer, M., Förtsch, C., Seidel, T., and Neuhaus, B. J. (2021b). Comparing two constructs for describing and analyzing teachers' diagnostic processes. *Stud. Educ. Eval.* 68:100973. doi: 10.1016/j.stueduc.2020.100973

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lawson, A. (2003). The nature and development of hypothetico-predictive argumentation with implications for science teaching. *Int. J. Sci. Educ.* 25, 1387–1408. doi: 10.1080/0950069032000052117

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teach. Teach. Educ.* 91:103059. doi: 10.1016/j.tate.2020.103059

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mercier, H., and Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi* 33, 513–524. doi: 10.1007/s11245-013-9217-4

Mercier, H., and Sperber, D. (2017). *The enigma of reason*. Cambridge, Massachusetts: Harvard University Press.

Monteiro, S. D., Sherbino, J., Schmidt, H., Mamede, S., Ilgen, J., and Norman, G. (2020). It's the destination: diagnostic accuracy and reasoning. *Adv. Health Sci. Educ.* 25, 19–29. doi: 10.1007/s10459-019-09903-7

Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., and Mamede, S. (2017). The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Acad. Med.* 92, 23–30. doi: 10.1097/ ACM.0000000000001421

Page, G., Bordage, G., and Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad. Med.* 70, 194–201. doi: 10.1097/00001888-199503000-00009

Pfeiffer, J., Meyer, C. M., Schulz, C., Kiesewetter, J., Zottmann, J., Sailer, M., et al (2019). "FAMULUS: Interactive Annotation and Feedback Generation for Teaching Diagnostic Reasoning," *in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*,. eds. S. Padó and R. Huang Stroudsburg, PA, USA: Association for Computational Linguistics., 73–78

Pit-ten Cate, I. M., Hörstermann, T., Krolak-Schwerdt, S., Gräsel, C., Böhmer, I., and Glock, S. (2020). Teachers' information processing and judgement accuracy: effects of information consistency and accountability. *Eur. J. Psychol. Educ.* 35, 675–702. doi: 10.1007/s10212-019-00436-6

Poznanski, B., Hart, K. C., and Graziano, P. A. (2021). What do preschool teachers know about attention-deficit/hyperactivity disorder (ADHD) and does it impact ratings of child impairment? *Sch. Ment. Heal.* 13, 114–128. doi: 10.1007/ s12310-020-09395-6

Praetorius, A. K., Berner, V. D., Zeinz, H., Scheunpflug, A., and Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *J. Educ. Res.* 106, 64–76. doi: 10.1080/00220671.2012.667010

Radkowitsch, A., Sailer, M., Schmidmaier, R., Fischer, M. R., and Fischer, F. (2021). Learning to diagnose collaboratively–effects of adaptive collaboration scripts in agent-based medical simulations. *Learn. Instr.* 75:101487. doi: 10.1016/j. learninstruc.2021.101487

Rieu, A., Leuders, T., and Loibl, K. (2022). Teachers' diagnostic judgments on tasks as information processing–the role of pedagogical content knowledge for task diagnosis. *Teach. Teach. Educ.* 111:103621. doi: 10.1016/j.tate.2021.103621

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., et al. (2022). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learn. Instr.* 101620, 1–10. doi: 10.1016/j.learninstruc.2022.101620

Sampson, V., and Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: current perspectives and recommendations for future directions. *Sci. Educ.* 92, 447–472. doi: 10.1002/sce.20276

Schmidmaier, R., Eiber, S., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., et al. (2013). Learning the facts in medical school is not enough: which factors predict successful application of procedural knowledge in a laboratory setting? *BMC Med. Educ.* 13, 1–9. doi: 10.1186/1472-6920-13-28

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educ. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Shaffer, D. W. (2017). *Quantitative ethnography*, Madison, Wisconsin: Cathcart Press.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Stadler, M., Sailer, M., and Fischer, F. (2021). Knowledge as a formative construct: a good alpha is not always better. *New Ideas Psychol.* 60:100832. doi: 10.1016/j. newideapsych.2020.100832

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer praxis [problems of evidence-based or rather evidence-oriented educational practice]. *Zeitschrift für Pädagogische Psychologie* 31, 99–110. doi: 10.1024/1010-0652/a000201

Stark, R., Kopp, V., and Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: two experimental studies in undergraduate medical education. *Learn. Instr.* 21, 22–33. doi: 10.1016/j.learninstruc.2009.10.001

Südkamp, A., Praetorius, A.-K., and Spinath, B. (2018). Teachers' judgment accuracy concerning consistent and inconsistent student profiles. *Teach. Teach. Educ.* 76, 204–213. doi: 10.1016/j.tate.2017.09.016

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* 48, 1273–1296. doi: 10.1007/s11165-016-9602-2

Toulmin, S. E. (1958). *The uses of argument*, Cambridge: Cambridge University Press.

Urhahne, D., and Wijnia, L. (2020). A review on the accuracy of teacher judgments. *Educ. Res. Rev.* 32:100374. doi: 10.1016/j.edurev.2020.100374

Van Gog, T., Paas, F., and Van Merriënboer, J. J. (2004). Process-oriented worked examples: improving transfer performance through enhanced understanding. *Instr. Sci.* 32, 83–98. doi: 10.1023/B:TRUC.0000021810.70784.b0

Vanlommel, K., Van Gasse, R., Vanhoof, J., and Van Petegem, P. (2017). Teachers' decision-making: data based or intuition driven? *Int. J. Educ. Res.* 83, 75–83. doi: 10.1016/j.ijer.2017.02.013

Vazire, S. (2017). Quality uncertainty erodes trust in science. Collabra. *Psychology* 3. doi: 10.1525/collabra.74

Walton, D. N. (1990). What is reasoning? What is an argument? *J. Philos.* 87, 399–419. doi: 10.2307/2026735

Wildgans-Lang, A., Scheuerer, S., Obersteiner, A., Fischer, F., and Reiss, K. (2020). Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment. *ZDM* 52, 241–254. doi: 10.1007/s11858-020-01139-9

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 22 November 2022
DOI 10.3389/feduc.2022.996039

Check for updates

OPEN ACCESS

EDITED BY
Ingo Kollar,
University of Augsburg,
Germany

REVIEWED BY
Michael Rochnia,
Universität Wuppertal,
Germany
Katharina Engelmann,
University of Hildesheim,
Germany

*CORRESPONDENCE
Tim Surma
tim.surma@thomasmore.be

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 16 July 2022
ACCEPTED 02 November 2022
PUBLISHED 22 November 2022

CITATION
Surma T, Camp G, de Groot R and Kirschner PA (2022) Novice teachers' knowledge of effective study strategies. *Front. Educ.* 7:996039.
doi: 10.3389/feduc.2022.996039

Novice teachers' knowledge of effective study strategies

Tim Surma[1]*, Gino Camp[2], Renate de Groot[2] and Paul A. Kirschner[1,2]

[1]Expertise Centre for Education and Learning, Thomas More University of Applied Sciences, Antwerp, Belgium, [2]Faculty of Educational Sciences, Open University of the Netherlands, Heerlen, Netherlands

This survey research, assessed whether novice secondary school teachers knew and understood the effectiveness of empirically-supported learning strategies, namely spaced practice, retrieval practice, interleaved practice, using multimodal representations, elaborative interrogation and worked-out examples. These 'proven' strategies can be contrasted with frequently used learning strategies that have been found to be less effective, such as re-reading, taking verbatim notes, highlighting/underlining, summarizing, and cramming. This study broadens previous research on teachers' knowledge of learning strategies by both refining and extending the methodology used in the scenario studies, and by administering it to a different, previously unexplored population. Novice teachers enrolled in a teacher training program (N=180) in Flanders, Belgium were presented with a three-part survey, consisting of open-ended questions, learning scenarios and a list of study strategies. The results show that misconceptions about effective study strategies are widespread by novice teachers and suggests that they are unaware of several specific strategies that could benefit student learning and retention. While popular but less effective strategies such as highlighting and summarising were commonly named by them in open-ended questions, this was not the case for proven effective strategies (e.g., studying worked-out examples, interleaving, and using multi-modal representations) which were not or hardly mentioned. We conclude that this study adds to the growing literature that it is not only students, but also novice teachers who make suboptimal metacognitive judgments when it comes to study and learning. Explicit instruction in evidence-informed learning strategies should be stressed and included in both teacher professional development programs and initial teacher training.

KEYWORDS

learning strategies, study strategies, teaching, teacher education, memory

Introduction

Educators are often asked for advice on how to improve their students' self-study behavior. This requires teachers to expand their teaching of subject-specific information with teaching their students how to best process this information (i.e., how to study; Weinstein and Mayer, 1986). Research into human cognition has provided information on concrete learning strategies that support student learning (e.g., Dunlosky et al., 2013), but has also shown that many learners have flawed mental models of how they learn, making them more likely to mismanage their learning (Bjork et al., 2013). Teachers are in the position to teach students how to optimize their use of study time to promote efficient and effective learning and better retention of knowledge and skills in both generic *learning to learn* lessons or within their subject-specific classes (Education Council, 2006). Since the beginning of the 21st century, learning strategy-instruction has indeed become part of several national curricula (Glogger-Frey et al., 2018). The idea is that teachers' use of the evidence-base on effective study-strategies when advising students can improve students' self-study behavior (see, e.g., Biwer et al., 2022, for a practical implementation of an evidence-based program). To move from the evidence into the actual design of pedagogical practices informed by this best-evidence, it is necessary to have a deep understanding of what, how, and when something works in optimal circumstances. To improve students' study behaviors, it is worth exploiting the most promising guidelines that have been shown to work for the largest possible group of pupils. Implementation of a so-called evidence-informed approach on teaching and learning, based on stable and robust scientific findings (best-evidence), then offers the chance to raise practice (see, e.g., Slavin, 2020). The question is, however, whether novice teachers have this accurate knowledge of the evidence on which they can base their practice. Knowledge, acquired during teacher education, can work as a starting point in their teaching career upon which the can gain further expertise during their ensuing professional career (Berliner, 2001). In this survey research, we assessed whether novice secondary school teachers, who recently graduated from initial teacher training, in Flanders (Belgium) have accurate knowledge of the effectiveness and non-effectiveness of particular study strategies.

Effective strategies for acquiring knowledge and skills

Research on teachers' knowledge is multifaced because of the multiple definitions given to the knowledge itself (Elbaz, 1983; Shulman, 1986; Darling-Hammond and Bransford, 2005). Teachers' knowledge about effective study-strategies is part of what Lee Shulman termed principles of teachers' propositional knowledge (i.e., 'know that', principles derived from empirical research and theory about learning and instruction (Shulman, 1986; Verloop et al., 2001)). Well over a century of laboratory and applied research in cognitive and educational psychology has brought us a number of well-established principles: certain learning strategies promote retention more and lead to more durable learning than others (Pashler et al., 2007; Dunlosky et al., 2013; Fiorella and Mayer, 2016). These strategies can be labeled as study strategies when students independently employ them to promote their learning by achieving goal oriented instructional tasks, often characterized by tests or exams (Winne and Hadwin, 1998; Dinsmore et al., 2016). Many experiments where learners are taught or encouraged to apply specific study strategies, such as rereading, spacing practice, summarizing or highlighting have been conducted to determine if and how they work and to determine which lead to longer-lasting learning (as opposed to achievement on exams). Several key reviews reach converging findings (Pashler et al., 2007; Dunlosky et al., 2013; Putnam and Roediger, 2018; Weinstein et al., 2018). In their extensive review, Dunlosky et al. (2013) discussed 10 frequently used and researched strategies: spaced practice, retrieval practice, interleaved practice, rereading, imagery use for text learning, keyword mnemonic, highlighting, summarization, self-explanation and elaborative interrogation. They assessed the effectiveness of these strategies for different age groups, subject areas, types of learning materials, study tasks and types of learning. Spaced practice and retrieval practice were, amongst others, qualified as useful strategies that promote learning, whereas highlighting, rereading, summarizing and keyword mnemonics were seen as strategies with low utility. Similarly, Pashler et al. (2007) identified seven effective learning and study strategies that overlap considerably with Dunlosky et al. (2013): spaced practice, studying worked examples, combining graphics with verbal descriptions, using concrete representations, retrieval practice and elaborative interrogation. These findings has led the National Council on Teacher Quality (NCTQ) to describe six learning and study strategies as the core of prospective teachers' knowledge base on effective learning processes, as their effectiveness is supported by evidence from multiple sources and replications, ranging from lab-based studies with paired associates as study materials to real classroom-settings with authentic study materials (Pomerance et al., 2016; Weinstein et al., 2018). In Table 1 these six strategies are presented and accompanied by an example of their implementation in students' self-study.

Distributed or spaced practice (i.e., study sessions of the same material are distributed across time) usually improves retention of that material in comparison to massing study of that same material in one long session, keeping total study time equivalent in both conditions. In a typical experiment, Nazari and Ebersbach (2019) compared two groups of secondary school students on learning mathematical calculations (basic probability) in either spaced fashion (i.e., three practice sessions of 15 min on three consecutive days) or massed fashion (i.e., one 45-min session delivered within a single day). Students in the spaced condition outperformed the students in the massed condition on post-tests after 2 and 6 weeks. Distributing practice extends the total time hypothesis (i.e., people tend to learn more as a simple function of time spent on the learning task; Ebbinghaus, 1964) with a timing aspect: introducing spacing gaps between study sessions enhances long term retention. This advantage is known as the spacing effect. For recent reviews, see (Carpenter, 2017; Wiseheart et al., 2019; Latimier et al., 2021).

A related strategy is *interleaved practice*, where learners alternate amongst several separate but related topics during one practice session as compared to blocked practice devoted to a single topic (Firth et al., 2021). When interleaving (also known as variability of practice; Van Merriënboer and Kirschner, 2018), practice of each

TABLE 1 Six effective study strategies applied to students learning.

Study strategy	Practical application
Distributed/spaced practice	Students can plan to restudy course materials on multiple days before an exam, rather than massing their study on the day and night before the exam.
Interleaving practice	After studying negative slopes in graphs, students can switch to studying positive, zero, and undefined slopes; next time, students can study the four in a different order, promoting discrimination and selecting appropriate strategies for problem solving.
Retrieval practice	When learning about social science, students can practice by recalling answers to questions rather than immediately looking up answers in a textbook.
Elaborative interrogation	When students are studying an expository text of the human circulatory cycle, students can ask and explain themselves why and how blood flows in a particular order.
Example-based study	Students can study worked-out examples to self-explain the procedure to solve quadratic equations.
Multi-modal learning	Students combine verbal and pictorial information when learning about the hydrological cycle.

specific topic or task is separated from the next occurrence by the practicing of other topics or tasks. For example, in study sequence A-B-C-B-A-C-A-B-C… there are three tasks between the first and second instance of A, one between the first and second instance of B, and so forth. Thus, by using interleaved practice, learners also achieve spacing effects but the reverse is not necessarily true. Simple spacing (A-A-A-interval-A-A-A) does not lead to interleaving. Interleaving practice is appropriate when students must learn to distinguish among concepts or terms, principles or types of problems that appear to be similar on the surface, or see deeper level similarities in concepts that appear on the surface to be different (e.g., when to use the formulae for acceleration, velocity, and resistance). For recent reviews, see, for example, Firth et al. (2021), Carvalho and Goldstone (2017), and Kang (2016).

Retention is also enhanced when learners engage in *retrieval practice (practice testing)* as a study strategy. Here students retrieve what they have learned either by testing themselves or by being tested by others such as peers or the teacher. Simply put, when students are tested on a particular learning material, they are required to retrieve it from their long term memory to get the correct answer. Note, these are no-stakes tests meant to support learning and not to assess learning (summative testing), to unfold the study process (formative testing) or as a means for self-evaluation. Retrieval strategies have been shown to be superior to non-retrieval strategies such as restudying, re-reading or copying the information, a benefit known as the testing effect (Adesope et al., 2017; Sotola and Crede, 2021).

Elaboration entails study strategies that foster conscious and deliberate/intentional connecting of the to-be-learned material with pre-existing (i.e., prior) knowledge (Hirshman, 2001). To take advantage of elaboration, students can, for instance, engage in what is known as elaborative interrogation (i.e., posing and answering questions about to-be-learned material). The practice of asking epistemic questions such as "why,," "when," and "how," can help increase students' understanding and retention of concepts (Ohlsson, 1996; Popova et al., 2014). Elaborative interrogation demands more than just recall of facts requiring learners to think about information on a deeper level, on such things as causal mechanisms and comparisons between important concepts (Pressley et al., 1987).

Learning from *multimodal representations* of to-be learned material (i.e., complementing text-based study materials with explanatory visual information such as graphs, figures and pictures) facilitates student learning and retention compared to studying single representations. Verbal and pictorial coding has additive effects on recall (Paivio, 1986; Camp et al., 2021; Mayer, 2021). Illustrations are especially helpful when the concept is complex or involves multiple steps (Eitel and Scheiter, 2015).

Finally, students learn more by alternating between studying *worked-out examples* (i.e., studying example problems with their solution) and solving similar problems on their own than they do when just given problems to solve on their own (Kalyuga et al., 2001; Renkl, 2002). Renkl et al., 1998; Kirschner et al., 2006; Van Gog et al., 2019). Students' procedural knowledge can be improved by replacing approximately half the practice problems with fully-worked-out examples and then removing steps, one at a time (i.e., partially worked-out examples) until only the problem remains. A common variation is to combine worked examples with prompts to allow students to explain the information to oneself (Bisra et al., 2018). Connecting concrete examples to more abstract representations also allows students to apply concepts in new situations (Weinstein et al., 2018).

Popular but less effective study strategies

Teachers' propositional knowledge about less effective study-strategies can also be useful; knowing which strategies are less effective should not be ignored in evidence-informed practice (Gorard, 2020). Research has shown that students often employ suboptimal study-strategies such as re-reading, taking verbatim notes, highlighting/underlining, and cramming (see, e.g., Morehead et al., 2016; Anthenien et al., 2018; Dirkx et al., 2019). However, in order to recognize, identify and evaluate these strategies when used by their students in order to eventually correct this, it is necessary for teachers to have an accurate understanding of them. Suboptimal strategies can be misleading when it comes to allocating study time in self-paced learning (Dunlosky et al., 2013).

Re-reading texts, an often used and suggested study strategy, is a passive study strategy as it does not require effortful processing of the text (Morehead et al., 2016; Dirkx et al., 2019). Moreover, it provides students with the false impression of successful learning due to the increased perceived fluency at second reading of the text (Rawson and Dunlosky, 2002). That is to say, when reading a text for a second time students recognize the information in the text but this is quite different from being able to remember it. A similar manifestation of this metacognitive overconfidence can be observed with students *copying* or rewriting notes or texts (Kobayashi, 2005). Here students passively engage in often verbatim copying of information which does not require a type of processing of the information that stimulates long-term retention, such as elaboration or retrieval processes. *Highlighting* or *underlining* is a popular study strategy because of its ease of use and its assumed potential for assisting the storage for important sections in text materials (Morehead et al., 2016; Dirkx et al., 2019). Although there is evidence to suggest that students recall highlighted information better than the non-highlighted information, in general, students' highlighting habits are mostly ineffective as they usually underline unessential information, or too much or too little information (Ponce et al., 2022). *Cramming* is a widely used study strategy where students mass their study sessions directly prior to exams or tests (Hartwig and Dunlosky, 2012). Massing study sessions, though fruitful for recall at a short retention interval (i.e., performance on a test), yields sub-standard recall in the long-term (i.e., learning).

Although summarizing and concept mapping could be seen as potential examples of active and generative study strategies (Fiorella and Mayer, 2016), the results of their use are often disappointing. Summarizing is the act of concisely stating key ideas from to-be-learned material using one's own words and excluding irrelevant or repetitive material. While summarizing is effective in certain domains and study tasks (e.g., summarizing short expository texts for history lessons), research has shown that there are a few important boundaries (e.g., procedural knowledge in for instance physics and chemistry is not appropriate for creating a summary as is vocabulary learning; Dunlosky et al., 2013). *Concept mapping* might be considered as a form of summarizing where a graphic organizer is created by identifying key words or ideas, by placing them in nodes, by drawing lines linking related terms and by writing about the nature of the relationship along those lines (Schroeder et al., 2018). Similar to summarizing, boundary conditions of the strategy have been identified. For instance, Karpicke and Blunt (2011) found that for studying text passages retrieval practice is more effective than concept mapping while observing the learning materials. Studies have also shown that students can struggle to create summaries or concept maps of sufficient quality if they have only received basic instructions (e.g., capturing the main points and on excluding unimportant material, see Rinehart et al., 1986; Bednall and James Kehoe, 2011; Schroeder et al., 2018) or have either not sufficiently practiced summarizing or concept mapping so as to acquire the necessary skills to do it well or lack the necessary prior knowledge to identify what is important.

However, as noted by Miyatsu et al. (2018), even the aforementioned more shallow strategies can be tweaked into a more effective approach by enriching or combining them with effective strategies. For instance, rewriting notes by reorganizing them elicits elaborative processing and studying one's summary followed by trying to reproduce it without the summary being visible takes advantage of the benefits of the testing effect. It is known that students who solely engage in less effective strategies (e.g., highlighting without engaging in retrieval practice) tend to reduce their potential of recall and transfer (Blasiman et al., 2017).

Why do students not know what is germane to their learning?

The accumulated knowledge from cognitive psychology about how to study effectively and how to avoid ineffective study strategies does not necessarily lead to improved learning behavior by students. The majority of self-report questionnaires reveals that students are often not aware of the advantages of retrieval practice, spaced practice, and elaboration strategies and do not often implement them in their self-regulated learning. Most students use strategies, such as repeatedly rereading their learning materials or massing their study, which hamper, rather than improve, their effectiveness as learners (see, e.g., Kornell and Bjork, 2007; Karpicke et al., 2009; McCabe, 2011; Hartwig and Dunlosky, 2012; Dirkx et al., 2019). This might be partially explained by two accounts. First, students (and teachers were former students) are susceptible to – often false – metacognitive intuitions or beliefs about learning which influences their knowledge (for an overview of biases and classic beliefs in human learning, see, e.g., Koriat, 1997; Bjork et al., 2013). For instance, monitoring judgments of learning is typically based on cognitive cues that learners consider to be predictive for their future memory performance, that is, they confuse initial performance with learning for long-term maintenance (Soderstrom and Bjork, 2015). Ineffective strategies such as massed practice (as opposed to spaced practice), blocked practice (as opposed to interleaved practice), rereading (as opposed to elaboration and retrieval practice) intuitively seem to be more satisfying and fluent because the learner makes quicker gains during initial study. These quick gains create "illusions of learning" such as the stability bias which make learners believe that their future performance will remain as high as during initial study (Kornell and Bjork, 2009).

Study strategies such as spaced practice, interleaved practice and retrieval practice reduce this illusion of learning. They can be grouped under the overarching concept of desirable difficulties, learning strategies that initially feel difficult in that they do cause errors and appear to slow down learning, but result in long lasting learning (Bjork, 1994). Even when learners experience memory benefits from these desirable difficulties, earlier research has shown a lack of awareness of the effectiveness of the strategies when predicting their own future learning while using spaced practice (Rawson and Dunlosky, 2011), retrieval practice (e.g., Roediger and Karpicke, 2006), and interleaved practice (e.g., Kornell and Bjork, 2008; Hartwig et al., 2022).

A second reason why students might not use the most effective study techniques is that students never learned how to study effectively or having learnt it, have not properly practiced it so as to make it a part of their repertoire, or struggle to maintain beneficial habits of studying (Fiorella, 2020). One influential source of such information is the teacher, who could provide students with metacognitive instructions (see further). Research suggests that teachers could improve students' knowledge about study strategies by embedding explicit strategy instruction into their subject-content teaching (Putnam et al., 2016; Rivers, 2021). However, several surveys indicate that only 20–36% of students report having been taught about study strategies (Kornell and Bjork, 2007; Hartwig and Dunlosky, 2012). In large international assessments, Flemish students self-report that only 55% of their teachers support their learning processes (OECD, 2019).

The case for explicit strategy instruction

Pintrich (2002) and Muijs and Bokhove (2020) suggest that explicit instruction of study-strategies should consist of pointing out the significance of a strategy ('know that', i.e., conceptual or propositional knowledge), how to employ the strategy in classroom settings ('know how', i.e., prescriptive or procedural knowledge), and monitoring and evaluating proper use of the strategy while providing instructional scaffolds. For instance, students in courses with explicit instruction on implementing retrieval practice in self-study were more likely to use the strategy compared to the control group who did not receive explicit instruction (McCabe, 2011). Biwer et al. (2020) compared two groups of undergraduate students who were randomly assigned to either a 12-week "Study Smart"-program where they received explicit instruction on metacognitive knowledge or a control group. During three sessions students learned about when and why particular learning strategies were effective; reflected on and discussed their strategy use, motivation, and goal-setting; experienced ineffective versus effective strategies (i.e., highlighting versus practice testing) and practiced the strategies in subject-specific courses. Students in the Study-Smart-condition gained more accurate knowledge of effective study strategies (e.g., rated methods based on retrieval practice as more effective and highlighting as less effective) and reported, for instance, an increased use of practice testing and less usage of ineffective study strategies such as highlighting and rereading.

If teachers do not have the propositional knowledge relating to effective study strategies, they cannot be expected to use, model them or explicitly teach students to use them. Willingham (2017) describes this as the necessity to "have a mental model of the learner": because the teacher can recognize the underlying mechanisms in instructional methods or study approaches (e.g., retrieval processes while using flashcards), they can also transfer these strategies to novel situations. Teacher knowledge has indeed been defined as a central element and precursor of teaching competence (for a full discussion on teacher knowledge for teaching and learning, see, e.g., Toom, 2017). Understanding the essential theoretical concepts of the strategies is required to notice, scaffold, and teach strategy-use in generic learning-to-learn courses or subject-specific courses (Glogger-Frey et al., 2018). Earlier studies on the use of research evidence find that teachers pay limited attention to best-evidence findings and rarely consult it to improve their practices (Dagenais et al., 2012; Walker et al., 2019). In addition, there is some evidence that teachers do not begin their careers with this foundational knowledge about effective strategies for learning and study. Research by the National Council for Teaching Quality in United States showed that the way in which essential information on effective learning is covered in the written study material used in in pre-service teacher education programs is inadequate (Pomerance et al., 2016). This was partially replicated by Surma et al. (2018) for Dutch and Flemish teacher education. They found that in general, teacher education textbooks and syllabi do not sufficiently cover essential learning strategies from cognitive psychology or, in some cases, do not cover them at all. For instance, only three teacher education programs (out of 24) provided textbooks and syllabi with a full coverage on spaced practice and retrieval practice (i.e., conceptual information, prescriptive information on how to apply the strategy in regular classrooms, and references to research). Such results indicate that teacher candidates may be under-informed, or not informed by their study materials about effective learning strategies.

In addition to research on the textbooks and syllabi used in teacher education, survey research is an often-used method to gain insight in teachers' knowledge. McCabe (2018) had academic support instructors rate a list of 36 study strategies for their effectiveness on 5-point Likert-scale (from not effective to extremely effective). Several effective study strategies were recognized as effective (e.g., retrieval practice, answering questions, spacing study sessions), whereas some (e.g., multi-modal learning, interleaved practice) were less recognized. Ineffective study strategies (e.g., rereading, copying notes verbatim) consistently had lower ratings. McCabe also asked the instructors to predict the outcomes of four learning scenarios where two contrasting study strategies were contrasted, each one describing a ecologically valid/realistic educational situation. Learning scenarios are a type of vignette-based research, which is becoming more popular in social science studies because it allows respondents to react to context-specific cues such as real-life classroom conditions (Aguinis and Bradley, 2014). The use of learning scenarios to grasp instructors' knowledge has since then been replicated and extended for other populations, such as pre-service teachers and in-service teachers (Halamish, 2018; Firth et al., 2021), university instructors (Morehead et al., 2016) and medical faculty (Piza et al., 2019). The results were mixed, with some educators capable of both identifying some effective strategies (e.g., retrieval practice contrasted with the more passive restudying, Firth et al., 2021) and simultaneously being unsuccessful in distinguishing an effective strategy from a less effective one (e.g., interleaved practice versus blocked practice in all off the aforementioned studies). Table 2 provides a summary of all the studies on metacognitive judgments of learning strategies using scenario-methodology.

Broadening the research base on knowledge of effective study strategies

The current research broadens previous research on teachers' knowledge of effective study strategies by both refining and extending the methodology used in the earlier scenario studies, and by administering it to a different, previously unexplored population (i.e., novice teachers). Earlier studies using the scenario-method had some limitations regarding the number of study strategies being assessed/rated, the sampling method used, and the lack of open-ended questions that were presented to the respondents.

First, most studies only examined a limited number of learning scenarios where more effective study strategies were contrasted with less effective ones (Morehead et al., 2016; Halamish, 2018; Firth et al., 2021). Only spacing, testing, and interleaving were included in each study, which are all examples of study strategies within the desirable difficulties paradigm (see Table 2). Other study strategies with a robust evidence base, such as studying worked-out examples, elaboration and using multi-modal representations were rarely or not assessed by teachers. Moreover, retrieval practice, for example, has not been assessed in relation to a non-passive study strategy (such as concept mapping). Increasing the number of scenarios is particularly interesting because the study strategies can also be interpreted as instructional strategies from the teacher's perspective. For instance, teachers can use retrieval practice by integrating regular low stakes quizzes in their classrooms (Agarwal et al., 2021). As such, the knowledge about the strategies in the scenarios also provides insight into the teacher's pedagogical knowledge. In the present research we introduce the participants to seven learning scenario's which tackle all the aforementioned limitations.

Second, in previous studies the sample ranged from pre-service teachers to more experienced teachers (Halamish, 2018; Firth et al., 2021), but did not explicitly gauge the knowledge of novice teachers (i.e., teachers who very recently graduated from teacher training institutions; see participants). This is valuable because novice teachers have not benefitted from wide-ranging practical classroom experience nor professional development programs, both of which might be influential to clarify how human memory works in the classroom. Earlier research did not find significant differences between pre-serve and in-service teachers (Halamish, 2018; Firth et al., 2021). This study adds to a baseline measurement of novice teacher knowledge, which might contribute to the understanding of the impact of teacher education on imparting the essential knowledge and skills to start the profession.

Third, authors of the previous studies indicated that the sampling of teachers was probably not consistently representative due to selection bias arising from convenience sampling (Halamish, 2018; Firth et al., 2021). In Firth's study Firth et al. (2021), data were collected from students in one teacher training college and in-service teachers were sampled using self-selection. Halamish (2018), recruited respondents by self-selection through a call in an online teacher discussion group. We used cluster sampling, where the sample population is selected in groups (clusters) based on location and timing.

Finally, it is also worth pointing out that earlier survey research used closed-answer questioning (McCabe, 2011; Morehead et al., 2016; Blasiman et al., 2017; Halamish, 2018; Firth et al., 2021) and, thus, did not ask for spontaneous recommendations on effective study strategies using open-ended questions (with notable exceptions for McCabe (2018), who asked academic support-centers to prioritize three learning strategies, and Glogger-Frey et al. (2018), who limited their research to comprehension-oriented learning strategies). McCabe (2018) found limited evidence for the use of terms from cognitive psychology (such as retrieval practice or metacognition) which could indicate that the academic support-center heads were not familiar with the evidence-base in the field of effective learning and studying. Open-ended questions examine the respondents' organization of the knowledge schemes present. If teachers have sufficient in-depth knowledge of effective learning strategies, they will be able to prioritize and coherently explain why one strategy

TABLE 2 Metacognitive judgments of learning strategies using scenario-methodology.

	McCabe (2011)	Morehead et al. (2016)	Morehead et al. (2016)	McCabe (2018)	Halamish (2018)	Halamish (2018)	Firth et al. (2021)	Firth et al. (2021)
Country respondents	US	US	US	US	ISR	ISR	UK	UK
respondents	Under-graduate	Undergraduate	University level Instructors	Academic Study-advisors	Pre-service teachers	In-service teachers	Pre-service teachers	In-service teachers
Retrieval vs. restudying	30%	49%	62%	59%	49%	48%	4.82*	4.7*
Interleaving vs. blocking		16%	13%		23%	12%	3.18*	2.93*
Spacing vs. massing	10%	69%	74%	23%	28%	40%	5.27*	4.45*
Dual coding vs. single coding				52%				
Generating vs. non generating				80%				

Percentages of respondents who preferred the study strategy with empirical support in the scenario. *Scores of respondents on a 7-point Likert scale (1, the evidence-based study strategy is not effective - 7: the evidence-based study strategy is very effective).

is preferred to another. It is therefore expected that novice teachers can access their knowledge about effective learning-strategies according to the knowledge structures they possess. One would expect that the most effective learning strategies (such as spaced practice and retrieval practice) would be recalled first (Glogger-Frey et al., 2018). This is especially important because an adequate knowledge organization is predictive for the accessibility of that information at a later stage (Prawat, 1989). Open questions should also be positioned at the beginning of the survey because measuring this coherent knowledge is more challenging when the respondents have not already been shown a list of study strategies: prior knowledge is activated by the list, which can lead to bias in the assessment.

The current study

Taken together, the results of earlier research on teachers' knowledge of study strategies indicates that teacher knowledge might not be sufficient or even available to equip their students with effective study strategies. It is hypothesized that, based on earlier research, novice teachers might not be aware of the effectiveness of study-strategies such as retrieval practice, interleaved practice and spaced practice and that spontaneous study advice might include less effective strategies. Given the methodological concerns in the particular context of survey research in the area of teachers' propositional knowledge of evidence-based study strategies, more research is needed. The present study examines knowledge about the effectiveness of study strategies within novice secondary school teachers and further examines whether these teachers' spontaneous study-strategy advice is underpinned by research into human learning. This study thereby gives insight into the baseline level of knowledge of novice teachers and extends the methodology used in previous research by adding learning scenarios and open-ended questions.

Materials and methods

Participants

Participants were 240 novice teachers who followed an introductory course for novice teachers in secondary education, organized in two provinces in Flanders, Belgium from 19 Flemish teacher education institutions encompassing both bachelor and master-level teacher education programs. Novice teachers were defined, based on the theory of stages of expertise development, as practicing teachers with comparable in-group and between-group professional experience before they reached the stadium of advanced beginners, which is reached at approximately 1.5 years of experience, above which an increased teachers expertise level can be expected (see, e.g., Sabers et al., 1991).

The participants were informed about the research and that the survey data would be used for research purposes. The participants were then asked to consent to their responses being used in this research. One participant did not consent and was excluded from all analyses. Of the remaining 239, 59 participants indicated they had more than 2 years of teaching experience and were excluded from the analysis. This resulted in a final sample of 180 respondents (Median age = 25; SD = 6.5; Mean 25.7; male = 62; female = 118).

Procedure

The survey was administered to a large population at an annual kick-off meeting for all novice teachers in two provinces where 19 teacher education institutions were represented. Permission from the Flemish pedagogical support network was asked and obtained to conduct the research. There was no response bias, as most teachers attending the meeting were expected to participate by their school leaders. The pen and pencil survey, which took approximately 30 min to complete, was administered live during the meeting. The survey was completed by the participants anonymously.

The open-ended questions were placed at the start of the survey in order to identify the study strategies that teachers would 'spontaneously' recommend (i.e., recall from their long-term memory) before being primed by the learning scenarios or lists of study strategies. Respondents then completed the second part (i.e., seven learning scenarios) and the final part (i.e., study strategy list) of the survey before providing demographic information (age, gender, type of teacher education, teacher education institute, years of teaching experience, subject-domain of teaching). Respondents were restricted from viewing the remaining parts of the survey and could not return to earlier answered questions to limit prior questions influencing subsequent answers.

Materials

The instrument used in this study consisted of three major parts: open-ended questions on study strategy advice; learning scenarios based on the learning scenarios as described by McCabe (2011, 2018), Morehead et al. (2016), and Halamish (2018); and a list of study strategies (based on McCabe, 2018) that respondents had to rate for effectiveness.

Open-ended questions

First, participants were asked to write down three study strategies they would recommend to their students to help them pass a subsequent test. They were instructed to think about general, not subject-specific study strategies. So as not to influence respondents with the direction of their response, no answer categories were provided. In the second open-ended question, more context-specific cues were added by articulating that the test would take place in 3 weeks and the student had already studied the material once, prompting participants to deliberately consider spaced and/or retrieval practice as preferred study strategy. For the second question, teachers were asked to recommend one single study strategy to their students.

Following the open design of qualitative studies (Creswell and Creswell, 2017), the data from the first open question was analyzed before moving to the second open question. First, the first author read every answer to gain a general overview. As a second step, the first author followed a process of mixed coding, both theoretical (i.e., based on the 15-category coding scheme of Dirkx et al., 2019, as described below) and *in vivo* (i.e., based on the participants' responses). Third, after coding 20 questions, the first author cleaned the codes, and made a final lists of codes with relevant example statements. This process resulted in a coding scheme consisting of 16 categories. Fourth, data from the free-response question about the three most recommended study strategies were then classified into 16 categories. The first 10 codes in the coding frame by Dirkx et al. (2019) correspond to the 10 learning strategies discussed by Dunlosky et al. (2013). The following four codes correspond to strategies that were not covered by the above-mentioned article but are often reported as being used as a study strategy by students. The categories added by Dirkx and colleagues were copying (i.e., copying of course materials; see also Blasiman et al., 2017), generating examples (see Karpicke et al., 2009), cramming (as opposed to spaced practice), and solving practice problems (i.e., solving problems provided in students' learning materials such as textbooks and electronic learning environments). Another final category was added after the second phase of coding, namely the code in which recommendations are collected that form the 'behind-the-scenes of studying' and that are not dominated by information-processing, such as time-management, avoidance of behaviors counterproductive for learning, concentration, study aids, attitude, self-discipline, intrinsic or extrinsic motivation (Credé and Kuncel, 2008). These constructs are also important in research on learning and metacognition but go beyond the scope of this article, which focuses on cognitive learning strategies that facilitate long-term learning.

After the coding process, the percentage of teachers with a response in each category was calculated. When teachers specified more study strategies than asked for, the additional strategies were nevertheless included in the results. For instance, the recommendation students should test themselves several times before the test, consists of two study strategies (i.e., retrieval practice and spaced practice). Two researchers assessed 25% of the surveys whether the students' responses were an example of one of the strategies that would fit into the coding frame of Dirkx et al. (2019). The coders discussed their findings, and intercoder reliability was found to be 82%, which was satisfactory. When inconsistencies were uncovered, the researchers re-reviewed the recommendations until they reached agreement. To establish intercoder reliability, the researchers reanalyzed the same selection of responses after a period of 4 months and obtained a 96% level of agreement with previous coding results. The first author coded the remaining surveys twice.

Learning scenarios

The second part of the survey consisted of seven hypothetical study scenarios, each describing two students using two different study-strategies, one empirically validated as being effective and one not. Each scenario was based on a educationally relevant study that investigated the effectiveness of study strategies (see Table 3). The participants were asked which strategy they would recommend to their students to achieve long-term learning (i.e., better outcomes as measured by delayed-test scores) given a particular situation.

For example, one scenario contrasting spaced practice and massed practice presented the following situation: Two students are preparing for a written test in 3 weeks. They have to study one chapter, comprising both theory and practice problems. Student A spaces their practice and study over the 3 weeks. Student B studies and practices intensively just prior to the test (i.e., the night before). All told, they study an equal amount of time. Rate the effectiveness of both students' study strategies for long term retention.

In each scenario, participants used a 5-point Likert-scale to score each strategy of each student in the scenarios. The use of separate scores per strategy made it possible to assess both the absolute perceived effectiveness of each strategy and the difference in perceived effectiveness between the strategies. The authentic context provided in the scenarios was designed to activate prior knowledge about cognitive learning strategies. The retrieving and interleaving scenarios were drawn from previous surveys (McCabe, 2011, 2018; Halamish, 2018) with minor modifications in wording to make the learning scenarios more appropriate for Flemish respondents. This can be seen as replicating and extending the evaluation of learning scenarios presented in the aforementioned studies. The remaining five scenarios were novel (spacing vs. massing; worked examples vs. problem solving; dual coding vs. single coding; elaborative interrogation vs. rereading; retrieving vs. mind mapping), with similar style and length, and were reviewed by a team of international experts in cognitive science and translatory research in order to validate their contents. After an iterative process of three rounds of feedback, full consensus was reached on the content and wording of the new scenarios.

TABLE 3 Seven learning scenarios.

Comparison of study strategies (effective versus less effective)	Inspired by
1. Retrieving vs. restudying	Roediger and Karpicke (2006) Experiment 1.
2. Spacing vs. Massing	Carpenter et al. (2009)
3. Interleaving vs. blocking	Rohrer and Taylor (2007)
4. Worked examples vs. problem solving	Sweller and Cooper (1985)
5. Dual coding vs. single coding	Mayer and Gallini (1990)
6. Elaborative interrogation vs. rereading	Smith et al. (2010)
7. Retrieving vs. mindmapping	Karpicke and Blunt (2011)

Study strategy list

In the final part of the survey, participants were provided with a list of 22 specific study strategies (obtained and adapted from McCabe, 2018) and they were asked to rate on a 5-point Likert scale, on average, how effective they thought each strategy was for their students' learning. The list was slightly refined by adding some elaborated comments to the initial statements by McCabe. For example, in the original study strategy list 'Using pictures' is adapted to 'Search pictures in order to clarify difficult concepts', as the first statement did not describe how pictures should be used.

Statistical analyses

All survey data was analyzed *via* SPSS. The alpha level was set to 0.05 for all statistical tests reported. For the analysis of the learning scenarios, paired samples t-tests were used to compare the mean ratings given to the empirically validated and non-empirically validated study strategies for each scenario and the resulting effect size are reported with Cohen's *d* (De Winter and Dodou, 2010). Positive effect sizes showed effects supporting the evidence-based study strategy, while negative effect sizes showed effects supporting the non-evidence-based strategy. Hinge-points for small, medium or large effects were 0.2, 0.5, and 0.8, respectively. The data from the seven scenarios were combined to form an overall accuracy score for each participant. For each scenario question, each individual participant was coded as a 0 if the non-empirically validated strategy was given a higher rating than the empirically validated strategy and a 1 if the empirically validated scenario was given a higher rating than the non-empirically validated scenario. Accuracy scores ranged from a minimum score of 0 (zero correct scenario judgments) to a maximal score of 7 (all scenarios were judged correctly). The overall accuracy comparing groups (e.g., masters vs. bachelors and gender) across all scenarios were calculated *via* chi-square tests.

For the analysis of the study scenarios, descriptive statistics were calculated. Paired *t*-tests were used to compare items that rely on the same strategy (e.g., "test yourself with practice tests" and "use flashcards to test yourself" both rely on the testing effect).

Results

To identify relevant clustering in the dataset, a number of exploratory analyses were first carried out. There were no significant results from analyses comparing correct strategy endorsements from the learning scenarios among self-reported teacher education types (collapsing into three categories for universities, universities of applied sciences or adult education programs; ($\chi^2 = 6.141$; $p > 0.05$)), nor bachelor/master level ($\chi^2 = 4.872$; $p = 0.56$) nor were strategy endorsements correlated with teachers years of experience (i.e., 0 or 1 year teaching experience; $\chi^2 = 6.244$; $p = 0.396$) nor were strategy endorsements correlated with age ($\chi^2 = 154.732$; $p = 0.256$) or gender ($\chi^2 = 6.620$; $p = 0.357$). It was not possible to compare the various teacher education institutions and subject domains due to a limited number of respondents per teacher education institution or subject domain. As a result, associations with the demographic factors mentioned earlier will not be examined further.

Learning strategy recommendations

For a full overview of the top-three recommendations that would be given to students if they were studying for a test, see Table 4. Here, we present the most notable results: Summarization was advised by 95% of the teachers. Less than half suggested taking a practice test and only 19 (10%) explicitly mentioned that repeating the subject matter in more than one session (spaced practice) was advantageous. In contrast, 38 teachers (21%) said that students should cram the material just before the test. Self-explanation was a relatively often suggested strategy (39%), especially in the context of trying to explain the subject matter to yourself or explaining it to someone else. Some effective strategies, such as studying worked examples, interleaving, and using multimodal representations were not or hardly mentioned.

Less effective study strategies such as copying notes, using mnemonics or re-reading were given less attention. When

TABLE 4 The frequency of recommended study strategies per open question.

	Open question 1 Study advice for a test (3 answers per participant allowed)	Open question 2: Study advice for a test in 3 weeks (1 answer per participant allowed)
Summarizing	172 (95%)	27 (15%)
Practice testing	81 (45%)	15 (8%)
Self-explaining	71 (39%)	7 (4%)
Highlighting	59 (33%)	2 (1%)
Cramming	38 (21%)	5 (3%)
Doing practice problems	34 (19%)	57 (32%)
Rereading	30 (17%)	10 (6%)
Elaborative interrogation	18 (10%)	0 (0%)
Spaced practice	19 (10%)	95 (53%)
Organizational & practical advice	11 (6%)	26 (14%)
Copying	10 (6%)	0 (0%)
Keyword mnemonics	6 (3%)	0 (0%)
Imagery use – multimodal coding	4 (2%)	0 (0%)
Thinking of real-life examples	4 (2%)	1 (1%)
Interleaved practice	1 (1%)	0 (0%)
	3.08 advices per teacher	1.34 advice per teacher

The first figure in each cell indicates the absolute frequency of how many times a particular strategy was recommended by a respondent. The second figure indicates the percentage of respondents who recommended the study strategy.

highlighting was mentioned as a recommendation (33%), it was in combination with another study strategy, such as rereading and summarizing (e.g., "highlight the most important information while rereading"; "make a summary using the highlighted text.").

In the second open question, when it was explicitly stated that the test would only take place in 3 weeks and students had already studied once, spacing of study moments was explicitly mentioned by 53% of the teachers. Taking a practice test, however, was only suggested by 15 teachers (8%). Note that teachers were only allowed to provide one study-advice on the second open-ended question. Less effective study strategies such as copying notes, highlighting and cramming were hardly mentioned. Similar to McCabe (2018), there was limited evidence for the use of terms originating from cognitive or educational psychology in both open questions; that is, there was no mention of concepts such as "retrieval," "metacognition," "testing effects," etc.

Learning scenarios

Novice teachers in the current study made predictions about learning outcomes for scenarios representing seven evidence-based study strategies. In Table 5 the descriptive and inferential statistics per scenario are presented. In all cases, the responses ranged from the minimum (1) to the maximum (5). For five of the seven scenarios, participants provided mean ratings indicating their endorsement of the evidence-based strategy. Interleaved practice and retrieval practice were not seen as being effective in scenarios when compared with blocked practice and mind mapping, respectively. Retrieval practice was judged as being effective in comparison with restudying.

Study strategy list

Ratings of the strategy's perceived effectiveness (rated on a 5-point scale with 5 indicating highest effectiveness) are found in Table 6. The study strategies that are described in the literature as the least effective (i.e., copying notes, cramming, rereading …) are also rated the lowest by novice teachers. Novice teachers consider study strategies that are based on spaced practice, retrieval practice, elaboration, multimodal representations, and worked examples to be effective. Generative study strategies such as summarizing and mind mapping are also evaluated as being effective. Items 6 "test yourself with practice tests" and 12 "use flashcards to test yourself" which both rely on the underlying mechanism of retrieval practice were not perceived equally effective ($t(179) = 8{,}85$, $p < 0.01$). A similar pattern I for items related to spacing (i.e., items 5 and 11; $t(179) = 3.10$, $p < 0.01$) and interleaving (i.e., items X and X; $t(179) = 2.420$, $p < 0.05$) and rereading (i.e., items 5 and 11; $t(179) = 3.10$, $p < 0.01$). Items concerning elaboration (i.e., items 3 and 7; $t(179) = 0$, $p = 1.00$) and marking (i.e., items X and X; $t(179) = 0.533$, $p = 0.594$) were perceived equally effective.

Discussion

This study explored novice teachers' knowledge of effective study strategies. The results of a three-part survey in which participants were asked to provide study advice for their students (open-ended questions) and assess the effectiveness of given study strategies (closed questions) were presented. The results showed that some misconceptions about effective study strategies are widespread within novice teachers albeit with a dissimilar pattern compared to previous empirical research. The results were consistent across demographic factors. For instance, why teachers who have recently completed a master's program do not tend to have a broader knowledge of effective study strategies. This can be explained by the curriculum used: a master's program in teacher education in Flanders does not encapsulate a more in-depth package of, for instance, educational psychology, but mainly expands subject-specific learning content. Overall, we found two main results. First, there is considerable variability in the perceived effectiveness of the most effective study strategies when comparing answers from open questions (i.e., section 1 of this survey) and closed questions (i.e., sections 2 and 3 in this survey; learning scenarios and study strategy list). Second,

TABLE 5 Mean ratings (and standard deviations) for empirically validated learning strategies (EV) and non-empirically validated learning strategies (non-EV) for the learning scenario questions.

Learning scenario	EV		Non-EV		Comparison		% EV
	M1	SD	M2	SD	T	Cohens *d*	
Testing (EV) vs. restudying	4.07	0.68	2.27	0.83	20.60*	2.37	90
Spacing (EV) vs. Massing	4.44	0.87	2.54	0.87	18.73*	2.18	87
Interleaving (EV) vs. blocking	3.51	1.10	3.39	0.91	0.97	0.12	44
Worked examples (EV) vs. problem solving	4.03	0.98	2.98	1.07	8.36*	1.02	65
Dual coding (EV) vs. single coding	4.67	0.56	2.46	0.89	26.99*	2.97	96
Elaborative interrogation (EV) vs. rereading	4.50	0.64	2.45	0.77	27.54*	2.89	96
Testing (EV) vs. mindmapping	3.43	0.93	4.31	0.70	9.81*	1.07	13

M, Mean; SD, Standard Deviation; EV, empirically validated learning strategy (EV), Non-EV, non empirically validated learning strategy *$p < 0.05$. Responses range from 1 = very ineffective to 5 = very effective.

TABLE 6 Perceived effectiveness of learning strategies as reported by novice teachers.

	Mean	Std. Deviation
1. Use concrete examples to explain difficult concepts.	4.39	0.610
2. Search for images to clarify difficult concepts.	4.35	0.672
3. Study by explaining the subject matter to others.	4.32	0.821
4. Make a summary, mind map or outline of the subject matter.	4.29	0.757
5. Study the same material several times spaced in time.	4.28	0.748
6. Test yourself through practice tests.	4.28	0.652
7. Ask yourself who-what-why-how.. questions.	4.18	0.654
8. Find similarities or differences in the subject matter.	4.18	0.719
9. Use mind maps, summaries or diagrams.	4.14	0.797
10. Practise by answering questions about the subject matter.	4.12	0.617
11. Try to study the same subject repeatedly spaced in time.	4.11	0.781
12. Use flash cards to test yourself.	3.98	0.756
13. Develop mnemonic devices (such as rhymes) while studying	3.97	0.869
14. Study by imagining the material as you study	3.87	0.756
15. Use examples that explain how to solve an exercise	3.69	0.749
16. Mix up exercises of different types	3.67	0.995
17. Vary the order in which you practice within one study session	3.49	0.952
18. Underline or highlight the most important elements of the course material	3.42	0.978
19. Revisit the parts you have underlined or marked	3.39	0.928
20. Read the course material out loud	3.10	1.023
21. Read the course material several times	2.73	0.960
22. Study the subject matter all at once for a longer period of time	2.09	0.777
23. Copy the course material verbatim.	1.72	0.825

Responses range from 1 = very ineffective to 5 = very effective.

teachers often have incomplete knowledge about strategies that do not tend to produce durable learning; they sometimes prefer strategies in their study recommendations that have been shown not to work. In what follows, we elaborate on these two observations.

Perceived effectiveness of the most effective strategies

This study contrasts with prior work in that respondents were asked to answer open-ended questions on effective strategy-use before assessing learning scenarios contrasting two commonly used study strategies. There was considerable variation between strategy recommendations of highly effective study strategies in the open-ended questions (requiring recall from long-term memory) and the endorsement of these strategies in closed questions (possibly requiring only recognition). Results show that the respondents very often - but not always - provided appropriate judgments (i.e., preferring the strategy which is backed up by evidence) when they had to weigh two study-strategies against each other, but the same effective study-strategies were not recommended spontaneously to their students in the open-ended questions. Strong endorsements in learning scenarios does not automatically turn into obvious recommendations. *Spaced practice*, for instance, was mentioned as a strategy by less than half of the teachers after it was prompted in the second open-ended question (i.e., that students had already studied for the test once and that the test would take place within 3 weeks), while the majority of the respondents identified spaced practice as a more effective strategy than massed practice in a learning scenario. A similar tendency was observed in the third section of the survey, where items referring to the spacing effect (i.e., "study the same materials several times spaced in time") were considered highly effective. Likewise, *retrieval practice* was assessed as effective when contrasted with a rather passive study strategy (i.e., rereading) but was suggested as a strategy by less than half of the teachers in the first open-ended question. *Interleaved practice, elaboration, using worked examples,* and *using multi-modal representations* were also marginally recommended in the open-ended questions. However, when they were presented in opposition to a less effective study strategy in the learning scenarios, all except for interleaved practice were appropriately and almost unanimously identified as effective.

If novice teachers were presented forced-choice questions, in many cases they will opt for the right answer, which paints an relatively optimistic picture. That is, they remember or are capable of discerning in a paired comparison what works (i.e., they might possess the tacit knowledge) but cannot freely recall it when only prompted to do so (i.e., they might not possess deep conceptual propositional knowledge). This limits the chance that those not freely recalling the strategy will use the strategy in their teaching repertoire is probably negligible. Possessing certain propositional knowledge is known to precede competently handling the pedagogical skills related to the knowledge areas in real classroom situations (Munby et al., 2001). In optimal circumstances, they should also spontaneously recommend the strategy to their students, which is not entirely the case with strategies such as spaced practice, retrieval practice, interleaved practice, using multimodal representations, and using worked examples. This also confirms the claim for the introduction of open-ended questions as a methodological improvement for measuring learners' knowledge about study strategies: performing well on the learning scenarios does not necessarily imply that teachers spontaneously transfer their knowledge to more ecologically valid settings.

Another noteworthy observation was that even within one study single study strategy such as retrieval practice, there were considerable differences in perceived effectiveness. For instance, concept mapping, which is essentially a generative strategy, is considered to be more effective than retrieval practice in a learning scenario, while the memory benefits of retrieval practice (i.e., engaging immediately in trying to remember after a first reading) are more profitable over time than merely generating concept maps from open books (Karpicke and Blunt, 2011; Camerer et al., 2018). Nevertheless, when contrasted with rereading, retrieval practice yielded superior results. In the first open-ended question, retrieval practice was advised by less than half of the teachers, but we were unable to determine from the responses whether retrieval practice was conceived as merely self-testing (a strategy for self-evaluation at the very end of the study process) or as a study strategy to strengthen one's memory. This suggests that the respondents might not be fully aware of the cognitive principles supporting strategies such as retrieval practice (Rivers, 2021). This limits novice teachers' to generalize the strategies to novel situations and instructional methods (Willingham, 2017). Whether teachers' and learners are aware of the full advantages of retrieval practice and for explanations why retrieval practice is not considered a study strategy but merely an self-evaluation strategy, should be tackled by future research (see, e.g., Rivers, 2021).

When novice teachers had to assess the effectiveness from a list of 36 study strategies, on the whole, the most effective strategies were more often rated higher than those with a weaker evidence-base. A notable exception – again - is interleaving, where both items were rated low in effectivity ("Mix up exercises of different types"; "Alternate the order in which you practice within one study session"). The lower accuracy of the strategy endorsements related to mixing up study sequence (i.e., interleaving) is consistent with earlier research (McCabe, 2018; Firth et al., 2021). Some well recognized study strategies such as interleaving are counterintuitive to people as they pose difficulties during the initial learning process (Bjork, 1994; Clark and Bjork, 2014). Metacognitive insight into desirable difficulties may be different from that of other effective strategies and require explicit instruction and practice as some of the advantages do not appear to be obvious for learners and teachers at first sight (Soderstrom and Bjork, 2015). Conditions of retrieval practice and interleaved practice that often facilitate long-term retention may appear unhelpful in the short term as they appear to impede current performance.

One might suspect that when the information on effective strategies is presented clearly and in contrast to less effective strategies, the former will appear obvious in hindsight. However, this does not explain why Flemish novice teachers assess the study scenarios using desirable difficulties (i.e., spaced practice, retrieval practice, interleaved practice) differently than other populations. Compared to earlier studies with similar scenarios, Flemish novice teachers seem to be notably more accurate in identifying desirable difficulties than their mostly Anglo-Saxon counterparts. Table 2 shows three study scenarios which were replicated for seven different population groups in different countries. The explanation for these differences may be grounded in the fact that novice teachers recently graduated from teacher education and topics regarding memory and cognition are still vivid in their minds. If, however, that was the case, more-effective strategies should have been spontaneously mentioned and more subject-specific terms from cognitive psychology should have been generated in the open questions. This is also at odds with the findings of Surma et al. (2018) on the contents of teacher education textbooks and their accompanying syllabi. This research therefore also identifies possible geographical and curricular issues in surveys on respondents' knowledge: generalization about teachers' mental models of learning over countries and related teacher education curricula do not seem to be self-evident.

Perceived effectiveness of the least effective strategies

The respondents tended to suggest strategies that have been shown not to work while avoiding strategies that do work. For example, the vast majority of novice teachers recommend summarizing as a principal study strategy while this strategy is described by Dunlosky et al. (2013) as a low-utility strategy. In the list of study strategies, however, while summarizing was also seen as highly effective, highlighting and cramming were listed among the least effective strategies. Copying notes was not often spontaneously mentioned in the open questions, nor was it strongly appreciated in the study strategies list.

The reasons for this dispersed perception of effectiveness for summarizing versus copying/ highlighting/cramming may have several explanations. First, Pressley et al. (1989, p.5) stated that "summarizing is not one strategy but a family of strategies." When a participant notes that summarizing is a robust strategy, it is not necessarily known what the participant considers to be summarizing (i.e., declarative knowledge: for one student, summarizing is perceived as schematizing single words while for another it might be making a verbatim transcription of their textbook) and the way in which summarizing proceeds (i.e., procedural knowledge: do I summarize with the textbook open or closed? Do I summarize after I have already studied the material thoroughly? Do I use a summary to review afterwards or to test myself? Do I summarize aloud or in writing?). For copying and highlighting, the conceptual and procedural interpretation appears to be more straightforward. In reality, the manifestations of summarizing as a study strategy are probably more diverse and prone to individual differences than the narrow definition that researchers assign to the concept (Miyatsu et al., 2018). Despite the fact that teachers do prefer summarizing over retrieval practice, this choice is unlikely a symptom of their knowledge of effective studying because learners appear, based on earlier research, not always fully aware of the boundary conditions of certain study strategies (Bjork et al., 2013).

A second explanation as to why the novice teachers spontaneously suggest suboptimal strategies can be found in a theory-practice gap. Study strategies are often studied in cognitive science literature as "singletons," that is as individual and generic phenomena, whereas in ecologically valid situations, a given study strategy is often sequentially linked within a series of other study strategies and linked with specific type of learning content. For instance, a student who first reads the learning material, rereads the material while highlighting relevant information, summarizes, and finishes by testing themselves, uses a number of strategies labeled as less effective (i.e., rereading, highlighting, summarizing). As noted earlier by Miyatsu et al. (2018), ineffective strategies, under certain conditions, can be potent. For example, a strategy labeled as ineffective such as massed or blocked practice will sometimes result in a good performance on an immediate test even though it does little for long-term retention and distracts the learner by providing suboptimal judgments of future learning. So far, research on learning strategies has been fairly myopic, focusing on study strategies in isolation but not often tracing optimal combinations or study arrangements in holistic ecologically valid settings (Dirkx et al., 2019). Follow-up research should look at how learners perceive effective study strategies from a semantic point of view, which strategies they choose depending on the type of learning content or subject area, how they combine study strategies chronologically, and why they do so. A more qualitative research design may be appropriate for this purpose.

Limitations

One must be careful when interpreting the results of this study, because multiple factors could have contributed to the discrepancy between the results of the open-ended and closed questions, and the lack of consistency regarding the perceived effectiveness of the study strategies. The limitations with respect to semantics (i.e., do all respondents interpret the term summarizing identically?), the focus on individual strategies (i.e., students are likely to use more than one study strategy during the study process) and the geographical differences (i.e., Flemish novice teachers score better on scenarios that probe desirable difficulties than respondents from other countries) were outlined earlier. The validity of a measurement instrument is not established in one or two (sets of) studies. For example, in follow-up studies learning scenarios can be added that contrast popular and frequently used strategies such as summarizing with other generative strategies such as mind mapping to gain a more fine grained image of novice teachers' knowledge of study strategies. A more qualitative approach can be used to determine how teachers interpret certain (combinations of) study strategies. Finally, and to state the obvious: Responses are self-reported and may not reflect novice teachers true educational advice given in real classrooms.

Conclusion

There remains a noticeable gap between the typical way learners perceive study strategies and the empirical evidence regarding their effect on learning, and novice teachers seem to be no different than their peers elsewhere. The results from this study add to the growing literature that not only students, experienced teachers, university instructors, and pre-service teachers can be suboptimal in their judgments (Morehead et al., 2016; Halamish, 2018; McCabe, 2018; Firth et al., 2021). Overall, our data suggests that Flemish novice teachers are consistent in evaluating given study strategies (specifically: spaced practice, multimodal representations, and elaboration), but are less able to spontaneously formulate study-advices about the same study strategies. Other aspects of the results are more complex. Novice teachers appeared to be less consistent in their evaluation of study strategies that rely on the desirable difficulties framework. Indeed, strategies with the strongest evidence-base, such as spaced practice and retrieval practice, were not often spontaneously recommended in open-ended questions. Since there are large discrepancies between spontaneously recommended study strategies and the effectiveness scores of the same strategies in closed questions, it is possible that that novice teachers do not yet exhibit a coherent image of the learners cognitive architecture. Student teachers knowledge of learning strategies has been previously described as 'knowledge in pieces' (Glogger-Frey et al., 2018, p. 228), and the same conundrums are found in novice-teachers study strategy knowledge. Our indications of the lack of sophistication in novice teachers' knowledge highlights the need for teaching them about and training them in the use of evidence-based strategies (McCabe, 2018).

Teacher learning and their classroom skills should be seen as and a dynamic process and a continuum rather than an judgment of teachers' knowledge at a fixed time (Blömeke et al., 2015). From the perspective of translational research— the amalgam of processes and activities associated with the use of findings from empirical research to incorporate best-evidence guidelines into everyday practice (see, e.g., Gorard, 2020) – the presented study offers an opportunity to examine curricula in both teacher education and continuing professional development whether they disseminate the most consistent research results regarding learning processes. Where best-evidence is used as part of initial teacher education and continuing professional development curricula, teacher performance is found to be superior (Brown and Zhang, 2016). Explicit strategy instruction in teacher education and continuous professional development may thus provide a tangible solution for this 'knowledge in pieces' in novice teachers. Explicit instruction about the concepts, use and advantages of employing empirically supported learning strategies thus might promote teachers' understanding of the mental model of the learner (Willingham, 2017). As argued by Lawson et al. (2019), learning about learning and cognition should perhaps be seen as a separate knowledge domain so that pre-service and in-service teachers can transfer that propositional knowledge both implicitly and explicitly to their students. It is important that all teachers have deep conceptual knowledge of study strategies as teachers might be considered as 'memory workers' who have the responsibility of teaching their students how learning happens and how to use effective study strategies to create lasting learning.

Data availability statement

The datasets presented in this article are not readily available because in the informed consent, it was stated that the dataset would only be used for this particular research. Requests to access the datasets should be directed to tim.surma@thomasmore.be.

Ethics statement

The studies involving human participants were reviewed and approved by Open University of the Netherlands. The patients/participants provided their written informed consent to participate in this study.

Author contributions

TS, GC, PK, and RG contributed to the conception and the design of the study. TS analyzed the data and wrote, revised, and reviewed sections of the manuscript. GC, RG, and PK provided feedback during the writing process. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to thank Peter Verkoeijen, Pedro Debruyckere, Dominique Sluijsmans, Kristel Vanhoyweghen, Tine Hoof, Eva Maesen, and Laurie Delnoij for their feedback during various stages of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adesope, O. O., Trevisan, D. A., and Sundararajan, N. (2017). Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87, 659–701. doi: 10.3102/0034654316689306

Agarwal, P. K., Nunes, L. D., and Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: a systematic review of applied research in schools and classrooms. *Educ. Psychol. Rev.* 33, 1409–1453. doi: 10.1007/s10648-021-09595-9

Aguinis, H., and Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organ. Res. Methods* 17, 351–371. doi: 10.1177/1094428114547952

Anthenien, A. M., DeLozier, S. J., Neighbors, C., and Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Soc. Psychol. Educ.* 21, 303–322. doi: 10.1007/s11218-017-9412-z

Bednall, T. C., and James Kehoe, E. (2011). Effects of self-regulatory instructional aids on self-directed study. *Instr. Sci.* 39, 205–226. doi: 10.1007/s11251-009-9125-6

Berliner, D. C. (2001). Learning about and learning from expert teachers. *Int. J. Educ. Res.* 35, 463–482. doi: 10.1016/S0883-0355(02)00004-6

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., and Winne, P. H. (2018). Inducing self-explanation: a meta-analysis. *Educ. Psychol. Rev.* 30, 703–725. doi: 10.1007/s10648-018-9434-x

Biwer, F., de Bruin, A., and Persky, A. (2022). Study smart – impact of a learning strategy training on students' study behavior and academic performance. 1–21. *Adv. Health Sci. Educ.* 1–11. doi: 10.1007/s10459-022-10149-z

Biwer, F., Oude Egbrink, M. G., Aalten, P., and de Bruin, A. B. (2020). Fostering effective learning strategies in higher education–a mixed-methods study. *J. Appl. Res. Mem. Cogn.* 9, 186–203. doi: 10.1016/j.jarmac.2020.03.004

Bjork, R. A. (1994). "Memory and metamemory considerations in the training of human beings," in *Metacognition*. eds. J. Metcalfe and A. Shimamura (Cambridge, MA: The MIT Press), 185–205.

Bjork, R. A., Dunlosky, J., and Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annu. Rev. Psychol.* 64, 417–444. doi: 10.1146/annurev-psych-113011-143823

Blasiman, R. N., Dunlosky, J., and Rawson, K. A. (2017). The what, how much, and when of study strategies: comparing intended versus actual study behaviour. *Memory* 25, 784–792. doi: 10.1080/09658211.2016.1221974

Blömeke, S., Hoth, J., Döhrmann, M., Busse, A., Kaiser, G., and König, J. (2015). Teacher change during induction: development of beginning primary teachers' knowledge, beliefs and performance. *Int. J. Sci. Math. Educ.* 13, 287–308. doi: 10.1007/s10763-015-9619-4

Brown, C., and Zhang, D. (2016). Is engaging in evidence-informed practice in education rational? What accounts for discrepancies in teachers' attitudes towards evidence use and actual instances of evidence use in schools? *British Educational Research Journal* 42, 780–801.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644. doi: 10.1038/s41562-018-0399-z

Camp, G., Surma, T., and Kirschner, P. A. (2021). "Foundations of multimedia learning," in *The Cambridge handbook of multimedia learning*. eds. R. E. Mayer and L. Fiorella (Cambridge, UK: Cambridge University Press), 17–24.

Carpenter, S. K. (2017). "Spacing effects on learning and memory," in *Cognitive Psychology of Memory: Learning and Memory. Vol. 2*. ed. J. T. Wixted, (Oxford, UK: Elsevier).

Carpenter, S. R., Mooney, H. A., Agard, J., Capistrano, D., DeFries, R. S., Díaz, S., et al. (2009). Science for managing ecosystem services: beyond the Millennium Ecosystem Assessment. *PNAS* 106, 1305–1312.

Carvalho, P. F., and Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. Journal of experimental psychology. *Learn. Memory Cogn.* 43, 1699–1719. doi: 10.1037/xlm0000406

Clark, C. M., and Bjork, R. A. (2014). "When and why introducing difficulties and errors can enhance instruction" in *Applying science of learning in education: Infusing psychological science into the curriculum*. eds. V. A. Benassi, C. E. Overson and C. M. Hakala. Society for the Teaching of Psychology, 20–30.

Credé, M., and Kuncel, N. R. (2008). Study habits, skills, and attitudes: the third pillar supporting collegiate academic performance. *Perspect. Psychol. Sci.* 3, 425–453. doi: 10.1111/j.1745-6924.2008.00089.x

Creswell, J., and Creswell, D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods approaches* (*5th ed.*) London: Sage Publications.

Dagenais, C., Lysenko, L., Abrami, P., Bernard, R., Ramde, J., and Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: a review of empirical research. *Evid. Policy* 8, 285–309. doi: 10.1332/174426412X654031

Darling-Hammond, L., and Bransford, J. (2005). *Preparing Teachers for a Changing World: What Teachers should Learn and be Able to Do*. San Francisco: Jossey-Bass.

De Winter, J. F., and Dodou, D. (2010). Five-point likert items: t test versus Mann-Whitney-Wilcoxon. *Pract. Assess. Res. Eval.* 15, 1–16. doi: 10.7275/bj1p-ts64

Dinsmore, D. L., Grossnickle, E. M., and Dumas, D. (2016). "Learning to study strategically," in *Handbook of Research on Learning and Teaching. 2nd Edn*. eds. R. E. Mayer and P. A. Alexander. (New York: Routledge).

Dirkx, K. J. H., Camp, G., Kester, L., and Kirschner, P. A. (2019). Do secondary school students make use of effective study strategies when they study on their own? *Appl. Cogn. Psychol.* 33, 952–957. doi: 10.1002/acp.3584

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* 14, 4–58. doi: 10.1177/1529100612453266

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. eds. H. A. Ruger, C. E. Bussenius and E. R. Hilgar. New York, NY: Dover Publications.

Education Council (2006). Recommendation of the European Parliament and the council of 18 December 2006 on key competencies for lifelong learning (Brussels, Official Journal of the European Union), 30 December.

Eitel, A., and Scheiter, K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educ. Psychol. Rev.* 27, 153–180. doi: 10.1007/s10648-014-9264-4

Elbaz, F. (1983). *Teacher thinking: A study of practical knowledge*. New York: Nichols.

Fiorella, L. (2020). The science of habit and its implications for student learning and well-being. *Educ. Psychol. Rev.* 32, 603–625. doi: 10.1007/s10648-020-09525-1

Fiorella, L., and Mayer, R. E. (2016). Eight ways to promote generative learning. *Educ. Psychol. Rev.* 28, 717–741. doi: 10.1007/s10648-015-9348-9

Firth, J., Rivers, I., and Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Rev. Educ.* 9, 642–684. doi: 10.1002/rev3.3266

Glogger-Frey, I., Ampatziadis, Y., Ohst, A., and Renkl, A. (2018). Future teachers' knowledge about learning strategies: Misconcepts and knowledge-in-pieces. *Think. Skills Creat.* 28, 41–55. doi: 10.1016/j.tsc.2018.02.001

Gorard, S. (2020). *Getting Evidence into Education. Evaluating the Routes to Policy and Practice*. London, UK: Routledge.

Halamish, V. (2018). Pre-service and in-service teachers' metacognitive knowledge of learning strategies. *Front. Psychol.* 9:2152. doi: 10.3389/fpsyg.2018.02152

Hartwig, M. K., and Dunlosky, J. (2012). Study strategies of college students: are self-testing and scheduling related to achievement? *Psychon. Bull. Rev.* 19, 126–134. doi: 10.3758/s13423-011-0181-y

Hartwig, M. K., Rohrer, D., and Dedrick, R. F. (2022). Scheduling math practice: students' underappreciation of spacing and interleaving. *J. Exp. Psychol. Appl.* 28, 100–113. doi: 10.1037/xap0000391

Hirshman, E. (2001). "Elaboration in memory," in *International Encyclopedia of the Social & Behavioral Sciences*. eds. N. J. Melser and P. B. Baltes (Oxford, England: Pergamon)

Kalyuga, S., Chandler, P., and Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educ. Psychol.* 21, 5–23. doi: 10.1080/01443410124681

Kang, S. H. (2016). "The benefits of interleaved practice for learning," in *From the Laboratory to the Classroom: Translating the Science of Learning for Teachers*. eds. J. C. Horvath, J. M. Lodge and J. Hattie (London, UK: Routledge).

Karpicke, J. D., and Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* 331, 772–775. doi: 10.1126/science.1199327

Karpicke, J. D., Butler, A. C., and Roediger, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own? *Memory* 17, 471–479. doi: 10.1080/09658210802647009

Kirschner, P. A., Sweller, J., and Clark, R. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1207/s15326985ep4102_1

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemp. Educ. Psychol.* 30, 242–262. doi: 10.1016/j.cedpsych.2004.10.001

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *J. Exp. Psychol. Gen.* 126, 349–370. doi: 10.1037/0096-3445.126.4.349

Kornell, N., and Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychon. Bull. Rev.* 14, 219–224. doi: 10.3758/BF03194055

Kornell, N., and Bjork, R. A. (2008). Learning concepts and categories: is spacing the "enemy of induction"? *Psychol. Sci.* 19, 585–592. doi: 10.1111/j.1467-9280.2008.02127.x

Kornell, N., and Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *J. Exp. Psychol. Gen.* 138, 449–468. doi: 10.1037/a0017350

Latimier, A., Peyre, H., and Ramus, F. A. (2021). Meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educ. Psychol. Rev.* 33, 959–987. doi: 10.1007/s10648-020-09572-8

Lawson, M. J., Vosniadou, S., Van Deur, P., Wyra, M., and Jeffries, D. (2019). Teachers' and students' belief systems about the self-regulation of learning. *Educ. Psychol. Rev.* 31, 223–251. doi: 10.1007/s10648-018-9453-7

Mayer, R. E. (2021). "Cognitive theory of multimedia learning," in *The Cambridge Handbook of Multimedia Learning. 3rd Edn*. eds. R. E. Mayer and L. Fiorella. (New York, NY: Cambridge University Press).

Mayer, R. E., and Gallini, J. K. (1990). When is an illustration worth ten thousand words? *J. Educ. Psychol.* 82, 715–726. doi: 10.1037/0022-0663.82.4.715

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Mem. Cogn.* 39, 462–476. doi: 10.3758/s13421-010-0035-2

McCabe, J. (2018). What learning strategies do academic support centers recommend to undergraduates? *J. Appl. Res. Mem. Cogn.* 7, 143–153. doi: 10.1016/j.jarmac.2017.10.002

Miyatsu, T., Nguyen, K., and McDaniel, M. A. (2018). Five popular study strategies: their pitfalls and optimal implementations. *Perspect. Psychol. Sci.* 13, 390–407. doi: 10.1177/1745691617710510

Morehead, K., Rhodes, M. G., and DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory* 24, 257–271. doi: 10.1080/09658211.2014.1001992

Muijs, D., and Bokhove, C. (2020). *Metacognition and Self-regulation: Evidence Review*. London. Available at: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/metacognition-and-self-regulation-review/ (Accessed January 15, 2022).

Munby, H., Russell, T., and Martin, A. K. (2001). "Teachers' knowledge and how it develops," in *Handbook of Research on Teaching. 4th Edn*. ed. V. Richardson. (Washington DC: American Educational Research Association).

Nazari, K. B., and Ebersbach, M. (2019). Distributing mathematical practice of third and seventh graders: applicability of the spacing effect in the classroom. *Appl. Cogn. Psychol.* 33, 288–298. doi: 10.1002/acp.3485

OECD (2019). *PISA 2018 results (volume I): What students know and can do*. Paris: OECD Publishing.

Ohlsson, S. (1996). "Learning to do and learning to understand," in *Learning in Humans and Machines*. eds. P. Reimann and H. Spada (Oxford, UK: Pergamon).

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. Oxford, UK: Oxford University Press.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., et al. (2007). *Organizing Instruction and Study to Improve Student Learning (NCER 2007–2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Pract.* 41, 219–225. doi: 10.1207/s15430421tip4104_3

Piza, F., Kesselheim, J. C., Perzhinsky, J., Drowos, J., Gillis, R., Moscovici, K., et al. (2019). Awareness and usage of evidence-based learning strategies among health professions students and faculty. *Med. Teach.* 41, 1411–1418. doi: 10.1080/0142159X.2019.1645950

Pomerance, L., Greenberg, J., and Walsh, K. (2016). *Learning about Learning: What Every New Teacher Needs to Know*. Washington, D.C: National Council on Teacher Quality.

Ponce, H. R., Mayer, R. E., and Méndez, E. E. (2022). Effects of learner-generated highlighting and instructor-provided highlighting on learning from text: a meta-analysis. *Educ. Psychol. Rev.* 34, 989–1024. doi: 10.1007/s10648-021-09654-1

Popova, A., Kirschner, P. A., and Joiner, R. (2014). Effects of primer podcasts on stimulating learning from lectures: how do students engage? *Br. J. Educ. Technol.* 45, 330–339. doi: 10.1111/bjet.12023

Prawat, R. S. (1989). Promoting access to knowledge, strategy, and disposition in students: a research synthesis. *Rev. Educ. Res.* 59, 1–41. doi: 10.3102/00346543059001001

Pressley, M., Johnson, C. J., Symons, S., McGoldrick, J. A., and Kurita, J. A. (1989). Strategies that improve children's memory and comprehension of text. *Elem. Sch. J.* 90, 3–32. doi: 10.1086/461599

Pressley, M., McDaniel, M. A., Turnure, J. E., Wood, E., and Ahmad, M. (1987). Generation and precision of elaboration: effects on intentional and incidental learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 13, 291–300. doi: 10.1037/0278-7393.13.2.291

Putnam, A. L., and Roediger, H. L. (2018). "Education and memory: seven ways the science of memory can improve classroom learning," in *The Steven's Handbook of Experimental Psychology and Cognitive Neuroscience*. ed. J. T. Wixted (New York: Wiley).

Putnam, A. L., Sungkhasettee, V. W., and Roediger, H. L. (2016). Optimizing learning in college: tips from cognitive psychology. *Perspect. Psychol. Sci.* 11, 652–660. doi: 10.1177/1745691616645770

Rawson, K. A., and Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 69–80. doi: 10.1037/0278-7393.28.1.69

Rawson, K. A., and Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? *J. Exp. Psychol. Gen.* 140, 283–302. doi: 10.1037/a0023956

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learn. Instr.* 12, 529–556. doi: 10.1016/S0959-4752(01)00030-5

Renkl, A., Stark, R., Gruber, H., and Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemp. Educ. Psychol.* 23, 90–108. doi: 10.1006/ceps.1997.0959

Rinehart, S. D., Stahl, S. A., and Erickson, L. G. (1986). Some effects of summarization training on reading and studying. *Read. Res. Q.* 21, 422–438. doi: 10.2307/747614

Rivers, M. L. (2021). Metacognition about practice testing: a review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educ. Psychol. Rev.* 33, 823–862. doi: 10.1007/s10648-020-09578-2

Roediger, H. L., and Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Rohrer, D., and Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instr. Sci.* 35, 481–498. doi: 10.1007/s11251-007-9015-8

Sabers, D. S., Cushing, K. S., and Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensional, and immediacy. *Am. Educ. Res. J.* 28, 63–88. doi: 10.3102/00028312028001063

Schroeder, N. L., Nesbit, J. C., Anguiano, C. J., and Adesope, O. O. (2018). Studying and constructing concept maps: a meta-analysis. *Educ. Psychol. Rev.* 30, 431–455. doi: 10.1007/s10648-017-9403-9

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.2307/1175860

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educ. Psychol.* 55, 21–31. doi: 10.1080/00461520.2019.1611432

Smith, B. L., Holliday, W. G., and Austin, H. W. (2010). Students' comprehension of science textbooks using a question-based reading strategy. *J. Res. Sci. Teach.* 47, 363–379. doi: 10.1002/tea.20378

Soderstrom, N. C., and Bjork, R. A. (2015). Learning versus performance: an integrative review. *Perspect. Psychol. Sci.* 10, 176–199. doi: 10.1177/1745691615569000

Sotola, L. K., and Crede, M. (2021). Regarding class quizzes: a meta-analytic synthesis of studies on the relationship between frequent low-stakes testing and class performance. *Educ. Psychol. Rev.* 33, 407–426. doi: 10.1007/s10648-020-09563-9

Surma, T., Vanhoyweghen, K., Camp, G., and Kirschner, P. A. (2018). The coverage of distributed practice and retrieval practice in Flemish and Dutch teacher education textbooks. *Teach. Teach. Educ.* 74, 229–237. doi: 10.1016/j.tate.2018.05.007

Sweller, J., and Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cogn. Instr.* 2, 59–89. doi: 10.1207/s1532690xci0201_3

Toom, A. (2017). "Teachers' professional and pedagogical competencies: a complex divide between teacher work, teacher knowledge and teacher education," in *The SAGE Handbook of Research on Teacher Education*. eds. J. Clandinin and J. Husu (London, UK: Sage Publications).

Van Gog, T., Rummel, N., and Renkl, A. (2019). "Learning how to solve problems by studying examples," in *The Cambridge Handbook of Cognition and Education*. eds. J. Dunlosky and K. A. Rawson (Cambridge, UK: Cambridge University Press).

Van Merriënboer, J. J. G., and Kirschner, P. A. (2018). *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design* (*3rd Rev. ed*). New York: Routledge.

Verloop, N., van Driel, J., and Meijer, P. (2001). Teacher knowledge and the knowledge base of teaching. *Int. J. Educ. Res.* 35, 441–461. doi: 10.1016/S0883-0355(02)00003-4

Walker, M., Nelson, J., Bradshaw, S., and Brown, C.. (2019). Teachers' engagement with research: what do we know? Available at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Teachers_engagement_with_research_Research_Brief_JK.pdf (Accessed November 7, 2022).

Weinstein, Y., Madan, C. R., and Sumeracki, M. A. (2018). Teaching the science of learning. *Cogn. Res.* 3, 2–17. doi: 10.1186/s41235-017-0087-y

Weinstein, C. E., and Mayer, R. E. (1986). "The teaching of learning strategies," in *Handbook of Research in Teaching*. ed. C. M. Wittrock (New York, NY: Macmillan Publishing Company).

Willingham, D. T. (2017). A mental model of the learner: teaching the basic science of educational psychology to future teachers. *Mind Brain Educ.* 11, 166–175. doi: 10.1111/mbe.12155

Winne, P. H., and Hadwin, A. F. (1998). "Studying as self-regulated learning," in *Metacognition in Educational Theory and Practice*. eds. D. J. Hacker, J. Dunlosky and A. C. Graesser (Mahwah, NJ: Erlbaum).

Wiseheart, M., Kim, A. S. N., Kapler, I. V., Foot-Seymour, V., and Küpper-Tetzel, C. E. (2019). "Enhancing the quality of student learning using distributed practice," in *The Cambridge Handbook of Cognition and Education*. eds. J. Dunlosky and K. A. Rawson (Cambridge, UK: Cambridge University Press).

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/feduc.2022.996029

Check for updates

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University,
Germany

REVIEWED BY
Vicki S. Napper,
Weber State University,
United States
Samuel Merk,
University of Tübingen,
Germany

*CORRESPONDENCE
Hendrik Lohse-Bossenz
hendrik.lohse-bossenz@ph-heidelberg.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 16 July 2022
ACCEPTED 13 October 2022
PUBLISHED 24 November 2022

CITATION
Lohse-Bossenz H, Bloss C and Dörfler T (2022) Constructing multi-theory vignettes to measure the application of knowledge in ambivalent educational situations.
Front. Educ. 7:996029.
doi: 10.3389/feduc.2022.996029

Constructing multi-theory vignettes to measure the application of knowledge in ambivalent educational situations

Hendrik Lohse-Bossenz*, Christopher Bloss and Tobias Dörfler

Institute for Psychology, University of Education Heidelberg, Heidelberg, Germany

Research on evidence-based argumentation shows that (pre-service) teachers have difficulties in orienting their actions to existing theories and empirical evidence. This article addresses the knowledge content needed for this and presents a vignette-based procedure. Within each vignette, two different theoretical perspectives are addressed. The behavior of a teacher can be either suitable or unsuitable from both perspectives or more or less suitable depending on the perspective. In study 1, the procedure is piloted and in study 2, an intervention on a specific area of knowledge takes place. The results show that participants differentiate the vignettes as expected. The intervention leads to corresponding increases in knowledge, which likely relates to a change in the evaluations. The presented approach is discussed with regard to possible applications in the context of research on evidence-based argumentation.

KEYWORDS

evidence-based argumentation, teacher education, vignettes, measurement approach, theoretical knowledge

Introduction

The decisions involved in planning, delivering, and evaluating school lessons are characterized by high levels of uncertainty (Floden and Buchmann, 1993). In the face of this uncertainty, teachers may rely on a variety of sources to make their pedagogical decisions: scientific theories, scientific evidence, subjective theories, beliefs, anecdotes, recipes, or even gut feelings (Stark, 2017; Kiemer and Kollar, 2021). Given that information linked to these specific sources is acquired *via* specific knowledge-building processes, its epistemic status varies, for example, with respect to trustworthiness and credibility (Fenstermacher, 1994). Although the idea that scientific evidence might be valuable in solving practical problems is controversial

(Brown and Rogers, 2015), the use of educational evidence to explicate the reasons for pedagogical judgments seems to be beneficial, at least in cases of classroom problems that appear to occur repeatedly.

Several sets of findings indicate that (pre-service) teachers encounter challenges on two levels of scientific reasoning (Csanadi et al., 2021). On the process level, they may struggle to engage in the inquiry process (Klahr and Dunbar, 1988) or to follow the trajectory of epistemic processes suggested by Fischer et al. (2014). This means that they might not collect enough evidence before engaging in evaluation of that evidence; as a result, the process is unsystematic and speculative. On the content level, they may not be able to relate scientific knowledge from relevant domains to actual classroom incidents, because they lack the requisite knowledge that would enable them to make this transfer, or to do so in a suitable manner (Brown and Rogers, 2015; Hetmanek et al., 2015; Hartmann et al., 2016).

To date, research has seldom addressed situations in which different lines of actions (in the sense of conflicting evidence) are available. Consider, for example, a typical classroom situation in which two students become angry with one other and are arguing at a time when all the students have been asked to work quietly on their worksheets. From a classroom management perspective, the teacher should intervene immediately to enforce the classroom rule that time on task should be maximized (see Lenske et al., 2016 for evidence on the influence of classroom management on students' learning gains). From the perspective of the development of a healthy classroom climate and peer relationships, the teacher may instead let the class stop working and use this situation to explicitly address productive ways to solve peer conflicts (Jennings and Greenberg, 2009). Here, two educational goals with unique theoretical and empirical backgrounds come into play and lead to divergence in the actions that the teacher could potentially take. Such situations are highly prevalent in educational settings, and teachers face the challenge of weighing the benefits of possible actions against one another and coming to a decision tailored to the specific situation.

Presumably, if teachers make decisions on the basis of the knowledge that is accessible to them, they may not even perceive certain cues in the situation that would have led to another decision (a reflection of insufficient knowledge). In other cases, in which the teacher does have access to knowledge from different fields, the process by which they evaluate strands of evidence which may lead to different decisions (fragmented knowledge) is not well understood.

To address these issues, this article presents the construction and validation of a vignette-based instrument, involving items presenting scenarios in which decisions may vary depending on the theoretical perspectives on which the decision is based. After describing the theoretical background and the process of constructing the instrument, we present two validation studies indicating that convergent and divergent theoretical perspectives lead to systematic differences in decision-making (Study 1) and that knowledge input influences judgments in a manner that indicates deeper evaluation of the cues corresponding to the newly-acquired knowledge (Study 2).

Theoretical background and research aims

There seems to be an increasing demand for teachers and policymakers to orient their respective educational and political decisions more towards evidence rather than relying on other sources such as subjective theories or anecdotes (Davies, 1999; Bromme et al., 2014). Following Stark (2017) and other researchers, we consider *evidence* to broadly consist of both theories and obtained empirical results that are valued by an individual as being of high scientific quality. That means that evidence does not have an independent existence in an objective sense, outside the judgment of individuals who attribute to it the specific property of meeting scientific standards (Bromme et al., 2014).

Research in the domain of evidence-based education often makes reference to the medical profession, where the parallel term *evidence-based medicine* is employed (Sackett et al., 1996). Although the feasibility of transferring theoretical perspectives from the medical to the teaching profession is under debate (Stark, 2017), the basic idea that teachers use evidence in their argumentation for or against specific decisions seems plausible. In their description of evidence-based argumentation, Csanadi et al. (2021) differentiate between *content* and *process* levels. The *content level* relates to knowledge which is used for evidence-based argumentation. On this level, strands of knowledge with variable epistemic status (Fenstermacher, 1994) are brought to bear. In addition to scientific theories and empirical results, subjective theories or case knowledge can also be put to use as sources in the argumentation process (Kiemer and Kollar, 2021). The *process level* itself can be further subdivided into the selection and the use of specific sources (Kiemer and Kollar, 2021). In turn, the use of specific sources consists of further subprocesses, including problem identification, hypothesis generation, and drawing conclusions (Fischer et al., 2014).

Recent research on the process level has provided insight into the ways in which (pre-service) teachers use or do not use evidence. For instance, Hetmanek et al. (2015) have demonstrated that pre-service teachers – despite being provided with the necessary information – do not use scientific evidence in their case analysis. Concerning the content level, recent studies have directly compared types of source to explore their specific role in the argumentation process. For example, Kiemer and Kollar (2021) have demonstrated that scientific theories are used more often than anecdotes or subjective theories in case analysis.

Research gap

To date, there has been a paucity of research concerning the comparison of sources with comparable epistemic status, such as convergent or divergent scientific theories. In such situations, heuristics like 'scientific theories are more trustworthy than subjective theories' provide no value. Instead, the evidence has to be evaluated with respect to the specific situation at hand and different strands must potentially be weighted differently in order to arrive at a decision. To this end, relevant information in the scenario, typically referred to as cues, must be observed and ultimately taken into account in the argumentation process. Furthermore, research on evidence-based argumentation in the domain of education has mainly focused on generic issues in teaching, such as motivation or general instruction. Subject-specific theories are seldomly addressed as sources of evidence.

To address these gaps, we developed an approach using multi-theory vignettes. The basic idea is to present situations that can be perceived differently from different perspectives. By defining two perspectives and their related core principles *a priori* in the process of constructing the vignettes, we can explicitly model participants' decision-making processes and formulate hypotheses concerning their reactions to the situations depicted.

Construction of multi-theory vignettes

Vignettes as a test format are becoming increasingly popular in the field of teacher education (Brovelli et al., 2014). Under this approach, each vignette consists of a scenario that presents an authentic situation from a lesson in school involving specific issues which necessitate the activation of professional knowledge in order to address them, and they are considered to be a suitable tool to assess situational knowledge or the ability of participants to access their knowledge in specific situations. In particular, research in the field of professional vision regularly employs this approach (Santagata and Angelici, 2010; Meschede et al., 2017).

Under our multi-theory vignette (MTV) approach, we constructed a set of vignettes containing cues that would be relevant from two different perspectives: the first falling primarily under the scope of a specific model of teaching games in Physical Education (PE), and the second falling primarily under the scope of self-determination theory (SDT). The core principle from the former perspective is that of *complexity reduction*: most teaching approaches in the domain of PE agree that sporting games need to be reduced in complexity when they are integrated into a school's curriculum (Kolb, 2005). Therefore, the teachers' behaviors depicted in our vignettes can be considered suitable from a PE teaching perspective if they involve a cue indicating some kind of complexity reduction. The core principle from the latter perspective is the *fulfilment of basic psychological needs*. If students' basic psychological needs are fulfilled, this appears to enhance their sense that their actions are self-determined and to increase their intrinsic motivation (Ryan and Deci, 2000). Research indicates that satisfaction of basic psychological needs is associated with self-determined motivation (Chen and Jang, 2010; Goldman et al., 2017; Hu and Zhang, 2017) and positive learning outcomes (Baeten et al., 2013; McEown et al., 2014; Salmi and Thuneberg, 2019). Therefore, teachers' behaviors depicted in our vignettes that address students' psychological needs can be considered suitable from the SDT perspective.

As complexity reduction and need satisfaction are conceptually unrelated and are principles that arise from different theoretical perspectives, we combined both perspectives with their core principles in our vignettes. A *convergent vignette* would depict a pedagogical situation in which the action of the fictitious teacher is either suitable (the core principles are fulfilled) or unsuitable (the core principles are not fulfilled) according to both perspectives. We adopted a labelling scheme in which convergent vignettes depicting suitable teacher behavior were labelled SS because they suggested a **s**uitable teacher action as seen from both perspectives. In contrast, convergent vignettes depicting unsuitable actions were labelled UU, as they suggested an **u**nsuitable teacher action from both perspectives. A *divergent vignette* would depict a teacher action that is suitable from one of the perspectives and unsuitable from the other. These vignettes were labelled U_gS_m if they depicted an action which could be considered **s**uitable or need-supporting from the perspective of **m**otivational psychology or SDT, but an **u**nsuitable action from the perspective of teaching **g**ames; or S_gU_m if they depicted an action which could be regarded as suitable from the perspective of teaching games but unsuitable from the perspective of SDT or motivational psychology.

A total of 10 experts in the field of sports science with a focus on teaching games and 11 experts in motivational psychology were asked to evaluate our categorizations of 26 drafted vignettes as illustrating suitable or unsuitable actions from their expert perspective. To this end, they were informed beforehand of which teacher actions we considered to be suitable or unsuitable in terms of complexity reduction and need satisfaction. The sports science experts were not informed of the SDT interpretation of the vignettes, nor were they asked to rate the vignettes with this perspective in mind, and vice versa. The experts were also asked to name possible alternative actions for the teacher in each vignette. In general, the experts considered the vignettes to be authentic and suitable for our research purposes. However, some disagreement emerged concerning the suitability of the actions described, and experts from both fields suggested alternative actions for the teacher in a number of cases. It became clear that sports science experts with a focus on teaching games weight motivational considerations more heavily than psychology experts weight sports science considerations. After discussing all the results, excluding 10 vignettes, and slightly reformulating some vignettes, we arrived at a final set of 16 vignettes, four of each type (UU, SS, U_gS_m, and S_gU_m).

Example multi-theory vignette

> There is not much activity happening on the field where 24 fifth-graders are playing dodgeball: only a few students are actively taking part by running, dodging the ball, trying to catch it and throwing it at their opponents. One of the less active players, who has already had to leave the active zone of the field, is now outside in the passive zone (from where it is possible to return to the active zone by successfully throwing the ball at an opponent). She is standing close to the teacher and says to him: "This game is sooo boring…," looking expectantly at the teacher.
>
> The teacher replies that it would not be so boring if she, the girl, took part in it more actively. When the first round of the game has finished and the second round is about to begin, the teacher reduces the size of the field.

This vignette was constructed and used in our test as a divergent vignette (S_gU_m). From the perspective of teaching games, the teacher reacts rather appropriately to the lack of activity among his students by reducing the field size (complexity reduction). This lack of activity is evident in the vignette through the descriptions of the many passive players on the field and also the girl's claim of boredom. Although it cannot be assumed that the teacher's response here represents the ideal reaction, it is certainly a possible solution to a lack of activity during a ball game. However, from the SDT perspective, the teacher's reaction to the girl's complaint is inappropriate because he does not address the basic psychological needs of the student in this situation (need satisfaction). His answer makes it clear that he would prefer the girl to eliminate her negative emotions as quickly as possible. Additionally, he gives an unclear instruction by telling the girl that she should take part more actively: it can be assumed that the girl does not know what 'taking part more actively' means. Therefore, the student's psychological needs are not satisfied.

General hypotheses on multi-theory vignette ratings

As described, each vignette contained a problem, a dilemma, or a challenge to which a fictitious teacher's reaction was depicted. Each ended with a description of the teacher's actions, which were generally verbal, but sometimes non-verbal. As part of the instrument, participants were then asked to rate the fictitious teacher's action in relation to the statement 'The teacher's action is suitable', with higher ratings indicating greater perceived suitability. Convergent vignettes (i.e., those in which the actions are either suitable or unsuitable from both perspectives) are rather clear, and we thus expected participants to provide polarized ratings: UU vignettes should receive the lowest rating and SS vignettes the highest rating, indicating high unsuitability and high suitability, respectively. In contrast, we expected ratings for divergent vignettes (i.e., those in which the suitability of the actions varied depending on the perspective adopted) to be close to the middle of the scale, as participants should be undecided. An example train of thought for the participant might be: "This is an appropriate way of dealing with the issue [complexity reduction], but the way he talks to his students does not seem right… [no need satisfaction]." However, our objective was to establish a method of identifying the type of knowledge brought to bear by different participants in providing their ratings by investigating individual differences in the ratings of divergent vignettes. Specifically, if a participant judges the actions depicted in U_gS_m vignettes to be more suitable than those depicted in S_gU_m vignettes, it can be concluded that their knowledge of SDT seems to have been of greater importance in their decision; conversely, if a participant judges the actions depicted in S_gU_m vignettes to be more suitable than those depicted in U_gS_m vignettes, it can be concluded that cues relating to the perspective of PE teaching seem to have been more salient to them. Furthermore, by examining changes in these differences over time, we expected to be able to measure the effects of knowledge-building and application.

The present study

We conducted two studies to test the validity of the MTV instrument described above (Borsboom et al., 2004). In Study 1, we aimed to pilot the instrument with a sample of student teachers and a sample of sports science students. We expected the vignette ratings to exhibit the distribution described above, with SS vignettes receiving the highest ratings, UU vignettes the lowest, and US and SU vignettes receiving intermediate ratings. We also expected that the exact pattern would be dependent on the sample: specifically, we hypothesized that U_gS_m vignettes would be associated with lower suitability judgments than S_gU_m vignettes by sports science students and vice versa for student teachers. In Study 2, we tested the hypothesis that a knowledge intervention providing information on SDT would elicit an increase in the difference between participants' U_gS_m and S_gU_m ratings.

Study 1: Pilot

Methods

Student teachers (Sample 1)

Sample 1 consisted of 153 pre-service teachers (127 female) from a university of specializing in education studies. The mean age was 21.85 years ($SD = 3.06$); 78.5% were in semester 3 of their studies or below, and the remaining 22.5% were in semesters 4 to 13. The vast majority (79.7%) were working towards a Bachelor of Arts in either primary or secondary education; a smaller number (18.3%) were working towards a master's degree in education.

Sports science students (Sample 2)

Sample 2 consisted of 48 sports science students (27 female), with a mean age of 21.10 years (*SD* = 2.15). Most were working towards a Bachelor of Science (87.5%); the remainder were working towards a Master of Science (12.5%). Approximately, 77% were in either their first or their third semester of study; only two (4.2%) had advanced beyond semester five.

Instruments

Participants completed the MTV instrument, in which they were presented with 16 MTVs (four in each condition, sample item in section 2.4) and were asked to rate the statement 'The teacher's reaction is suitable' in relation to each vignette. Ratings were given on a five-point Likert scale ranging from 1 (*I completely disagree*) to 5 (*I completely agree*). The instrument was administered online to both participant groups. Language complexity of vignettes (measured by the LIX-index, Lenhard and Lenhard, 2014-2022) seems comparable for all types of vignettes: $LIX_{UU} = 42.15$ ($SD = 6.35$); $LIX_{US} = 49.49$ ($SD = 3.20$); $LIX_{SU} = 45.94$ ($SD = 8.53$), $LIX_{SS} = 47.60$ ($SD = 6.48$); Kruskal-Wallis-H = 2.54, $df = 3$, $p = 0.47$; Bayes factor for ANOVA with H_1: $LIX_{UU} = LIX_{US} = LIX_{SU} = LIX_{SS}$ compared to unconstraint model H_u indicates moderate evidence for equality assumption ($BF_{iu} = 2.94$). Subsequently, each participant provided a self-assessment of their knowledge of both SDT and the teaching of games in PE on a four-point (sample 1) or five-point (sample 2) Likert scale. To compare these ratings between scales, we transformed individual ratings to scores on a scale ranging from 0 (*no knowledge*) to 1 (*advanced knowledge*).

Analyses

To explore whether participants' ratings followed the hypothesized patterns, mean ratings for each vignette were calculated. Next, a mean score was computed for each set of vignettes in the same condition (i.e., UU, SU, US, or SS); these scores can be interpreted as representing the mean rating for vignettes within each condition or cluster. Next, we conducted Welch's *t*-tests to compare the mean ratings given by participants in samples 1 and 2 for each condition. We were particularly interested in these comparisons for the conditions involving divergent vignettes. Finally, we calculated a Baysian analysis of variance with repeated measures (Gu et al., 2018; Hoijtink et al., 2019): mean scores for conditions represented the within-subject factor with four levels and sample was the between-subject factor with two levels (sample 1 and sample 2) and the following informed hypotheses:

> H1: $\mu_{UU1} = \mu_{US1} = \mu_{SU1} = \mu_{SS1}$ (all means are equal in sample 1).
> H2: $\mu_{UU2} = \mu_{US2} = \mu_{SU2} = \mu_{SS2}$ (all means are equal in sample 2).
> H3: $\mu_{UU1} < \mu_{SU1} < \mu_{US1} < \mu_{SS1}$ (means are ordered with SU being lower than US in sample 1).
> H4: $\mu_{UU1} < \mu_{US1} < \mu_{SU1} < \mu_{SS1}$ (means are ordered with US being lower than SU in sample 1).
> H5: $\mu_{UU2} < \mu_{SU2} < \mu_{US2} < \mu_{SS2}$ (means are ordered with SU being lower than US in sample 2).
> H6: $\mu_{UU2} < \mu_{US2} < \mu_{SU2} < \mu_{SS2}$ (means are ordered with US being lower than SU in sample 2).
> H7: $\mu_{UU1} - \mu_{UU2} = \mu_{US1} - \mu_{US2} = \mu_{SU1} - \mu_{SU2} = \mu_{SS1} - \mu_{SS2}$ (mean differences for clusters are equal across samples indicating no interaction effect).

For each hypothesis, we calculated Bayes factors compared to the unconstrained hypotheses using the R-package bain (Hoijtink et al., 2019).

Results

Our research objective in Study 1 was to collect evidence on the validity of our proposed instrument by comparing participants' ratings of the suitability of the behaviors depicted in the convergent and divergent vignettes to our hypotheses regarding the expected pattern of ratings.

Preliminary analysis: Comparison of groups

As access to relevant knowledge was expected to influence evidence-based argumentation *via* its influence on the ability to identify relevant cues, participants were asked to rate their knowledge concerning the content of the vignettes; these ratings are summarized in Table 1. With respect to knowledge of SDT, both samples (i.e., both student teachers and sports science students) gave rather low self-reports (with mean ratings being 0.11 and 0.15, respectively); there was no significant difference between the groups, $t(82) = 0.94$, $p = 0.35$, Cohen's $d = 0.15$. However, as expected, sports science students reported having significantly more knowledge of teaching games ($M = 0.40$, $SD = 0.25$) than did student teachers ($M = 0.06$, $SD = 0.18$), $t(64) = 8.77$, $p < 0.001$, Cohen's $d = 1.69$.

Vignette ratings

Average ratings for each vignette are presented in Table 2. In line with the hypotheses, convergent vignettes of the UU type received the lowest ratings (sample 1: $M = 2.39$, $SD = 0.70$; sample 2: $M = 2.42$, $SD = 0.67$). In other words, both groups were in agreement on their judgments of teacher behaviors which we had constructed to represent unsuitable actions from both perspectives. Participants from each group rated the individual UU vignettes (UU1–UU4) slightly differently, but the groups were approximately in agreement on the overall ordering, with vignette UU3 receiving the lowest overall rating and vignettes UU1 and UU2 receiving the highest ratings within this condition. A similar pattern was observed for convergent vignettes of the SS type,

TABLE 1 Self-reported knowledge of self-determination theory (SDT) and teaching games in physical education.

Type of knowledge	Student teachers			Sports science students			Welch's *t*-test and effect size			
	N	*M*	*SD*	*N*	*M*	*SD*	*t*	d.f.	*p*	Cohen's *d*
SDT	150	0.11	0.21	48	0.15	0.21	0.94	82	0.35	0.15
Teaching games	152	0.06	0.18	48	0.40	0.25	8.77	64	<0.001	1.69

TABLE 2 Vignette ratings: descriptive statistics and group comparisons.

Vignettes*	Student teachers (*N* = 153)		Sports science students (*N* = 48)		Welch's *t*-test and effect size			
	M	*SD*	*M*	*SD*	*t*	d.f.	*p*	*d*
UU1	2.95	1.10	2.85	1.05				
UU2	2.91	1.19	3.13	1.05				
UU3	1.56	0.77	1.77	1.07				
UU4	2.12	1.15	1.91	1.01				
	2.39	0.70	2.42	0.67	0.27	81	0.79	0.04
US1	3.71	1.12	3.57	1.15				
US2	3.31	1.35	3.38	1.34				
US3	1.68	0.94	1.42	0.74				
US4	2.71	1.17	2.67	1.02				
	2.84	0.60	2.73	0.51	1.25	91	0.22	0.19
SU1	4.11	1.07	3.93	1.16				
SU2	2.60	1.31	3.30	1.23				
SU3	2.52	1.25	2.65	1.19				
SU4	3.62	1.13	3.56	1.33				
	3.23	0.73	3.35	0.83	0.90	71	0.37	0.16
SS1	4.01	1.15	3.64	1.26				
SS2	4.07	1.02	4.09	1.09				
SS3	3.81	1.15	3.75	1.26				
SS4	3.14	1.27	3.14	1.22				
	3.77	0.71	3.61	0.79	1.25	72	0.21	0.22

*U, unsuitable; S, suitable, first letter indicating sports science perspective, second letter indicating psychological perspective.

which received the highest overall ratings (sample 1: $M = 3.77$, $SD = 0.71$; sample 2: $M = 3.61$, $SD = 0.79$).

The conditions containing divergent vignettes (US and SU) received intermediate ratings from participants in both groups, a result that was also in line with the hypotheses. Additionally, participants in both groups judged the actions in US vignettes ($M_1 = 2.84$; $M_2 = 2.73$) to be slightly less suitable than those in SU vignettes ($M_1 = 3.23$; $M_2 = 3.35$). Within these conditions, the rank order of the suitability of individual vignettes was constant across both groups, although the mean ratings varied.

Repeated measures ANOVA indicated a significant main effect for the within-subject factor "vignette condition" with $F(3) = 119.334$, $p < 0.001$, partial $\eta^2 = 0.377$. Post-hoc comparisons with Bonferroni correction revealed significant differences ($p < 0.001$) between each condition (UU, US, SU, and SS).

Comparison of ratings by student teachers and sports science students

The results of an independent-samples Welch's *t*-test indicated that there was no significant difference between the two groups in their ratings of UU vignettes. The corresponding effect size (Cohen's $d = 0.04$) indicated that the difference was below the threshold to be considered even a small effect. Similarly, both groups gave comparable judgments in response to the SS vignettes, representing items in which the teacher action was intended to represent a suitable response from both perspectives. There was no statistically significant difference between the ratings given by each group on this condition, and the (statistically insignificant) standardized mean difference (Cohen's $d = 0.22$) was just above the threshold of what is considered to be a small effect.

Concerning the divergent vignette conditions, once again no significant effect of group was observed. A comparison on the descriptive level of the within-group difference between ratings of the SU and US vignettes across groups indicated that there was a larger difference in the case of sports science students, who self-reported having greater knowledge of PE teaching (ΔSU–US = 0.61, $SD = 0.87$), compared to pre-service teachers (ΔSU–US = 0.38, $SD = 0.88$). However, this difference in differences was not significant, $t(79) = 1.62$, $p = 0.11$, Cohen's $d = 0.27$.

In the repeated measures ANOVA results from the paired Welch-tests could be replicated by a non-significant main effect for sample [$F(1) = 0.158$, $p = 0.692$]. Further, a non-significant interaction between condition and sample [$F(3) = 1.518$, $p = 0.209$] leads to the assumption that judgments did not depend on the sample. The results from the frequentist approach were supported by bayesian evaluation of informed hypotheses: Bayes factors indicated strong evidence for H4 ($BF_{iu} = 20.98$, means are ordered with US being lower than SU in sample 1), H6 ($BF_{iu} = 21.52$, means are ordered with US being lower than SU in sample 2), and H7 ($BF_{iu} = 13.49$, mean differences for clusters are equal across samples indicating no interaction effect).

Overall, our results indicated that participants were able to identify relevant cues in the vignettes in judging the suitability of specific teacher actions, and this led to a pattern of ratings that conformed to the hypotheses. However, differences between the two groups in terms of the mean ratings they gave were observed only on the level of individual vignettes, with no differences observed in the groups' average ratings over any of the aggregated conditions (UU, US, SU, or SS). There was a tendency in the case of the divergent vignette conditions towards a difference between the groups, in the hypothesized direction, but this did not reach statistical significance. This lack of systematic differences between the groups may be attributable to the fact that the participants had not had enough opportunities to build a sufficient knowledge base in their respective fields. To address this issue, we conducted Study 2, in which a specific knowledge intervention was implemented.

Study 2: Intervention

Methods

This study was carried out in the course of a unit of teaching taken by students as part of their degrees in education studies. We developed a short intervention and tested whether this changes how participants perceive the suitability of teacher actions in our MTVs.

Design and sample

To investigate whether knowledge input would change participants' judgments in relation to our MTVs, we employed a pre–post intervention design. At the beginning of the unit, a pre-test including a similar MTV instrument to the one used in Study 1 was administered to participants. After participating in the knowledge intervention, they also completed an identical post-test. The entire procedure, consisting of the pre-test, knowledge intervention, and post-test, took place during the regular 90-min session for delivery of the unit in question.

The 46 participants (72% female) were recruited from a single university specializing in education studies and had mean age of 24.59 years (*SD* = 1.88). All participants had already obtained a bachelor's degree in the field of education, either at the primary (65%) or the secondary (35%) level, and at the time of the study, they were working towards a master's degree in education. Most (approximately 85%) were in the first year of their master's studies; 36 had already completed an obligatory semester of practical teacher training in a school, and 10 had yet to do so. Of the 48 potential participants who initiated participation by beginning the pre-test, 46 (96%) completed the entire pre-test. Following the knowledge intervention, 42 participants began the post-test. Ultimately, full data (i.e., a linked pre-test and post-test) were available for 38 participants.

Intervention

The intervention administered in Study 2 aimed to enhance student teachers' knowledge of SDT, and specifically their understanding of the core principle of basic psychological needs and the ways in which teachers might foster need satisfaction in the classroom setting. The intervention was embedded in a seminar on learning and motivation theories, in the form of a unit which lasted approximately 60 min. The unit began with an overview of SDT, including presentations of cognitive evaluation theory and organismic integration theory as sub-theories (Ryan and Deci, 2000). Subsequently, the focus was on providing an explanation of competence, autonomy, and social relatedness as basic psychological needs which foster self-determined forms of motivation. The key overall message, therefore, was that teachers design the motivational climate of their classrooms in such a way as to fulfil certain basic needs to variable extents. Finally, based on this theoretical perspective, various possible actions specifically linked to the satisfaction of basic needs were presented. Participants were free to ask questions during the unit and to make comments based on their own ideas or understanding. Questions and comments were handled discursively; nevertheless, this intervention overall can be considered to have been rather directive. The intervention was administered as an online course *via* the platform Zoom.

Instrument

The pre-test and post-test included an online questionnaire with several parts. In addition to collecting several demographic variables (e.g., gender and age), we asked participants to self-rate their knowledge of SDT on a scale from 1 (*no knowledge*) to 6 (*advanced knowledge*). Additionally, we constructed a knowledge test on the topic of SDT, consisting of nine multiple-choice items whose content was directly linked to the content of the knowledge intervention. Participants' responses to each item were coded as 0 = *incorrect* or 1 = *correct*, and overall test scores were calculated by summing these values, resulting in a range of possible test scores from 0 to 9.

The central component of both tests was the MTV instrument. Due to time restrictions, we divided the 16 vignettes into two comparable subsets, each containing eight items consisting of two from each condition (UU, SU, US, and SS). Participants were randomly assigned (*via* random assignment to breakout sessions in Zoom) to complete one of the subsets of items for the pre-test and each completed the same subset again in the post-test phase. In each case, participants rated the teacher's actions described in each vignette on a six-point Likert scale ranging from 1 (*The teacher's action is unsuitable*) to 6 (*The teacher's action is suitable*). The purpose of using a six-point scale, rather than a five-point scale as in Study 1, was to avoid the possibility of participants selecting the midpoint of the scale, thus encouraging them to choose at least a specific direction for their evaluation. We anticipated that this would allow us to observe any changes in their decision-making more clearly.

Analyses

To test our hypothesis that the knowledge intervention would lead to changes in participants' MTV ratings, we first examined whether knowledge gains had occurred using a paired-samples *t*-test to compare pre- and post-test scores on self-assessed knowledge and knowledge test scores. Subsequently, we computed mean scores for each condition of the vignette instrument (UU, SU, US, and SS) and carried out paired-samples *t*-tests comparing participants' pre- and post-test judgments for each condition. To quantify the effects, we also computed Cohen's *d* as a measure of effect size.

Results

Our research objective for Study 2 was to investigate whether an intervention involving knowledge input would alter participants' judgments of the suitability of teacher behaviors in the MTVs.

TABLE 3 Vignette ratings in the intervention study and results of comparisons between pre- and post-test ratings.

Vignette*		Pre-test		Post-test		Paired *t*-test			
	N	*M*	*SD*	*M*	*SD*	*t*	d.f.	$p_{\text{one-sided}}$	$d_{\text{RM. pooled}}$
UUA1	19	1.84	0.83	1.74	0.99				
UUA2	19	2.26	1.05	2.16	1.07				
UUB1	19	2.84	1.26	2.42	1.17				
UUB2	19	2.21	1.27	1.79	1.18				
UU	38	2.29	0.89	2.03	0.96	3.02	37	0.01	−0.49
SUA1	19	3.11	1.49	3.05	1.31				
SUA2	19	2.42	1.02	2.58	1.07				
SUB1	20	2.90	1.29	3.05	1.54				
SUB2	19	4.32	1.16	4.11	1.29				
SU	39	3.22	0.98	3.18	1.13	0.25	38	0.40	−0.05
USA1	19	3.95	1.31	4.47	1.35				
USA2	19	3.47	1.26	3.74	1.56				
USB1	19	4.16	1.01	3.84	1.46				
USB2	20	3.40	1.27	3.85	1.31				
US	39	3.76	0.76	3.99	0.98	−1.56	38	0.06	0.24
SSA1	19	5.05	1.03	4.89	1.15				
SSA2	19	4.79	1.03	5.05	0.71				
SSB1	19	4.58	1.35	4.63	1.38				
SSB2	19	4.47	1.65	4.84	1.26				
SS	38	4.72	1.00	4.86	0.98	−0.89	37	0.19	0.13

*U, unsuitable; S, suitable, first letter indicating sports science perspective, second letter indicating psychological perspective, A and B indicate to which subset the respective vignettes belonged.

Effect of the intervention on self-determination theory knowledge

To test for an intervention effect independently of the MTV instrument, we analyzed participants' knowledge gains between the pre- and post-tests on the basis of their self-assessments and the more objective knowledge test. The results indicated that participants made sizeable knowledge gains: average self-assessed knowledge scores increased significantly from 2.31 ($SD = 0.80$) to 3.87 ($SD = 0.89$), $t(38) = 10.71$, $p < 0.001$, Cohen's $d = 1.60$; and average scores on the SDT knowledge test increased from 5.26 ($SD = 1.41$) to 6.46 ($SD = 1.07$), $t(38) = 5.45$, $p < 0.001$, Cohen's $d = 0.80$.

Effect of the intervention on MTV ratings

The intervention aimed to increase participants' knowledge of SDT, and more specifically their understanding of the core principle of the satisfaction of basic needs. Therefore, we expected that, following the intervention, participants would be better able to identify the SDT-related cues included in the MTVs, and thus that they may judge the actions depicted in UU and S_gU_m items to be less suitable, whereas they may judge those depicted in U_gS_m and SS items to be more suitable. Descriptive statistics (Table 3) indicated that participants' ratings of items in all conditions shifted in the expected directions; Figure 1 further illustrates the changes in average ratings and dispersion values between the pre- and post-test.

However, statistical comparisons of the mean differences indicated that these did not reach the level of statistical significance ($p < 0.05$), with the exception of condition UU, $t(37) = 2.30$, $p < 0.05$, Cohen's $d = -0.49$, $BF_{iu} = 1.98$ [weak evidence for H_i: $\mu_{Pre} > \mu_{Post}$]. According to the standardized mean differences (Cohen's *d*), the changes in rating for the US and SS conditions were rather small ($d = 0.24$ and $d = 0.13$, respectively),

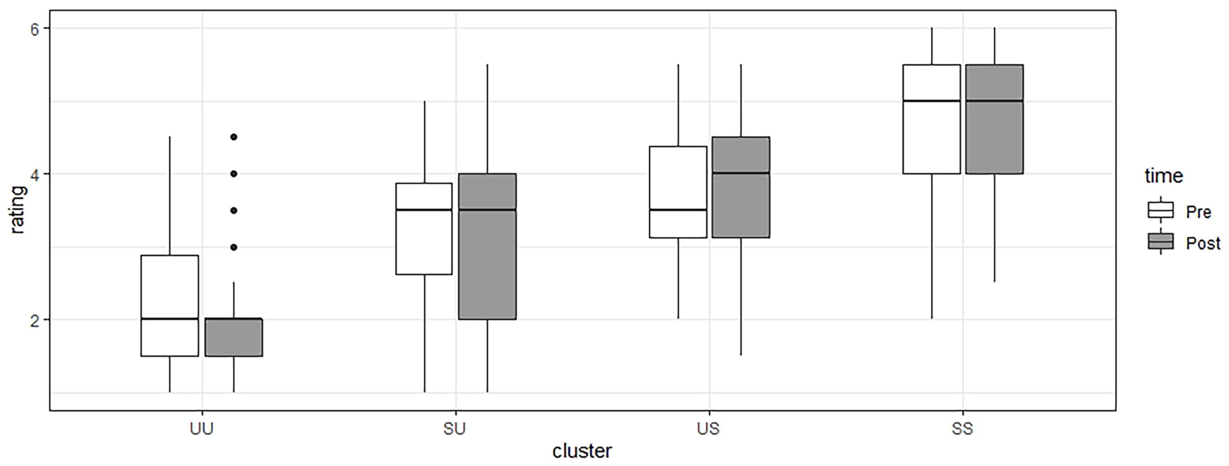


FIGURE 1
Comparisons of mean absolute ratings and mean difference in ratings for each condition between pre- and post-test.

while the change in rating for the SU condition was close to zero ($d = -0.05$). Figure 2 presents an illustration of Bayes factors for H1: $\mu_{Pre} > \mu_{Post}$, H2: $\mu_{Pre} = \mu_{Post}$, and H3: $\mu_{Pre} > \mu_{Post}$ against the unconstraint model.

Discussion

In this study, we tested MTVs as a tool for the presentation, in a measurement instrument, of authentic situations including content cues linked to two different theoretical perspectives. In particular, we constructed vignettes that could be evaluated from the perspective of teaching games in PE and from the perspective of self-determined motivation, which would involve application of the core principles of complexity reduction and need satisfaction, respectively. From each of these perspectives, the specific teacher action depicted in each vignette could be considered to be either suitable or unsuitable, producing four types of vignettes, two convergent and two divergent. In two studies, we demonstrated that participants' judgments of the suitability of the teacher's behaviors varied as expected according to vignette type. Furthermore, a brief knowledge intervention elicited change in participants' judgments between a pre-test and a post-test.

Multi-theory vignettes as a research tool

Evidence-based education as a field of research has gained in importance, with an increasing focus on the process level: that is, on evidence-based argumentation (Csanadi et al., 2021). In this domain, existing research indicates that (pre-service) teachers vary in their approaches to the selection and use of different sources of evidence (Kiemer and Kollar, 2021). Those sources can be considered to vary in their epistemic status (Fenstermacher, 1994) and utility value (Kiemer and Kollar, 2021). However, there is little existing research addressing the selection and use of competing strands of evidence that are of comparable epistemic status.

From the perspective of evidence-based argumentation, participants presented with our MTVs are confronted with a rather weakly-defined scenario: although some contextual information is provided, much other information has to be inferred. Nevertheless, the available information may lead to the generation of different hypotheses (step 3 in the process model proposed by Fischer et al., 2014) and, in the case of divergent MTVs, possible re-evaluation of one's thoughts (step 7). This specific step, in which two closely comparable hypotheses must be evaluated and weighed up to arrive at a decision or the solution to a problem, allows for a deeper exploration of the extent to which teachers make use of their knowledge of specific theories and empirical evidence. Here, a discrepancy between formal logic and participants' response becomes obvious: Formal logic would forecast that vignettes which contain at least one unsuitable action should be rated equally low – irrespective of other likely suitable actions. Instead, participants seem to apply a compensatory approach where suitable actions may compensate for unsuitable ones. In the present study, we further explored the influence of a specific knowledge intervention: specifically, we provided participants with information on a particular scientific theory as a source of evidence. However, future research may use other sources of evidence and explore the extent to which they can influence participants' judgments.

According to the perspective of professional vision as a knowledge-based ability (van Es and Sherin, 2002), participants presented with MTVs notice specific cues in the scenario, which then lead to a reasoning process. Both noticing and reasoning are considered to be knowledge-based processes, which means – drawing on process models of selective attention – that cues in the scenarios can only be noticed if the corresponding knowledge is

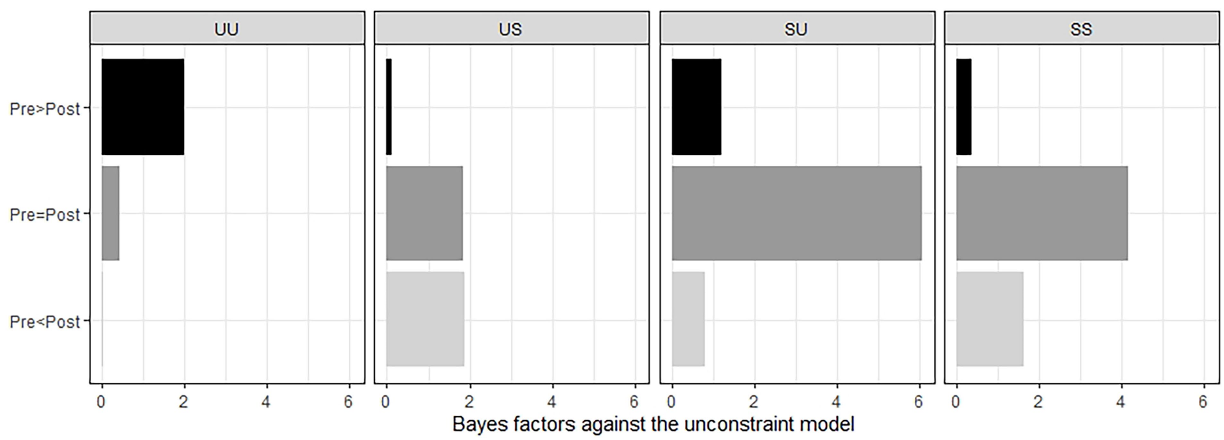


FIGURE 2
Bayes factors for different hypotheses compared to unconstraint model depending on vignette clusters.

represented in the cognitive system (see also Loibl et al., 2020). In Study 1, sports science students had higher levels of self-reported knowledge in the domain of teaching games and tended to orient their decisions to the suitability of the teacher's actions from that perspective to a greater extent. Furthermore, in Study 2, the intervention enhancing participants' SDT knowledge elicited an increase in their ratings of the suitability of the teacher's actions in divergent vignettes depicting behaviors which would be considered suitable from an SDT perspective, and a decrease in their ratings of the suitability of actions that would be considered unsuitable from that perspective. However, it remains unclear whether these results can be attributed to changes in their noticing or their reasoning processes. Nevertheless, MTVs may represent a potential tool for further differentiation between these processes, an issue which is seldom addressed in research in the domain of professional vision (Gold and Holodynski, 2017; Meschede et al., 2017).

Limitations

Despite the promising initial results, several limitations to the present study warrant further consideration. First, the participant samples recruited for the pilot (Study 1) were limited in scope. Both groups consisted of university students who were rather inexperienced in their fields of study. Therefore, the ratings they provided may have had a tendency to represent 'common sense' ratings rather than the results of systematic argumentation. Nevertheless, the mean ratings followed the hypothesized pattern. Additionally, we did not explicitly control for content, tone, and sentiments between and within vignette clusters. Further studies are needed to explore the extent to which specific knowledge may contribute to participants' judgments and to explore possible context factors.

Second, the MTVs that we constructed encompassed only two theoretical perspectives, namely teaching games in PE and SDT. Therefore, the results should be replicated across other theoretical perspectives in different content domains. However, we consider the instrument presented here to be a prototype, with reference to which many other MTVs can be generated in accordance with specific research questions. Additionally, the integration of multiple theoretical perspectives allows for research that crosses subject matter domains.

Third, in our analyses we theoretically assume that the judgments with a specific combination are uni-dimensional. Due to different knowledge domains (teaching ball games and self-determination theory) as well as different aspects within each vignette, this assumption could be questioned. However and as first evidence, dimensionality analysis within the structural equation modelling framework revealed that a four-dimensional model (UU, SU, US, and SS) fits better to our data of sample 1 than a one-dimensional model ($\Delta\chi2 = 28.051$, $\Delta df = 6$, $p < 0.001$). Nevertheless, testing the uni-dimensional assumption within each vignette and more complex latent structures across vignettes warrant further attention. Additionally, we calculated mean ratings within each vignette cluster. With this approach, we possibly reduced heterogeneity on the level of vignette ratings. Future studies might inspect sources for differences on the vignette level and may relate them to knowledge differences, for instance.

Fourth, the results of the knowledge intervention study are limited in scope due to a deficiency in the design: specifically, we did not include a knowledge intervention focusing on knowledge about teaching games in PE (with the core principle of complexity reduction). Under the view we have presented here, we would hypothesize that gaining knowledge in this area would influence participants' judgments in the opposite direction. Furthermore, it would be of interest to investigate the outcome if participants are exposed to both interventions. This approach would allow for a deeper exploration of the processes involved in weighing up multiple hypotheses.

Contributions

Despite the limitations mentioned above, this study demonstrates that it is possible to construct vignettes that manipulate participants' judgments of the suitability of the actions of a teacher which can be considered under a specific theoretical perspective. Although this appears to be a rather research-oriented endeavor, the situations presented in the vignettes are of substantial importance to teachers' everyday experiences in the classroom, where it is very common for them to be confronted with situations in which there is no clearly correct or incorrect response, but rather competing solutions with comparable value. It therefore seems that it would be valuable to obtain further insight into the argumentation processes involved in such situations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

HL-B wrote the manuscript, project management, and supervision. CB was responsible for vignette development and data collection. TD supervision and editing drafts of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work originates from the Research Training Group "Diagnostic Competences of Teachers", which is funded by the Ministry of Science, Research, and the Arts Baden-Württemberg, Germany. The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and the Arts and the University of Education Heidelberg in the funding programme Open Access Publishing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.996029/full#supplementary-material

References

Baeten, M., Dochy, F., and Struyven, K. (2013). The effects of different learning environments on students' motivation for learning and their achievement. *Br. J. Educ. Psychol.* 83, 484–501. doi: 10.1111/j.2044-8279.2012.02076.x

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bromme, R., Prenzel, M., and Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Z. Erzieh.* 17, 3–54. doi: 10.1007/s11618-014-0514-5

Brovelli, D., Bölsterli, K., Rehm, M., and Wilhelm, M. (2014). Using vignette testing to measure student science teachers' professional competencies. *Am.J. Educ. Res.* 2, 555–558. doi: 10.12691/education-2-7-20

Brown, C., and Rogers, S. (2015). Knowledge creation as an approach to facilitating evidence informed practice: examining ways to measure the success of using this method with early years practitioners in Camden (London). *J. Educ. Chang.* 16, 79–99. doi: 10.1007/s10833-014-9238-9

Chen, K.-C., and Jang, S.-J. (2010). Motivation in online learning: testing a model of self-determination theory. *Comput. Hum. Behav.* 26, 741–752. doi: 10.1016/j.chb.2010.01.011

Csanadi, A., Kollar, I., and Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Davies, P. (1999). What is evidence-based education? *Br. J. Educ. Stud.* 47, 108–121. doi: 10.1111/1467-8527.00106

Fenstermacher, G. D. (1994). Chapter 1: the knower and the known: the nature of knowledge in research on teaching. *Rev. Res. Educ.* 20, 3–56. doi: 10.3102/0091732X020001003

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i3.96

Floden, R. E., and Buchmann, M. (1993). Between routines and anarchy: preparing teachers for uncertainty. *Oxf. Rev. Educ.* 19, 373–382. doi: 10.1080/0305498930190308

Gold, B., and Holodynski, M. (2017). Using digital video to measure the professional vision of elementary classroom management: test validation and methodological challenges. *Comput. Educ.* 107, 13–30. doi: 10.1016/j.compedu.2016.12.012

Goldman, Z. W., Goodboy, A. K., and Weber, K. (2017). College students' psychological needs and intrinsic motivation to learn: an examination of self-determination theory. *Commun. Q.* 65, 167–191. doi: 10.1080/01463373.2016.1215338

Gu, X., Mulder, J., and Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: a general method for testing informative hypotheses. *Br. J. Math. Stat. Psychol.* 71, 229–261. doi: 10.1111/bmsp.12110

Hartmann, U., Decristan, J., and Klieme, E. (2016). Unterricht als Feld evidenzbasierter Bildungspraxis? *Z. Erzieh.* 19, 179–199. doi: 10.1007/s11618-016-0712-4

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M. R., Gräsel, C., et al. (2015). Wozu nutzen Lehrkräfte welche Ressourcen? Eine Interviewstudie zur Schnittstelle zwischen bildungswissenschaftlicher Forschung und professionellem Handeln im Bildungsbereich. *Unterrichtswissenschaft* 43, 194–210.

Hoijtink, H., Gu, X., and Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *Br. J. Math. Stat. Psychol.* 72, 219–243. doi: 10.1111/bmsp.12145

Hoijtink, H., Mulder, J., van Lissa, C., and Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychol. Methods* 24, 539–556. doi: 10.1037/met0000201

Hu, P., and Zhang, J. (2017). A pathway to learner autonomy: a self-determination theory perspective. *Asia Pac. Educ. Rev.* 18, 147–157. doi: 10.1007/s12564-016-9468-z

Jennings, P. A., and Greenberg, M. T. (2009). The prosocial classroom: teacher social and emotional competence in relation to student and classroom outcomes. *Rev. Educ. Res.* 79, 491–525. doi: 10.3102/0034654308325693

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift Für Pädagogische Psychologie* 35, 127–141. doi: 10.1024/1010-0652/a000302

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Kolb, M. (2005). "Sportspiel aus sportpädagogischer Sicht" in *Beiträge zur Lehre und Forschung im Sport: Band 147. Handbuch Sportspiel.* eds. A. Hohmann, M. Kolb and K. Roth (Schorndorf: Hofmann), 65–83.

Lenhard, W., and Lenhard, A. (2014-2022). Berechnung des Lesbarkeitsindex LIX nach Björnson. (Dettelbach: Psychometrica) Available at: http://www.psychometrica.de/lix.hml.

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E., Liepertz, S., et al. (2016). Die Bedeutung der pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüleri/innen im Physikunterricht. *Z. Erzieh.* 19, 211–233. doi: 10.1007/s11618-015-0659-x

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgments by cognitive Modeling (DiaCoM). *Teach. Teach. Educ.* 91:103059. doi: 10.1016/j.tate.2020.103059

McEown, M. S., Noels, K. A., and Saumure, K. D. (2014). Students' self-determined and integrative orientations and teachers' motivational support in a Japanese as a foreign language context. *System* 45, 227–241. doi: 10.1016/j.system.2014.06.001

Meschede, N., Fiebranz, A., Möller, K., and Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: on its relation and differences between pre-service and in-service teachers. *Teach. Teach. Educ.* 66, 158–170. doi: 10.1016/j.tate.2017.04.010

Ryan, R. M., and Deci, E. L. (2000). Self-determination and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ* 312, 71–72. doi: 10.1136/bmj.312.7023.71

Salmi, H., and Thuneberg, H. (2019). The role of self-determination in informal and formal science learning contexts. *Learn. Environ. Res.* 22, 43–63. doi: 10.1007/s10984-018-9266-0

Santagata, R., and Angelici, G. (2010). Studying the impact of the lesson analysis framework on Preservice teachers' abilities to reflect on videos of classroom teaching. *J. Teach. Educ.* 61, 339–349. doi: 10.1177/0022487110369555

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer Praxis. *Zeitschrift Für Pädagogische Psychologie* 31, 99–110. doi: 10.1024/1010-0652/a000201

van Es, E., and Sherin, M. G. (2002). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *J. Technol. Teach. Educ.* 10, 571–596.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 07 December 2022
DOI 10.3389/feduc.2022.977788

OPEN ACCESS

EDITED BY
Martin Greisel,
University of Augsburg,
Germany

REVIEWED BY
Mario Mäeots,
University of Tartu,
Estonia
Tom Rosman,
Leibniz-Institute
for Psychology (ZPID), Germany

*CORRESPONDENCE
Katharina Engelmann
katharina.engelmann@uni-hildesheim.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 24 June 2022
ACCEPTED 04 November 2022
PUBLISHED 07 December 2022

CITATION
Engelmann K, Hetmanek A,
Neuhaus BJ and Fischer F (2022) Testing an
intervention of different learning activities
to support students' critical appraisal of
scientific literature.
Front. Educ. 7:977788.
doi: 10.3389/feduc.2022.977788

Testing an intervention of different learning activities to support students' critical appraisal of scientific literature

Katharina Engelmann[1]*, Andreas Hetmanek[2], Birgit J. Neuhaus[3] and Frank Fischer[4]

[1]Institute of Education, Universität Hildesheim, Hildesheim, Germany, [2]TUM School of Social Sciences and Technology, Technische Universität München, München, Germany, [3]Biology Education, Faculty of Biology, Ludwig-Maximilians-Universität München, München, Germany, [4]Department of Psychology, Ludwig-Maximilians-Universität München, München, Germany

In recent years, the call for an evidence-based practice has become more prevalent for educational professionals. However, educational professionals are rarely prepared for evidence-based practice; for example, teachers are not prepared to use and, thus, rarely do use scientific evidence in planning lessons. The knowledge and skills in appraising scientific literature, the basis of evidence-based practice, needs to be trained as early in professional education as possible. An effective training might start in university education of future educational professionals, engaging them in learning activities that foster their understanding of criteria that are used in appraising scientific literature and the skill to do so. However, we know little about the effect of different learning activities such as constructive or interactive learning in this context. Thus, this study investigated the influence of constructive versus interactive learning activities in the context of an intervention facilitating knowledge and skills in appraising scientific literature. This experimental study used a pre-posttest between-subject design with 105 participants. The students learned to evaluate scientific literature in an online learning environment. The results show that the inclusion of interactive versus constructive learning activities did not explain students' learning in the intervention. The results implicate that the learning activities might not play a major role with learning contents such as evidence-based practice. However, the gain in skills and knowledge from pre- to posttest shows promising achievements in preparing future educational professionals in their evidence-based practice.

KEYWORDS

evidence-based education, ICAP, appraisal of scientific literature, evidence evaluation, higher education

Introduction

Reasoning with scientific evidence to solve practical problems is one of the core competences in a knowledge society (Fischer et al., 2014). Specifically, professionals are expected to understand the development of knowledge in their field and to incorporate new knowledge into their practice after a careful evaluation of its origin. The so-called evidence-based practice is already established in medicine (Sackett et al., 1996) and is considered one of its most important milestones (Dickersin et al., 2007). Decisions in medical care of individual or public health decisions are commonly expected to be made with "conscientious, explicit and judicious use of current best evidence" (Sackett, 1997, p. 3).

For the past 20 years, the call for education to follow disciplines such as medicine and to place more importance on scientific evidence in practical decisions has been growing (Slavin, 2008; Cook et al., 2012; Bromme et al., 2014; Brown and Zhang, 2016; Cain, 2016; Stark, 2017; Thomm et al., 2021). Educational professionals, especially teachers, are increasingly expected to identify relevant research, systematically evaluate their findings and implement evidence-based practices in classrooms (Detrich and Lewis, 2013).

One central aspect of evidence-based practice is the critical appraisal of the validity and applicability of the evidence (Sackett, 1997). Consequently, there is an effort to facilitate appraisal of scientific literature in, for example, instructional interventions. So far, studies investigating such interventions are mainly conducted in the field of medicine (e.g., Bradley et al., 2005; Kulier et al., 2012; Reviriego et al., 2014; Molléri et al., 2018) while there is little evidence for fostering appraisal of scientific literature in educational professionals. Moreover, studies investigating one group of educational professionals, teachers, found that they rarely use scientific evidence in professional decisions (Hetmanek et al., 2015a) but rather refer to anecdotal evidence (Menz et al., 2021). Similarly, pre-service teachers rather choose anecdotal evidence than scientific evidence as information source when giving advice for teaching (Kiemer and Kollar, 2021).

Thus, in order to meet the call for more evidence-based practice in education (e.g., Slavin, 2008; Cook et al., 2012; Bromme et al., 2014; Brown and Zhang, 2016; Cain, 2016; Stark, 2017; Thomm et al., 2021), future professionals in education need to be better prepared to use scientific evidence. This paper presents an approach to target one central aspect of using scientific evidence, the critical appraisal of scientific literature.

Critical appraisal of scientific literature

Based on the definition of evidence-based medicine as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996, p. 71), we conceptualize evidence-based practice, independent from a specific discipline, as the *conscientious, explicit, and judicious use of current best evidence in making decisions about one's field of expertise*. Thus, educators' evidence-based practice will include the process of conscientiously, explicitly, and judiciously using the current best evidence in educational research in deciding, for example, on a teaching method in planning a lesson. One central task in working as a teacher is planning lessons. While experienced teachers do not need to plan every lesson from scratch, new learning goals or settings lead to teachers facing the decision on how to design a new lesson. In an evidence-based approach, at least some of these decisions would be made based on educational research that could provide a helpful insight "to identify the practices most likely to bring about positive student outcomes" (Cook et al., 2012, p. 498). In order to make evidence-based decisions in lesson planning, the educational professional starts by carefully analyzing the potential influencing factors, such as the learning goal of the lesson and characteristics of their pupils. The next steps include the search for potential evidence, appraising and selecting relevant, high quality evidence, and finally critically appraising the evidence and developing a lesson plan based on the insights gained from the evidence (see Trempler et al., 2015). While most educational professionals, for example teachers, are not able to spend the time needed to conduct a thorough search and evaluation of existing educational evidence in planning all educational interventions, these practices described above need to be integrated more often into the everyday practice of educational professionals to meet the call for more evidence-based practice in education.

A central aspect in this process is the critical appraisal of scientific literature (Sackett, 1997). In critically appraising scientific literature, the validity and usefulness of the evidence is evaluated (Sackett, 1997). For educational research, critically appraising the validity and usefulness of scientific literature can be adapted to appraising the quality of the research (validity) and the relevance (usefulness) of the research for a given problem (Hetmanek, 2014). Critically appraising the validity and usefulness of scientific literature requires (a) knowledge in criteria that are used in the appraisal as well as the (b) skill to correctly appraise the evidence. Research on sourcing and information integration from multiple resources includes further information such as the author of the text or the source of the evidence (e.g., Bråten et al., 2011; Thomm and Bromme, 2016). The evaluated information can be differentiated into first- and second-hand evaluation; in first-hand evaluation a claim is evaluated directly, while the second-hand evaluation targets the source, for example whether a claim or evidence is authored by a trustworthy expert (Bromme et al., 2010). Since first-hand evaluation requires prior domain-specific knowledge and skills, second-hand evaluation often is the necessary approach for laypeople (Bromme et al., 2010). While future educational professionals could also benefit from instruction in second-hand evaluation, because second-hand information such as trustworthiness of the source might also play a relevant role in educational professionals' evidence-based practice, this paper will focus on first-hand evaluation. Educational professionals possess knowledge and skills related to education, learning, and teaching; thus, they are not laypersons with regard to content knowledge and skills in the field of education. Furthermore they are, as for example defined by standards for teachers and teacher education in many countries, expected to know methods of educational research (KMK, 2019), know basics in research, and exhibit research literacy (Révai, 2018). Future educational professionals need not only to develop professional knowledge and skills in education, but also a basic understanding of the scientific background of education and the skill to appraise this evidence. It is therefore beneficial for future educational professionals to learn how to directly evaluate scientific evidence from their area of expertise: education.

There are several approaches that could help derive criteria to evaluate scientific evidence, of which we will focus on two: the QUESTS dimensions teaching practices in medicine (Harden et al., 1999) and an instrument measuring the appraisal of scientific literature in evidence-based practice (Hetmanek, 2014; Trempler et al., 2015). QUESTS is an acronym that stands for the six dimensions that play a role in evaluating evidence in medical education: *Q*uality, *U*tility, *E*xtent, *S*trength, *T*arget, and *S*etting. Quality refers to the rigor of the study design, with randomized controlled trials gaining more points than case studies or professional experience. The utility refers to the extent to which the object of investigation, for example an intervention, can be transferred to a different setting. The extent of the available evidence refers to the difference between multiple studies with similar outcomes or meta-analyses in comparison to single studies. The strength of a study refers to strength of the effect(s) found in a study. Furthermore, the dimensions on target and setting address the validity of a certain target or outcome and setting (Harden et al., 1999). The instrument measuring the appraisal of scientific literature in evidence-based practice (Hetmanek, 2014; Trempler et al., 2015) also includes the dimension of the validity of the target and the setting, differentiated in the dimension of quality between the rigor in conducting the study, and statistical rigor. The authors also added the fit of the intervention, the applicability of the intervention, appropriate measurement of the target, and fit of the participants with a given educational decision (Hetmanek, 2014; Trempler et al., 2015). In summary, these sets of criteria include relevance criteria and quality criteria. Relevance criteria target the fit between an educational decision and the evidence that is currently evaluated with regards to the teaching method, the learning objective, the participants, and the setting. Quality criteria target (a) whether the reported learning outcome is measured with an objective, reliable, and face valid test, (b) the statistical power, and (c) the design of the study (Hetmanek, 2014; Trempler et al., 2015). The criteria used to appraise scientific evidence might vary substantially between domains (see for discussion, e.g., Fischer et al., 2018); thus, these criteria are specific to critically appraising scientific research literature in education.

The cognitive processing of critically appraising scientific literature has not yet been specified in detail. Our conceptualization of the skill to critically appraise scientific literature is based on the model of information problem solving (e.g., Brand-Gruwel et al., 2005, 2009) that describes a process that is structurally similar to critically appraising scientific literature. Information problem solving describes skills needed to solve a problem by searching for information (for example on the internet), scan and process the information, and combine the information at the end of the process to solve a problem. Most important for this research is the fourth process of the information problem solving described as process information, during which one gains a deeper understanding of a piece of information and, as described in the subskill selecting, uses criteria to judge the usefulness and quality of the information (Brand-Gruwel et al., 2009). Selecting describes a skill that is central to the process of appraising scientific literature. Similarly to the selection of information on the internet, it is important in the critical appraisal of scientific literature to determine the usefulness of the evidence and the quality of information by judging how relevant a study is and how well it was conducted (Sackett, 1997; Hetmanek, 2014; Trempler et al., 2015). Appraising scientific literature provides the additional challenge that the information is given in the specific format of an empirical research article. Furthermore, the set of criteria used to appraise scientific evidence in terms of its relevance and quality is rather complex: For example, the fit between learning objectives indicates low relevance if one's own educational decision aims at facilitating a cognitive skill and if the evaluated study reports motivational outcome measures. In this example, one needs to be able to identify the learning objective in the educational decision as well as the learning objective or measurement of the dependent variable in the empirical research article, and come to the correct inference that there is no overlap between them. As a second example, the design of a study is considered to be of high quality if the teaching method that is investigated in the study is varied between conditions while there are no further confounding variables that also vary between conditions. Here, one needs to know what an unconfounded design is and be able to detect whether the design described in an empirical research article only varies the independent variable between conditions or if there are confounding factors.

Evaluating information in an information problem solving process on the internet is difficult for students of all ages and, thus, rarely criteria-led (summarized by Brand-Gruwel et al., 2009). An analysis of think-aloud data showed that secondary students rarely appraise sources during information problem solving, and if they did, they only used a small selection of criteria in only a small percentage of cases (Walraven et al., 2009). Thus, it is not surprising that higher education students exhibit difficulties when asked to appraise empirical research articles in education (Trempler et al., 2015). Research showed that support in form of scaffolding (Raes et al., 2012), whole-task trainings (Frerejean et al., 2019), or long-term intervention programs (Argelagos and Pifarré, 2012) can support learners in their information problem solving. Interventions in related fields showed the potential of fostering scientific and evidence-based reasoning: For example, a short intervention teaching evidence-based medicine showed to improve medical students' search for scientific literature (Gruppen et al., 2005), educational science students' scientific argumentation was improved by engaging them in activities around an elaboration tool (Stark et al., 2009), and an intervention teaching heuristics in appraising and using scientific evidence to pre-service teachers fostered their evidence-based argumentation (Wenglein et al., 2015).

While there is little research investigating interventions teaching critical appraisal of scientific evidence in education, we expect higher education students to lack knowledge of appropriate criteria as well as the skill to apply these criteria in appraising literature. In line with the findings of intervention studies on information problem solving as well as studies in interventions in scientific reasoning, we expect higher education students to benefit from a training of their knowledge and skill in appraising scientific research articles.

The role of different learning activities

In training critical appraisal of scientific literature, it might be advantageous to facilitate high cognitive engagement in learners. The so-called ICAP framework proposes that the way learners engage with learning material or the instruction influences their cognitive engagement and thereby their learning outcomes (Chi, 2009; Chi and Wylie, 2014; Chi et al., 2018). The hypothesis is based on a taxonomy of learning activities, ranging from passive, to active, to constructive, to interactive learning activities. Students learning passively receive information from teachers or learning material without further engaging with the information. Active learners engage with information to some degree by, for example, repeating or rehearsing it, taking notes, highlighting text, or stopping a video. Constructive learning activities are those activities in which students generate learning outputs, additional to the information given to them, for example in the form of formulating self-explanations, generating inferences, or drawing a concept map. Learning interactively describes at least two participants that take turns in a constructive learning process. In the ICAP framework, active learning activities are hypothesized to exceed passive learning activities because they require focused attention and, thus, more cognitive engagement by the learners than passive learning. Constructive learning activities are hypothesized to exceed active learning activities because they prompt a more active construction of individual knowledge. Interactive learning activities are hypothesized to exceed constructive learning activities because they require learners to frequently update their mental model because of the ongoing change in the information discussed (Chi and Wylie, 2014). This framework links the learners' activities to the cognitive processes that they are engaged in during learning, it does not directly focus on the cognitive activity of a learner (which is distinct from other concepts using similar vocabulary such as the "cognitive activity" describing a state of deep learning, see Klieme and Rakoczy, 2008). There is some evidence about the hierarchy of learning activities, suggested by Chi and colleagues, based on the analysis of prior research (Chi, 2009; Chi and Wylie, 2014; Chi et al., 2018). Based on the publications of Chi and colleagues, some intervention studies investigated the effect of instructions aiming at different learning activities and found new insights into the effect of learning activities: Adding an instruction on how to interact for short periods of time with peers during a physics lecture was found to improve students' conceptual knowledge about Newtonian dynamics concepts, but constructive instruction or a combination of constructive and interactive instructions were not found to be more beneficial than passive instruction (Henderson, 2019). An interactive learning activity also showed to foster better conceptual understanding of material science and engineering than a constructive learning activity (Menekse and Chi, 2019).

The beneficial role of interactive learning in comparison to constructive learning is called into question by meta-analyses that investigated whether constructive or interactive activities could be found in different instructional interventions: A meta-analysis investigating the effect of socio-cognitive scaffolding on domain-specific knowledge and collaboration skills included the presence of interactive prompts in a moderator analysis and found no significant difference for domain-specific knowledge nor collaboration skills (Vogel et al., 2017). Similarly, a meta-analysis on constructive and interactive instructions fostering scientific reasoning found no significant difference between the interventions on scientific reasoning outcome measures (Engelmann et al., 2016). Yet, a meta-analysis of interventions studies fostering domain knowledge with a preparing-to-teach and teaching intervention found that an interaction, and even the expectation of an interaction, was associated with a higher effect size than non-interactive teaching activities (Kobayashi, 2019). A meta-analysis of learning with videos found interactivity to be a significant moderator: There was no learning benefit found for interventions in which the control condition included more interactivity than the experimental condition and a particularly high effect on learning found for videos with interactive context (Noetel et al., 2021). Thus, while there is a body on literature supporting the ICAP hypothesis, the meta-analytic evidence found in different contexts indicates that the effect of learning activities on different outcome measures needs to be investigated more thoroughly.

Not only the outcome measures vary between studies, Chi and Wylie (2014) also found learning activities to be embedded in different learning situations and instructional approaches, such as individual or collaborative note taking, individual or collaborative building of concept maps, explaining examples or explaining own versus others answers. Beyond the investigation of learning activities, self-explaining in learning with examples has been shown to facilitate argumentation (Schworm and Renkl, 2007) and learning from texts in higher education teaching (Lachner et al., 2021), while there is also some research showing no benefit of self-explanations, for example in teaching critical thinking skills (van Peppen et al., 2018). However, a meta-analysis found a medium effect size of self-explanation prompts on learning (Bisra et al., 2018), providing evidence for the beneficial effect of self-explanations.

Self-explanations are most commonly used in problem solving, text comprehension tasks, or example-based learning (Bisra et al., 2018). Since example-based learning has been found to be beneficial in early skill acquisition (Renkl, 2014), we will focus on self-explanations in example-based learning. Example-based learning provides learners with the solution of a given problem and commonly the steps that lead to the solution (Renkl, 2014). Learning from examples is more effective if the learners self-explain the solution to a problem (summarized by Renkl, 2014). In the analysis of learning activities by Chi and Wylie (2014), self-explanations were often found to be constructive because they, for example, asked learners to explain steps in a worked-example to themselves. Comparisons between learning activities were mostly found between these constructive learning activities and passive activities, such as self-explaining versus rereading or explaining others' solution versus just watching the solution. Alternatively, they were also found between constructive learning activities and interactive learning activities, such as explaining alone versus explaining with a partner (Chi and Wylie, 2014).

The present study

In the present study, we aim at testing the effect of constructive versus interactive learning activities in the context of an intervention that facilitates critical appraisal of scientific literature. Thus, the development of the intervention was guided by instructional approaches in which learning activities (Chi and Wylie, 2014) were investigated: The effect of different learning activities was found in intervention studies in which learning was implemented with note taking, concept mapping, or self-explaining (Chi and Wylie, 2014). Since we planned to integrate the learning activities in our practice tasks that was mainly aimed at early skill acquisition, we developed these tasks to ask the participants to explain a model solution to themselves or a learning partner; i.e. a version of example-based learning, (see Renkl, 2014), that only provided the solution not the steps that led to the solution. Thus, we embedded the learning activities in a type of task that has shown to foster learning outcomes similar to the ones in this study and provided a fruitful learning environment to investigate the effect of different learning activities in prior studies (Chi and Wylie, 2014). Furthermore, we designed this intervention to be integrated into a scenario that models a realistic situation in which educational professionals need to make evidence-based decisions (Hetmanek, 2014; Trempler et al., 2015).

Research questions

There is evidence for the advantage of interactive learning activities in comparison to constructive learning activities (Chi and Wylie, 2014). However, the superiority of interactive activities over constructive activities has not yet been replicated in some fields related to the field of this study: For example, Engelmann et al. (2016) showed similar effect sizes in learning outcomes for interventions with constructive versus interactive learning activities. Similarly, Vogel et al. (2017) were not able to establish a difference in domain knowledge between interventions that prompted interactivity and those that did not. Thus, this study investigates the role of interactive versus constructive learning activities in supporting students in learning to appraise scientific evidence. The content of the intervention in this study was chosen since, in comparison to research in medicine (e.g., Bradley et al., 2005; Kulier et al., 2012; Reviriego et al., 2014; Molléri et al., 2018), there is little research on fostering evidence-based practice in education, for example on teachers making decisions in lesson planning based on scientific evidence. The existing literature supports the claim that educational professionals need more training in doing so: Prior research showed that teachers rarely use scientific evidence in professional decisions (Hetmanek et al., 2015a) and rather refer to anecdotal evidence (Kiemer and Kollar, 2021; Menz et al., 2021). Thus, this study also provides a first insight into learning the critical appraisal of scientific literature.

Our research questions are as follows:

Research question 1: To what extent does an intervention with interactive learning activities advance knowledge about scientific criteria in comparison to an intervention with constructive learning activities?

Research question 2: To what extent does an intervention with interactive learning activities advance the skill in critical appraisal of scientific literature in comparison to an intervention with constructive learning activities?

We hypothesized that the interactive learning activities facilitate a higher level of cognitive activities during the learning process in comparison to constructive learning activities and, therefore, lead to a higher gain in knowledge and skills (Chi and Wylie, 2014).

Materials and methods

Sample

The sample size was calculated before data collection started, using the software G*Power 3 (Faul et al., 2007). The calculation was based on a target power of 80% and a medium effect size ($d = 0.5$) for the within-between interaction effect of an ANOVA with repeated measures and between-factor design. The estimation of the expected effect size was based on a medium effect ($d = 0.64$) found between an interactive and a constructive learning condition in an intervention study (see Menekse et al., 2013). In the condition with an interactive approach, the unit of analysis was a pair of learners and in the condition with the constructive approach, the unit of analysis was the individual learner. Thus, we aimed at a minimal sample size of 102 participants.

Participants were recruited *via* printed notices that were posted around the university campus and received 20 Euro for their participation. All participants met the following criteria: They were students at a university in their bachelor's or master's studies of educational sciences, psychology, teacher training, or an equivalent subject, who have not yet received a master's degree. 105 students participated in this study ($M_{age}=24.50$ years, $SD=4.03$; 84% female), 49 students in teacher training, 34 psychology students, 14 educational sciences students, and 8 other students. The distribution of the participants' study programs was similar in both conditions; we found no indicator of a systematic differences between knowledge or skills in participants from different study programs. No participant needed to be excluded from the data analysis.

Design

This experimental study was conducted in a between-subject repeated measures design. The participants were randomly assigned to one of the conditions and they learned interactively ($n=56$) or constructively ($n=49$) during an intervention about reading scientific literature. The random assignment was based on a 2:1 ratio, since the unit of analysis in the interactive condition was the pair of learners and the unit of analysis in the constructive condition was the individual learner. However, participants were assigned to the condition before the appointments for the data collections were set (because only one condition could be implemented for each appointment) and the attendance was greater in the constructive condition; thus, the distribution of participants was unequal in favor of the constructive condition. The pre- and posttests were parallel tests. Which version participants received as pretest and which as posttest was counterbalanced. Due to organizational reasons, 43 participants received one order, 62 participants the other order of tests. Data indicated that the order of the tests was not relevant for the outcome.

Setting, procedure, and manipulation of the independent variable

The data collection took place at computer labs of a German university. At each data collection, a maximum of eight participants could take part in the session. The session took approximately 114 min to complete. An overview of the procedure of the experiment can be found in Figure 1.

The whole data collection was set in a fictional scenario in which the participants were asked to imagine themselves to be an educational professional who needs to make two educational decisions in designing a lesson: (a) present some content themselves or use the jigsaw technique and (b) use correct and erroneous or only correct video examples. The participants were asked to appraise a set of four pieces of scientific evidence from educational research in order to make each of the decisions. The evidence from educational research was presented in the form of structured briefs of scientific literature. The scenario was adapted from a competence test measuring skills of information selection and appraisal of scientific research articles (Hetmanek, 2014; Trempler et al., 2015). The first decision constituted the pretest and was referenced in the intervention. The second decision constituted the posttest. With which decision (a or b) learners worked with in the pretest and with which they worked with in the posttest was counterbalanced (see above). For each decision participants were given a short description of the learning content of the lesson, characteristics of the students and the setting. In a next step, they were asked to appraise a set of four pieces of scientific literature and make a decision regarding the teaching method based on this evidence. The intervention itself was nested between the two decisions building upon the first decision to create an authentic reference point for the participants.

The procedure was similar in both conditions: In a first phase, the participants were introduced to the scenario, the first decision, and filled out the pretest by evaluating the pieces of evidence related to the first decision. The second phase included the intervention. In the third phase, the participants were introduced to the second decision and filled out the posttest by evaluating the pieces of evidence related to the second decision.

The intervention started with a written four-page introduction to the criteria used in appraising scientific evidence. The introduction included the following criteria:

– evaluating the relevance of the evidence with respect to

a. the fit between the instructional methods that are part of the decision to be made and the instructional methods investigated in the piece of evidence,
b. the fit between the learning goal in the decision to be made and the dependent variable in the piece of evidence,
c. the fit between the learners' age or level of prior knowledge specified in the decision to be made and the age or level of prior knowledge of the participants in the piece of evidence, and
d. the fit between the setting in the decision to be made and the setting in the piece of evidence.

– evaluating the quality of the evidence was specified in

e. the quality of the measurement of the dependent variable,
f. the statistical robustness of the results, and
g. whether the effect found in the study could be attributed to the instructional method investigated in the piece of evidence.

For each criterion, the introduction provided a short description on how to evaluate the degree to which a piece of evidence is relevant and of high quality. Subsequently, the participants were presented with model solution of the appraisal

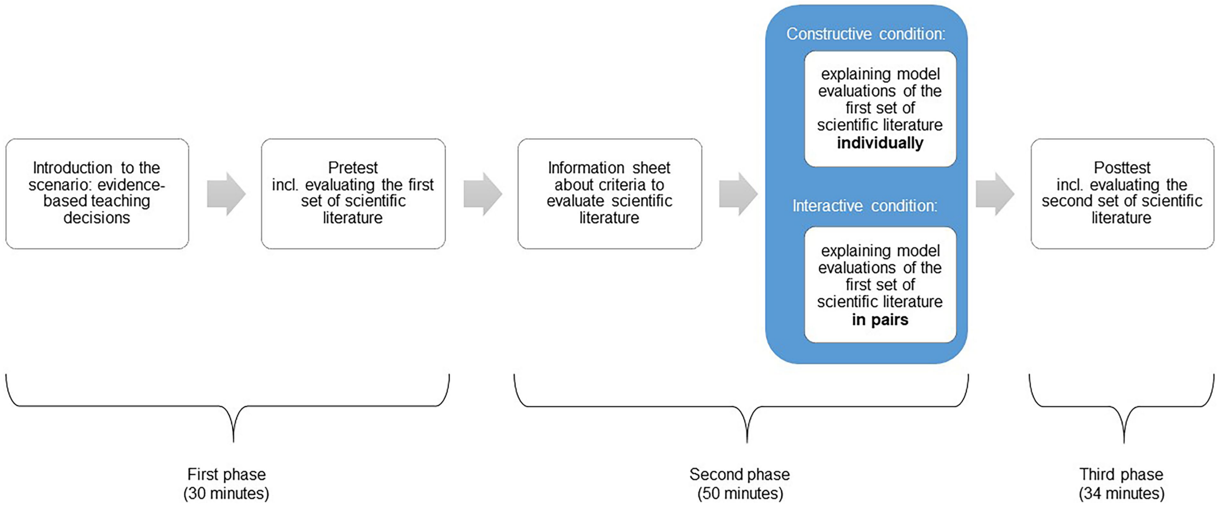


FIGURE 1
Procedure of the experimental study over 2 h.

of the first piece of evidence from their pretest given by a fictional character that was described as knowledgeable. All participants were asked to compare the model solution of each criterion to the description in the introduction text and explain to themselves or their learning partner (depending on the condition) which information in the piece of evidence led to the example appraisal and why the knowledgeable person came to this result. Since each pretest consisted of four pieces of evidence, the participants moved through four model solutions. The learning phase was limited to 50 minutes.

The whole experiment was conducted in an online learning environment, except for the written introduction to the appraisal criteria that was given to the participants on a sheet of paper at the beginning of the second phase and was taken from the participants at the end of the second phase, before the posttest was conducted.

The manipulation of the independent variable was realized during the intervention phase: (a) Participants in the interactive conditions were prompted to explain the model solutions to their learning partner. In order to make sure that the pairs actually learned interactively, they were alternately scaffolded to explain the model solutions to their learning partner or to question the solution provided by the partner. (b) Participants in the constructive condition were prompted to explain the model solution to themselves. All participants were asked to take notes.

Measurement of the dependent variables

Knowledge

The knowledge was measured with a test containing two open questions asking the participants to recall criteria to appraise the relevance and quality of scientific literature. The test was developed for this study. The answers given by the participants were saved in written form and were coded according to the following categories: For the first question regarding the relevance of a study, a point was given for mentioning the independent variable (e.g., "jigsaw technique," "task," "lesson"), the dependent variable (e.g., "measurement," "learning goal"), the characteristic of the students (e.g., "prior knowledge," "students") and the setting (e.g., "situation," "field of application"). During the coding process, we added one additional category: mentioning the research question or object of the investigation in the piece of evidence. We did so, because many participants mentioned this category instead of recalling the independent and dependent variable. One of the initial categories, coding whether participants mentioned the quality of a study as a criterion for the relevance of a study, was dropped because it could not be coded reliably. For the second question regarding the quality of a study, one point was given for mentioning the quality of the measurement (e.g., "objectivity," "reliability"), the clear design (e.g., "randomizing participants," "execution of the experiment," "validity"), and the statistical significance and/or power (e.g., "number of participants," "power," "statistical significance"). One point was given for each aspect mentioned, creating a knowledge score ranging between zero and nine points. The first author coded all data and the second author coded 10% of the data after a coding training to test the objectivity of the coding process. The agreement between coders showed to be ranged from acceptable to very good for all included categories (Cohen's kappa ranging from 0.77 to 1.00). The same scale was administered as pre- and posttest.

Skill

The skill in appraising scientific evidence was measured by asking the participants to appraise four pieces of evidence, each by evaluating the evidence on seven dimensions. Each dimension consisted of a multiple-choice item with four response options each (very high, somewhat high, somewhat low, and very low), one of which was correct. The items asked the participants to appraise different aspects of the scientific text, for example (translated into English) "The study described in the structured brief investigates an educational intervention that matches the educational intervention in my decision." or "The study described in the structured brief uses appropriate tests for its performance measures." The measured dimensions matched the criteria taught during the intervention. The test was adapted from a sub-scale of a test for evidence-based practice in education, developed by Hetmanek (2014) and Trempler et al. (2015). For each item, we saved the response that was selected by the participants. The responses were aggregated into a scale in which (a) a correct response option was counted as one point, (b) a false response option leaning into the same direction as the correct response option was counted as half a point, (c) a false response option leaning into the other direction as the correct response option was counted as zero points. Based on the 4 × 7 items, the aggregation resulted in a score ranging from zero to 28 points. The reliability of the scales was low to medium for the pretest ($\alpha = 0.53/0.54$) and posttest ($\alpha = 0.45/0.64$) for both versions of the test. However, we decided against selecting or discarding items to achieve a more homogeneous measure. Each of the items covers a different important aspect of critically appraising scientific literature, discarding one or more aspects of the test could have increased the Cronbach's alpha of the scale but would have decreased the validity of the test. Parallel scales were administered as pre- and posttest containing the exact same phrasing, targeting different sets of literature in pre- and posttest.

Statistical analysis

The analyses reported in this article were registered at the Open Science Framework (Engelmann et al., 2018). The registration was done after data collection but before any analysis was conducted.

In analyzing the learning pairs of the interactive condition, one person of each pair was randomly chosen and their data was used for the analysis.

Bayesian repeated measures ANOVAs with a between-subjects factor were used with priors kept at standard values, examining the BF_{10}, compared to the null model and investigating the effects across all models. The interpretation of the results was based on van den Bergh et al. (2020) and Wagenmakers et al. (2018a). The grouping variable (interactive condition versus constructive condition) was the independent variable in analyzing both research questions. The score in knowledge was the dependent variable for the first analysis, the score in the skill measurement was the dependent variable in the second analysis. Q-Q plots for both variables did not indicate non-normality. We decided to utilize a Bayesian approach in comparison to classical null hypothesis statistical testing because it provides the opportunity to quantify the evidence that the data provides for the null as well as the alternative hypothesis: "In the Bayesian framework, no special status is attached to either of the hypotheses under test" (Wagenmakers et al., 2018b, p. 46). We would also like to draw conclusions about outcomes that do not support our hypotheses. A Bayesian approach allows for that by providing evidence in favor of the null hypothesis, in favor of the alternative hypothesis, or neither (Wagenmakers et al., 2018b). The strength of the evidence is also provided (e.g., moderate evidence, strong evidence, very strong evidence) in comparison to an (arguably arbitrary) level of significance.

The sequential analyses add evidential trajectories, showing how the evidence for one of the hypotheses increases, decreases, or remains the same with each additional datapoint (Marsman and Wagenmakers, 2017). Thus, the sequential analyses also give additional information about the number of data points that were needed to reach a certain Bayes factor. Since the sequential analysis could not be conducted in the repeated measures design, the learning gain was calculated as the difference between pre- and posttest score and the sequential analysis was conducted using a Bayesian independent and paired sample *t*-test.

All tests were conducted using the software JASP, Version 0.13.1.0 (JASP Team, 2020). Additionally, Cohen's *d* of all reported mean differences were calculated (Cohen, 1988).

Results

Research question 1: To what extent does an intervention with interactive learning activities advance knowledge about scientific criteria in comparison to an intervention with constructive learning activities?

The descriptive values of knowledge are displayed in Table 1. The results of the Bayesian repeated measures analysis of variance can be found in Table 2 for the comparison of each model to the null model and Table 3 for the average across all models. The BF_{incl} (the inclusion Bayes factors) in Table 3 are of particular interest because they indicate the amount of evidence for the variable averaged over all models; thus, the BF_{incl} can be interpreted as the evidence found in the data supporting a certain variable (van den Bergh et al., 2020). The results showed extreme evidence (Schönbrodt and Wagenmakers, 2018) for an effect of the time on knowledge ($BF_{incl} = 592595.75$). However, the results regarding the condition ($BF_{incl} = 0.22$) and the interaction between time and condition ($BF_{incl} = 0.21$) provided moderate evidence for the null hypothesis (Schönbrodt and Wagenmakers, 2018). Thus, they provided evidence for the conclusion that knowledge is not

TABLE 1 Descriptive values of knowledge and skills before (pretest) and after (posttest) the intervention.

Time	Condition	*n*	Knowledge		Skill	
			M	*SD*	*M*	*SD*
Pretest	Constructive	46	1.61	1.41	11.82	2.97
	Interactive	28	1.43	1.00	10.73	3.27
Posttest	Constructive	46	3.02	2.15	14.84	3.52
	Interactive	28	2.89	2.03	14.25	3.44

TABLE 2 Model comparison for the Bayesian repeated measures ANOVA on knowledge.

Models	*P*(M)	*P*(*M*|data)	BF_M	BF_{10}	Error %
Null model (incl. subject)	0.20	9.062e-7	3.625e-6	1.00	
Time (pretest versus posttest)	0.20	0.76	12.32	832,984.82	1.96
Time + Condition	0.20	0.20	0.97	216,096.74	2.23
Time + Condition + Time * Condition	0.20	0.05	0.21	54,384.09	4.24
Condition (constructive versus interactive)	0.20	2.188e-7	8.750e-7	0.24	2.82

TABLE 3 Analysis of effects for the Bayesian repeated measures ANOVA on knowledge.

Effects	*P*(incl)	*P*(incl|data)	BF_{incl}
Time (pretest versus posttest)	0.60	1.00	592,595.75
Condition (constructive versus interactive)	0.60	0.25	0.22
Time * Condition	0.20	0.05	0.21

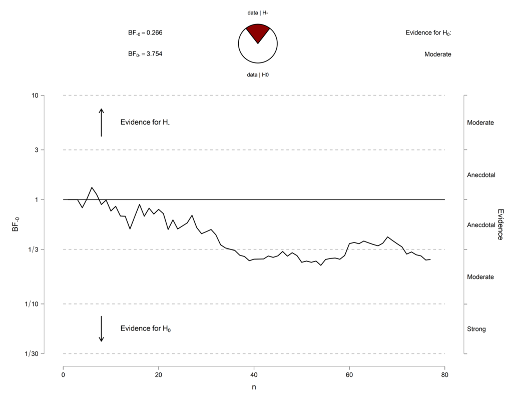


FIGURE 2
Results of the sequential analysis for the effect of the condition on knowledge.

affected by the difference between the constructive or interactive instruction. The results of the Bayesian independent and paired sample *t*-test regarding the learning gain in knowledge (see Figures 2, 3) are similar to the findings of the Bayesian repeated measures analysis of variance. The trajectories of the Bayes factors in the sequential analyses showed that these results were already present after approximately 40 participants.

The difference between the mean knowledge score in pre- and posttest would reflect a large effect size for constructive learners ($d = 0.76$) and a large effect size for interactive learners ($d = 0.91$); the difference between constructive and interactive learners in the pretest ($d = 0.15$) and posttest ($d = 0.06$) not even a small effect (see Cohen, 1988).

Research question 2: To what extent does an intervention with interactive learning activities advance the skill in critical appraisal of scientific literature in comparison to an intervention with constructive learning activities?

The descriptive values of the skill are displayed in Table 1. The Bayesian repeated measures analysis (see Table 4 for the

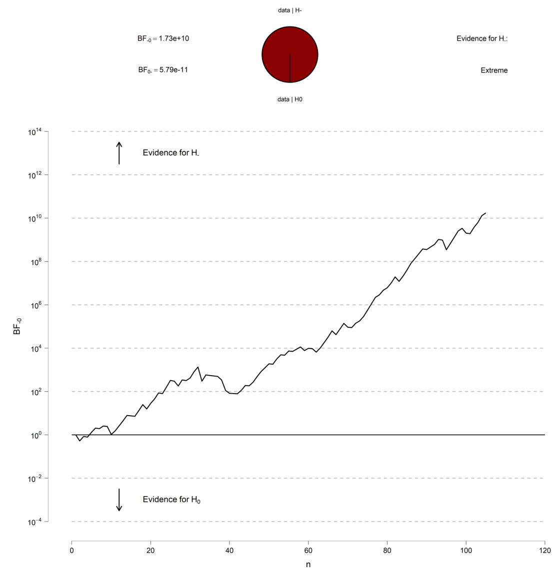


FIGURE 3
Results of the sequential analysis for the effect of time on knowledge.

comparison of each model to the null model and Table 5 for the average across all models) of variance showed extreme evidence (Schönbrodt and Wagenmakers, 2018) for an effect of the time on skill ($BF_{incl} = 3.020 \times 10^7$). However, the results regarding the condition ($BF_{incl} = 0.42$) and the interaction between time and condition ($BF_{incl} = 0.36$) provided anecdotal evidence for the null hypothesis (Schönbrodt and Wagenmakers, 2018). Thus, the data is not sufficiently informative to provide information on how skill might be affected by the difference between the constructive or interactive instruction. The results of the Bayesian independent and paired sample *t*-test regarding the learning gain in skill (see Figures 4, 5) are similar to the findings of the Bayesian repeated measures analysis of variance. The trajectories of the Bayes factors in the sequential analyses showed that these

TABLE 4 Model comparison for the Bayesian repeated measures ANOVA on skill.

Models	$P(M)$	$P(M \mid \text{data})$	BF_M	BF_{10}	Error %
Null model (incl. subject)	0.20	1.572e-8	6.286e-8	1.00	
Time (pretest versus posttest)	0.20	0.61	6.34	3.900e+7	1.30
Time + Condition	0.20	0.31	1.76	1.944e+7	2.70
Time + Condition + Time * Condition	0.20	0.08	0.36	5.185e+6	2.24
Condition (constructive versus interactive)	0.20	6.362e-9	2.545e-8	0.41	1.10

TABLE 5 Analysis of effects for the Bayesian repeated measures ANOVA on skill.

Effects	$P(\text{incl})$	$P(\text{incl} \mid \text{data})$	BF_{incl}
Time (pretest versus posttest)	0.60	1.00	3.020e+7
Condition (constructive versus interactive)	0.60	0.39	0.42
Time * Condition	0.20	0.08	0.36

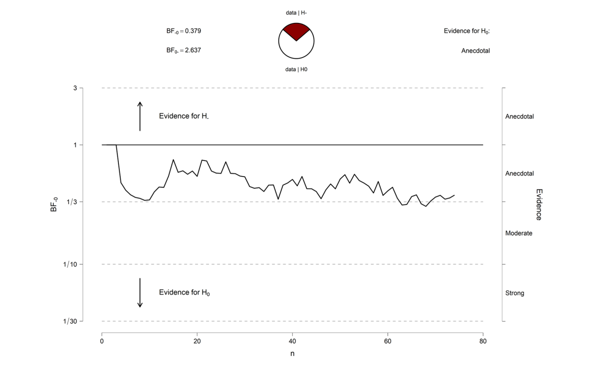


FIGURE 4
Results of the sequential analysis for the effect of the condition on skill.

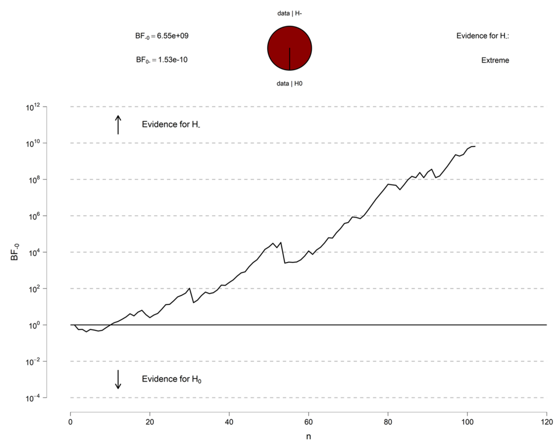


FIGURE 5
Results of the sequential analysis for the effect of time on skill.

results were already present after approximately 40 participants regarding the effect of time, but stayed variate regarding the effect of the condition.

The difference between the mean skill score in pre- and posttest would reflect a large effect size for constructive learners ($d = 0.93$) and a large effect size for interactive learners ($d = 1.05$); the difference between constructive and interactive learners in the pretest ($d = 0.35$) and posttest ($d = 0.17$) would reflect a small and not even a small effect size, respectively (see Cohen, 1988).

Discussion

This study aimed at testing the effect of an interactive versus a constructive approach in an intervention fostering knowledge and skills in appraising scientific literature.

The interactive condition did not outperform the constructive condition, contrasting our hypothesis. This is not coherent with the research on learning activities comparing passive, active, constructive, and interactive learning activities in interventions (Chi, 2009; Menekse et al., 2013; Chi and Wylie, 2014; Chi et al., 2018; Menekse and Chi, 2019). However, the Bayesian analysis only provided anecdotal to moderate evidence for the similarity of the constructive and interactive conditions. Based on these results, we conclude that the difference in learning with constructive or interactive learning activity alone is not sufficient to explain knowledge and skill gain in this content area. We expected the learners in the interactive conditions to benefit from frequently updating their mental model because of the ongoing change in the information discussed (Chi and Wylie, 2014) and we would expect that this mechanism would also take place in the intervention of this study. However, our operationalization of the learning environment might have affected the mechanism that is hypothesized to be caused by the interactive learning activities: prompting participants to frequently update their mental model because of the ongoing change in the information. The learners had to explain a model solution of the task (in appraising evidence) in both conditions. And while there was no other person in the constructive condition, there were written statements of another person that the students were asked to explain in both conditions. This factor could have caused students in both conditions to update their mental model more than once. In the ICAP framework it is hypothesized that there is a systematic relationship between the learning activity exhibited by the students and their cognitive process (Chi, 2009; Chi and Wylie, 2014; Chi et al., 2018). However, based on our results we would add that there might be other factors that also strongly influence this cognitive process, such as aspects of the learning task beyond the constructive versus interactive distinction. The prompt to update one's mental model more frequently might also be given by learning material that asks learners to explain new information.

Furthermore, we found students in both conditions to show a rise of, approximately, one standard deviation in knowledge and skills in appraising literature after an intervention that took less than an hour, far beyond the scope of a mere retest effect that could be expected in cognitive abilities (Scharfen et al., 2018). Participating in an intervention that asked learners to explain model solutions (thus, a short version of example-based learning, Renkl, 2014) to themselves or to learning partner (thus, self-explained the causal connection between the overview of how to appraise scientific evidences and the model solutions, *cf.* Bisra et al., 2018) seems to advance students' knowledge and skills in appraising scientific evidence. The results are consistent with prior research showing that scientific reasoning in general can be facilitated by interventions (Engelmann et al., 2016), more specifically, higher education students' scientific reasoning skills (e.g., Gruppen et al., 2005; Stark et al., 2009; Wenglein et al., 2015). Whether there is a causal relationship between this intervention and the learning gain in knowledge and skills needs to be tested in a future experiment.

One aspect for further research might be the domain specificity of the task. In this intervention, we only used educational research articles reporting experimental or quasi-experimental intervention studies, investigating an effect of an instruction or educational support. Also, we only included participants who were familiar with the general topics of this intervention: students in teacher education, educational sciences, and psychology. We did so, because some degree of domain-specific knowledge is necessary for the first-hand evaluation of the evidence (Bromme et al., 2010). The intervention and the measurement of knowledge and skills in appraising scientific literature was kept narrow in range. The material of intervention and tests were about learning with examples and the jigsaw technique, all studies employed a quantitative approach, and the structured briefs of scientific articles were structured similarly, focusing on one main research question (Hetmanek et al., 2015b). Thus, we did not examine to which extent the effect of the intervention could be transferable to appraising scientific literature in educational sciences that employ a different methodological approach or scientific literature that was presented in a different format. A wider range of scientific literature might benefit from a combined approach of first-hand and second-hand evaluation skills. Moreover, teachers are expected to read beyond educational sciences: this includes literature about the subjects that they are teaching, e.g., biology, mathematics, or history. The intervention presented in this paper might not be completely dependent on the content of the literature that is to be appraised, since we integrated two different educational topics in the intervention. However, a change in the methodological approach of the studies or in the format of the presentation might change the effect of this intervention. Further research in this type of intervention should systematically broaden the types of scientific research articles that are used as scientific evidence.

The study presented in this paper has several limitations. First, the posttest was conducted right after the intervention. Thus, we cannot make any generalization about long-term effects of facilitating the appraisal of scientific literature.

Second, the reliability of the skill measurement was relatively low. This can be explained by the conceptual breath of the scale. Each item in the skill measure targets a different aspect in which the scientific literature is evaluated. Each item covered one aspect that was also targeted in the intervention. The validity of the measure for the intervention would be decreased by removing any of the items to gain a higher reliability. Moreover, the measure was adapted from a validated scale (Trempler et al., 2015) by only changing the response options from a nine-point scale to four options. The wording of some items was slightly changed. Furthermore, the pattern of results is rather similar in knowledge and skills, showing a relevant difference between pre- and posttest and very little difference between the conditions. Thus, we do not expect the rather low reliability of the skill measure to have significantly impacted the results found in this study.

Third, while the intervention targeted a complex and large area of knowledge and skills, this intervention took less than 2 h. We designed the intervention to fit our scale in trying to focus on core knowledge and skills in appraising central aspects of scientific evidence. Still, the duration of the intervention might have been insufficient to teach a more comprehensive idea of appraising scientific literature. A longer intervention could give participants more time to practice appraising the evidence, which could have led to more knowledge gain (as has been discussed for decades by, e.g., Anderson, 1981; Berliner, 1990; van Gog, 2013). Specifically, students working interactively might need more time and instructional support in evidence-based argumentation (Csanadi et al., 2021). However, the intervention was kept short in order to compare the effect of this intervention to interventions of similar length: The study that compared constructive and interactive learning activities and found a higher learning gain for interactive learning in comparison to constructive learning with an effect size of $d = 0.64$ gave the students 25–30 min in the learning phase (Menekse et al., 2013).

Conclusion

This study showed the limitations of the hypothesis that interactive learning activities are accompanied with higher learning gains (Chi and Wylie, 2014) in teaching critical appraisal of scientific evidence. We suggest to expand the ICAP hypothesis to include more dimensions that influence the underlying cognitive processes, such as characteristics of a learning task in differentiating constructive and interactive learning activities. Based on the results of this study, we hypothesize that interactive learning might not require a person to discuss with, interaction might also be achieved with learning material that imitates an interaction or implements other ways for the learners to frequently update their mental model because of the ongoing change in the information discussed (*cf.* Chi and Wylie, 2014). Future studies are necessary to (a) investigate how the learning material must be designed and implemented to reach interactive learning activities that initiate learning processes similar to the learning process in interactive learning activities in cooperative settings and (b) investigate the interactive components in interventions, such as the constructive condition in this intervention to understand which aspects of the material is actually responsible for the beneficial learning process.

Evidence-based decisions are considered important for educational professionals. In this study, we implemented an intervention that facilitated future educational professionals to appraise scientific evidence in order to make evidence-based decisions in day-today practice. This study suggests that an intervention implementing constructive or interactive learning activities in studying sample solutions and self-explaining the examples facilitated higher education students' critical appraisal of scientific evidence. So far, the effect of the intervention was only observed in a measurement of appraising scientific literature. It would be interesting for future research to investigate the effect of this intervention on a broader measurement of scientific reasoning.

Data availability statement

The data supporting the conclusions of this article will be made available by the first author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

KE, AH, and FF developed the study concept and the study design. KE and AH developed and adapted the material and coded the data. KE planned the data collection, processed the data, performed the data analysis, interpreted the results, and took the lead in writing the manuscript. KE, AH, BN, and FF contributed to interpreting the results and writing the manuscript. All authors approved the final version of the manuscript for submission.

Funding

This work was supported by the Elite Network of Bavaria [K-GS-2012-209], the Center for Advanced Studies at Ludwig-Maximilians-Universität München, and Stiftung Universität Hildesheim.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anderson, L. W. (1981). Instruction and time-on-task: a review. *J. Curriculum Stud.* 13, 289–303. doi: 10.1080/0022027810130402

Argelagos, E., and Pifarré, M. (2012). Improving information problem solving skills in secondary education through embedded instruction. *Comput. Hum. Behav.* 28, 515–526. doi: 10.1016/j.chb.2011.10.024

Berliner, D. C. (1990). "What's all the fuss about instructional time" in *The nature of time in schools: Theoretical concepts, practitioner perceptions*. eds. M. Ben-Peretz and R. Bromme (Teachers College Press). 3–35.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., and Winne, P. H. (2018). Inducing self-explanation: a meta-analysis. *Educ. Psychol. Rev.* 30, 703–725. doi: 10.1007/s10648-018-9434-x

Bradley, P., Oterholt, C., Herrin, J., Nordheim, L., and Bjørndal, A. (2005). Comparison of directed and self-directed learning in evidence-based medicine: a randomised controlled trial. *Med. Educ.* 39, 1027–1035. doi: 10.1111/j.1365-2929.2005.02268.x

Brand-Gruwel, S., Wopereis, I., and Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Comput. Hum. Behav.* 21, 487–508. doi: 10.1016/j.chb.2004.10.005

Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004

Bråten, I., Strømsø, H. I., and Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learn. Instr.* 21, 180–192. doi: 10.1016/j.learninstruc.2010.02.002

Bromme, R., Kienhues, D., and Porsch, T. (2010). "Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) attained from others," in *Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice*. eds. L. D. Bendixen and F. C. Feucht (Cambridge: Cambridge University Press), 163–193.

Bromme, R., Prenzel, M., and Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik [educational research and evidence based educational policy]. *Z. Erziehungswiss.* 17, 3–54. doi: 10.1007/s11618-014-0514-5

Brown, C., and Zhang, D. (2016). Is engaging in evidence-informed practice in education rational? What accounts for discrepancies in teachers' attitudes towards evidence use and actual instances of evidence use in schools? *Br. Educ. Res. J.* 42, 780–801. doi: 10.1002/berj.3239

Cain, T. (2016). Research utilisation and the struggle for the teacher's soul: a narrative review. *Eur. J. Teach. Educ.* 39, 616–629. doi: 10.1080/02619768.2016.1252912

Chi, M. T. (2009). Active-constructive-interactive: a conceptual framework for differentiating learning activities. *Top. Cogn. Sci.* 1, 73–105. doi: 10.1111/j.1756-8765.2008.01005.x

Chi, M. T., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., et al. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cogn. Sci.* 42, 1777–1832. doi: 10.1111/cogs.12626

Chi, M. T., and Wylie, R. (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* 49, 219–243. doi: 10.1080/00461520.2014.965823

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (*2. Auflage*). Hillsdale, NJ: Erlbaum.

Cook, B. G., Smith, G. J., and Tankersley, M. (2012). "Evidence-based practices in education," in *APA Educational Psychology Handbook:Theories, Constructs, and Critical Issues Vol. 1*. eds. K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra and J. Sweller (American Psychological Association), 495–527.

Csanadi, A., Kollar, I., and Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Detrich, R., and Lewis, T. (2013). A decade of evidence-based education: where are we and where do we need to go? *J. Posit. Behav. Interv.* 15, 214–220. doi: 10.1177/1098300712460278

Dickersin, K., Straus, S. E., and Bero, L. A. (2007). Evidence based medicine: increasing, not dictating, choice. *BMJ* 334:s10. doi: 10.1136/bmj.39062.639444.94

Engelmann, K., Hetmanek, A., Neuhaus, B. J., and Fischer, F. (2018). Reading scientific articles: facilitating the evaluation of structured briefs of scientific literature. doi: 10.17605/OSF.IO/54F7J

Engelmann, K., Neuhaus, B. J., and Fischer, F. (2016). Fostering scientific reasoning in education–meta-analytic evidence from intervention studies. *Educ. Res. Eval.* 22, 333–349. doi: 10.1080/13803611.2016.1240089

Faul, F., Erdfelder, E., Lang, A. G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Fischer, F., Chinn, C. A., Engelmann, K., and Osborne, J. (Eds.). (2018). *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*. New York, NY: Routledge.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45.

Frerejean, J., Velthorst, G. J., van Strien, J. L., Kirschner, P. A., and Brand-Gruwel, S. (2019). Embedded instruction to learn information problem solving: effects of a whole task approach. *Comput. Hum. Behav.* 90, 117–130. doi: 10.1016/j.chb.2018.08.043

Gruppen, L. D., Rana, G. K., and Arndt, T. S. (2005). A controlled comparison study of the efficacy of training medical students in evidence-based medicine literature searching skills. *Acad. Med.* 80, 940–944. doi: 10.1097/00001888-200510000-00014

Harden, R. M., Grant, J., Buckley, G., and Hart, I. R. (1999). BEME guide no. 1: best evidence medical education. *Med. Teach.* 21, 553–562. doi: 10.1080/01421599978960

Henderson, J. B. (2019). Beyond "active learning": how the ICAP framework permits more acute examination of the popular peer instruction pedagogy. *Harv. Educ. Rev.* 89, 611–634. doi: 10.17763/1943-5045-89.4.611

Hetmanek, A. (2014). Evidenzbasierte Praxis im Bildungsbereich: Standortbestimmung und Vorarbeiten zur Förderung in drei empirischen Studien (Doctoral dissertation). München.

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M. R., Gräsel, C., et al. (2015a). Wozu nutzen Lehrkräfte welche Ressourcen? [For what do teachers use which kind of resource?]. *Unterrichtswissenschaft* 43, 193–208.

Hetmanek, A., Wecker, C., Trempler, K., Kiesewetter, J., Gräsel, C., Fischer, M. R., et al. (2015b). Structured briefs provide a means to communicate research more efficiently to practitioners. Paper presented at 16th Biennial EARLI Conference "Towards a Reflective Society", Limassol, Cyprus, 25–29 August 2015.

JASP Team (2020). *JASP (Version 0.13.1.0) [Computer software]*. Amsterdam: JASP Team

Klieme, E., and Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik* 54, 222–237.

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Z. Pädagog. Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

KMK (2019). Standards für die Lehrerbildung: Bildungswissenschaften. Available at: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf

Kobayashi, K. (2019). Learning by preparing-to-teach and teaching: a meta-analysis. *Jpn. Psychol. Res.* 61, 192–203. doi: 10.1111/jpr.12221

Kulier, R., Gülmezoglu, A. M., Zamora, J., Plana, M. N., Carroli, G., Cecatti, J. G., et al. (2012). Effectiveness of a clinically integrated e-learning course in evidence-based medicine for reproductive health training: a randomized trial. *JAMA* 308, 2218–2225. doi: 10.1001/jama.2012.33640

Lachner, A., Jacob, L., and Hoogerheide, V. (2021). Learning by writing explanations: is explaining to a fictitious student more effective than self-explaining? *Learn. Instr.* 74:101438. doi: 10.1016/j.learninstruc.2020.101438

Marsman, M., and Wagenmakers, E. J. (2017). Bayesian benefits with JASP. *Eur. J. Dev. Psychol.* 14, 545–555. doi: 10.1080/17405629.2016.1259614

Menekse, M., and Chi, M. T. (2019). The role of collaborative interactions versus individual construction on students' learning of engineering concepts. *Eur. J. Eng. Educ.* 44, 702–725. doi: 10.1080/03043797.2018.1538324

Menekse, M., Stump, G. S., Krause, S., and Chi, M. T. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *J. Eng. Educ.* 102, 346–374. doi: 10.1002/jee.20021

Menz, C., Spinath, B., and Seifried, E. (2021). Where do pre-service teachers' educational psychological misconceptions come from? *Z. Pädagog. Psychol.* 1:14.

Molléri, J. S., Bin Ali, N., Petersen, K., Minhas, N. M., and Chatzipetrou, P. (2018). Teaching students critical appraisal of scientific literature using checklists. In Proceedings of the 3rd European Conference of Software Engineering Education New York, NY: Association for Computing Machinery (pp. 8–17).

Noetel, M., Griffith, S., Delaney, O., Sanders, T., Parker, P., del Pozo Cruz, B., et al. (2021). Video improves learning in higher education: a systematic review. *Rev. Educ. Res.* 91, 204–236. doi: 10.3102/0034654321990713

Raes, A., Schellens, T., De Wever, B., and Vanderhoven, E. (2012). Scaffolding information problem solving in web-based collaborative inquiry learning. *Comput. Educ.* 59, 82–94. doi: 10.1016/j.compedu.2011.11.010

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cogn. Sci.* 38, 1–37. doi: 10.1111/cogs.12086

Révai, N. (2018). What difference do standards make to educating teachers? A review with case studies on Australia, Estonia and Singapore. *OECD Educat. Work. Pap.* 174:0_1-70.

Reviriego, E., Cidoncha, M. Á., Asua, J., Gagnon, M. P., Mateos, M., Gárate, L., et al. (2014). Online training course on critical appraisal for nurses: adaptation and assessment. *BMC Med. Educ.* 14, 1–10. doi: 10.1186/1472-6920-14-136

Sackett, D. L. (1997). "Evidence-based medicine," in *Seminars in Perinatology*, vol. *21* (Philadelphia, PA: WB Saunders), 3–5.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence-based medicine: what it is and what it is not. *BMJ* 312, 71–72. doi: 10.1136/bmj.312.7023.71

Scharfen, J., Peters, J. M., and Holling, H. (2018). Retest effects in cognitive ability tests: a meta-analysis. *Intelligence* 67, 44–66. doi: 10.1016/j.intell.2018.01.003

Schönbrodt, F. D., and Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review* 25, 128–142.

Schworm, S., and Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *J. Educ. Psychol.* 99, 285–296. doi: 10.1037/0022-0663.99.2.285

Slavin, R. E. (2008). Perspectives on evidence-based research in education—what works? Issues in synthesizing educational program evaluations. *Educ. Res.* 37, 5–14. doi: 10.3102/0013189X08314117

Stark, R. (2017). Probleme evidenzbasierter bzw.-orientierter pädagogischer Praxis. [Problems of evidence-based or rather evidence-oriented educational practice]. *Z. Pädagog. Psychol.* 31, 99–110. doi: 10.1024/1010-0652/a000201

Stark, R., Puhl, T., and Krause, U. M. (2009). Improving scientific argumentation skills by a problem-based learning environment: effects of an elaboration tool and relevance of student characteristics. *Evaluat. Res. Educat.* 22, 51–68. doi: 10.1080/09500790903082362

Thomm, E., and Bromme, R. (2016). How source information shapes lay interpretations of science conflicts: interplay between sourcing, conflict explanation, source evaluation, and claim evaluation. *Read. Writ.* 29, 1629–1652. doi: 10.1007/s11145-016-9638-8

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Z. Pädagog. Psychol.* 35, 173–184. doi: 10.1024/1010-0652/a000301

Trempler, K., Hetmanek, A., Wecker, C., Kiesewetter, J., Fischer, F., Fischer, M. R., et al. (2015). Nutzung von Evidenz im Bildungsbereich – Validierung eines Instruments zur Erfassung von Kompetenzen der Informationsauswahl und Bewertung von Studien [Use of evidence in education - validation of a tool for measuring competences in information selection and evaluation of studies]. *Z. Pädagogik* 61, 144–166.

van den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E. J., Derks, K., et al. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *LAnnee Psychol.* 120, 73–96. doi: 10.3917/anpsy1.201.0073

van Gog, T. (2013). "Time on task," in *International Guide to Student Achievement*. eds. J. Hattie and E. M. Anderman (New York: Routledge)

Van Peppen, L. M., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., Janssen, E. M., Koopmans, D., and van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Front. Educat.* 3:100. doi: 10.3389/feduc.2018.00100

Vogel, F., Wecker, C., Kollar, I., and Fischer, F. (2017). Socio-cognitive scaffolding with computer-supported collaboration scripts: a meta-analysis. *Educ. Psychol. Rev.* 29, 477–511. doi: 10.1007/s10648-016-9361-7

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., et al. (2018a). Bayesian inference for psychology. Part II: example applications with JASP. *Psychon. Bull. Rev.* 25, 58–76. doi: 10.3758/s13423-017-1323-7

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2018b). Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* 25, 35–57. doi: 10.3758/s13423-017-1343-3

Walraven, A., Brand-Gruwel, S., and Boshuizen, H. P. (2009). How students evaluate information and sources when searching the World Wide Web for information. *Comput. Educ.* 52, 234–246. doi: 10.1016/j.compedu.2008.08.003

Wenglein, S., Bauer, J., Heininger, S., and Prenzel, M. (2015). Kompetenz angehender Lehrkräfte zum Argumentieren mit Evidenz: Erhöht ein Training von Heuristiken die Argumentationsqualität. *Unterrichtswissenschaft* 43, 209–224.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/feduc.2022.976556

Teachers trust educational science - Especially if it confirms their beliefs

Kirstin Schmidt[1]*, Tom Rosman[2], Colin Cramer[3], Kris-Stephen Besa[4] and Samuel Merk[1]

[1]Faculty of Humanities and Human Science, Institute for School and Classroom Development, Karlsruhe University of Education, Karlsruhe, Germany, [2]Department for Research Literacy and User-Friendly Research Support, Leibniz Institute for Psychology, Trier, Germany, [3]Faculty of Economics and Social Sciences, Institute of Education, University of Tübingen, Tübingen, Germany, [4]Institute for Educational Science, Educational Science and Social Science, University of Münster, Münster, Germany

OPEN ACCESS

EDITED BY
Ingo Kollar,
University of Augsburg, Germany

REVIEWED BY
Georg Krammer,
University College of Teacher Education Styria, Austria
Leila Ferguson,
Kristiania University College, Norway

*CORRESPONDENCE
Kirstin Schmidt
kirstin.schmidt@ph-karlsruhe.de

SPECIALTY SECTION
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

RECEIVED 23 June 2022
ACCEPTED 07 November 2022
PUBLISHED 08 December 2022

CITATION
Schmidt K, Rosman T, Cramer C, Besa K-S and Merk S (2022) Teachers trust educational science - Especially if it confirms their beliefs.
Front. Educ. 7:976556.
doi: 10.3389/feduc.2022.976556

Teachers around the world are increasingly required by policy guidelines to inform their teaching practices with scientific evidence. However, due to the division of cognitive labor, teachers often cannot evaluate the veracity of such evidence first-hand, since they lack specific methodological skills, such as the ability to evaluate study designs. For this reason, second-hand evaluations come into play, during which individuals assess the credibility and trustworthiness of the person or other entity who conveys the evidence instead of evaluating the information itself. In doing so, teachers' belief systems (e.g., beliefs about the trustworthiness of different sources, about science in general, or about specific educational topics) can play a pivotal role. But judging evidence based on beliefs may also lead to distortions which, in turn, can result in barriers for evidence-informed school practice. One popular example is the so-called confirmation bias, that is, preferring belief-consistent and avoiding or questioning belief-inconsistent information. Therefore, we experimentally investigated (1) whether teachers trust knowledge claims made by other teachers and scientific studies differently, (2) whether there is an interplay between teachers' trust in these specific knowledge claims, their trust in educational science, and their global trust in science, and (3) whether their prior topic-specific beliefs influence trust ratings in the sense of a confirmation bias. In an incomplete rotated design with three preregistered hypotheses, $N = 414$ randomly and representative sampled in-service teachers from Germany indicated greater trust in scientific evidence (information provided by a scientific journal) compared to anecdotal evidence (information provided by another teacher on a teacher blog). In addition, we found a positive relationship between trust in educational science and trust in specific knowledge claims from educational science. Finally, participants also showed a substantial confirmation bias, as they trusted educational science claims more when these matched (rather than contradicted) their prior beliefs. Based on these results, the interplay of trust, first-hand evaluation, and evidence-informed school practice is discussed.

KEYWORDS

evidence-informed education, anecdotal evidence, confirmation bias, second-hand evaluation, scientific evidence, trust in science, teacher education

1. Introduction

Teachers can inform their professional practice using a vast number of information sources. To name just a few, they can refer to their own teaching experience, they can follow the advice of their colleagues, or they may refer to evidence obtained through educational science (Buehl and Fives, 2009). Around the world, policymakers (e.g., European Commission, 2007; Kultusministerkonferenz, 2014) as well as scientists (e.g., Bauer and Prenzel, 2012) increasingly value the latter and often consider it the most veracious body of knowledge because it is obtained through systematic and verifiable means (e.g., Williams and Coles, 2007; Bauer and Prenzel, 2012; Bauer et al., 2015; Brown et al., 2017). Such so-called evidence-informed school practice is considered to have large potential to improve school and teaching quality as well as student learning, for example by helping teachers to (a) make or change decisions pertaining to their teaching practice, (b) develop new practices, (c) inform leadership (Brown, 2020; Brown et al., 2022), or (d) effectively deal with problems that repeatedly come up in practice (Stark, 2017; Kiemer and Kollar, 2021). Empirical studies provide initial evidence for these considerations by showing that scientific evidence can actually inform teachers' practical decisions and actions (Cain, 2015), and that evidence-informed interventions, at least in specific contexts like formative assessment (Black and Wiliam, 2003) or mathematics learning (Doabler et al., 2014), can positively affect student achievement.

However, transforming scientific evidence into educational practice can be a challenging endeavor. Educational stakeholders increasingly take the position that scientific evidence can only *enrich* teachers' practical experience and contextual knowledge (Brown et al., 2017; Kiemer and Kollar, 2021). Thus, scientific evidence may not provide teachers with action-guiding recipes that can be directly used in everyday practice. Rather, it can be used to stimulate, reflect, or revise professional actions and decisions, which additionally makes them more transparent and objective (e.g., Bauer et al., 2015; Brown et al., 2017). In addition, considering evidence enables teachers to rationally justify their decisions and actions and consequently make them explicit—even to others, such as colleagues or parents (Bauer et al., 2015). The complexity of this endeavor is also reflected in the implementation steps of evidence-informed school practice. Evidence-informed practices require educational research literacy (e.g., Shank and Brown, 2007) which not only includes (1) accessing, (2) comprehending, and (3) critically reflecting the scientific evidence (e.g., its validity), but also (4) combining this evidence with prior knowledge before (5) using the evidence in practice (Shank and Brown, 2007; Brown et al., 2022).

This sophisticated theoretical view on teachers' professional use of scientific evidence contradicts reality: Even though policy guidelines all over the world (e.g., Kultusministerkonferenz, 2014 in Germany) emphasize the importance of evidence-informed school practice, it becomes apparent that teachers often do not follow such guidelines (e.g., Brown et al., 2017; Hinzke et al., 2020). Hence, the question arises what barriers teachers might face when it comes to realizing evidence-informed practice.

Previous studies identified various barriers ranging from lack of time to engage with evidence on top of other professional tasks, belief systems that devalue the importance, practicality, and usability of scientific evidence for teaching practice, as well as a lack of skills or knowledge to evaluate scientific evidence (see e.g., Gitlin et al., 1999; Williams and Coles, 2007; Thomm et al., 2021b). Thereby, different approaches exist that attempt to systematize these different barriers by using categorizations that range from rather broad distinctions between knowledge- and motivation-related barriers (e.g., Kiemer and Kollar, 2021) to more nuanced ones (e.g., van Schaik et al., 2018; Brown et al., 2022). While van Schaik et al. (2018), for example, differentiate between research knowledge level, individual teacher level, school organizational level, and communication level to systematize barriers, Brown et al. (2022) use the categories benefit, cost, and signification.

Specifically referring to knowledge-related barriers, we argue that these are often associated with the structure of modern societies, more precisely with the cognitive division of labor (Bromme et al., 2015). In general, the cognitive division of labor is defined as the uneven distribution and use of knowledge due to training of highly specialized experts (Bromme et al., 2010). This leads to the fact that teachers, in most countries, are trained as experts in education and learning, but not as educational scientists. Hence, student teachers, for example, have lower abilities in science-related areas compared to other students (Besa et al., 2020; Thiem et al., 2020) and are hardly trained in informing practice with scientific evidence (Ostinelli, 2009). Consequently, they often lack educational research literacy (Shank and Brown, 2007), making so-called first-hand evaluation of scientific evidence (i.e., assessing the veracity of scientific information by relying on objective criteria such as evaluating the study design) challenging for them (e.g., Bromme et al., 2010, 2015; Bromme and Goldman, 2014; Hendriks et al., 2015; Brown et al., 2017).

As a result, teachers often rely on so-called *second-hand evaluations*, which are defined as the assessment of the credibility and trustworthiness of the information's source (e.g., Bromme et al., 2010; Bromme and Goldman, 2014; Merk and Rosman, 2019). Hence, instead of analyzing whether the information itself is "true" or not, they evaluate if they can trust the information at hand or the person or body who conveys it (e.g., the researcher or specific science communication formats like clearing houses). Consequently, teachers' trust in educational science can be seen as a central predictor to which extent teachers positively evaluate and thus engage with scientific evidence (Hendriks et al., 2016;

Bromme et al., 2022). Hence, in the following, we analyze what factors influence teachers' trust in educational science. More specifically, we investigate whether teachers' trust in educational information is influenced by the source of information, their trust in educational science and in science in general, as well as by their prior topic-specific beliefs in the sense of a confirmation bias. While the first research question aims at aligning the study at hand with previous research on teachers' trust in science, the latter two extend existing research by transferring findings on teachers' beliefs from the domains of epistemic beliefs and cognitive biases to research on trust in science. This is to find out whether teachers' beliefs influence teachers' trust in educational science which, in turn, could also influence teachers' engagement with scientific evidence.

2. Trust-related barriers to evidence-informed practice

There are several definitions of trust (see Dietz and Den Hartog, 2006 for an overview), which often share the common overarching idea that there is a one-way dependency in which one party (the so-called trustor) trusts another party (the so-called trustee; Mayer et al., 1995; Dietz and Den Hartog, 2006; Blöbaum, 2016). Thereby, according to the popular definition of Mayer et al. (1995), the trustor agrees "*to be vulnerable to the actions of (...) [the other] party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" (p. 712).

Furthermore, there are different conceptualizations of trust such as trust as a belief, trust as an action, or trust as a decision (see again Dietz and Den Hartog, 2006 for an overview). In the following, we primarily focus on *trust as a belief*, i.e., the evaluation of trustworthiness of the potential trustee (Dietz and Den Hartog, 2006), which is why the question arises as to what causes trustworthiness in the first place. In this regard, three dimensions of trust are typically mentioned: (1) expertise, (2) integrity, and (3) benevolence. A party seems trustworthy if he/she is having high and relevant knowledge on the topic of interest (expertise), adheres to the rules, norms, and values of his or her profession (integrity) without ignoring the interest in the good of others (benevolence, e.g., Mayer et al., 1995; Dietz and Den Hartog, 2006; Hendriks et al., 2016; Bromme et al., 2022). These dimensions can be ascribed to characteristics of the trustee (Mayer et al., 1995; Dietz and Den Hartog, 2006; Blöbaum, 2016), whereby it depends on the trustors' perception how strong these characteristics are (Blöbaum, 2016). Characteristics of the trustor him- or herself, such as generalized trust in institutions or other beliefs, can also influence the degree of trusting a party (Mayer et al., 1995; Dietz and Den Hartog, 2006; Blöbaum, 2016).

Based on this, we argue that evaluating trustworthiness is accompanied by subjective and, therefore, less generalizable criteria, which could be prone to errors based on trustors' belief systems. Hence, we focus on three characteristics of teachers as trustors—(1) trust in different sources, (2) trust in educational science and science in general, as well as (3) prior individual beliefs about specific educational topics. While the former two are more directly related to second-hand evaluations, we argue that the latter (prior beliefs) are also important in this context since they might influence teachers' trust in information provided by educational science and therefore could act as (rather indirect) barriers to evidence-informed practice.

2.1. Trust in different knowledge sources

As outlined in the introduction, teachers can use a variety of sources to inform their practice. This ranges from anecdotal evidence (also called experiential sources, Bråten and Ferguson, 2015) that includes empirical information based on, for example, own personal experiences or experiences from a colleague (Buehl and Fives, 2009; Kiemer and Kollar, 2021) to more formalized sources such as lectures and formalized bodies of knowledge like scientific evidence (i.e., research findings, Buehl and Fives, 2009). The presentation or communication form of information is independent of the source. For example, anecdotal but also scientific evidence can be communicated both orally and in writing. Previous studies show that (future) teachers name and recognize the variety of information sources themselves (Buehl and Fives, 2009), but when it comes to practical decisions, they prefer anecdotal over scientific evidence (e.g., Gitlin et al., 1999; Parr and Timperley, 2008; Buehl and Fives, 2009; Cramer, 2013; Bråten and Ferguson, 2015; Zeuch and Souvignier, 2016; Menz et al., 2021; Groß Ophoff and Cramer, 2022). Furthermore, the preference of anecdotal evidence does not only seem to be associated with student teachers' motivation to learn in teacher training (Bråten and Ferguson, 2015), but it was also identified as the root of their beliefs and persistent misconceptions about specific topics from educational psychology (Menz et al., 2021).

In addition, the predominance of anecdotal evidence among (future) teachers is also evident in studies on trust in different sources: For example, Landrum et al. (2002) focused on an overall assessment of trust among student teachers by comparing, among others, the sources "scientific journals" vs. "teachers." In this descriptive study, the participants trusted information provided by other teachers more than information provided by scientific journals. In a two-step—first exploratory and then confirmatory—study by Merk and Rosman (2019) the results were more differentiated as student teachers deemed scientists in educational science as "smart but evil" (p. 6). Accordingly, they attributed more expertise but less benevolence and integrity to scientists than to practicing teachers. In a follow-up study, Rosman and Merk (2021) examined teachers' reasons for (dis-)trusting educational science vs. science in general. In line with the "smart but evil" pattern outlined above, teachers more strongly emphasized integrity and benevolence—compared to expertise—as reasons for distrusting educational scientists. Similarly, Hendriks et al. (2021) found differences in student teachers' reasons for (dis-)trusting information by educational psychology scientists or teachers. In their descriptive study, student teachers deemed educational psychology scientists not only less benevolent but also as having less expertise than teachers. When specifically looking for practical advice, student teachers rated teachers as more trustworthy than scientists—consistent across all three dimensions (expertise, integrity, benevolence).

To sum up, these results suggest that different knowledge sources influence (future) teachers' second-hand evaluation in a manner that anecdotal evidence provided by practicing teachers is perceived as more trustworthy than scientific evidence, leading to greater use of anecdotal evidence in practice. We argue that this preference may not only be caused by differences in the epistemological nature of anecdotal evidence (i.e., a non-scientific and possibly more "user-friendly" body of knowledge) compared to scientific evidence (which is often more abstract and theoretical) but also by the fact that individuals from one's own in-group are often evaluated more positively compared to individuals from out-groups since they share the same profession-related experiences (e.g., Mullen et al., 1992). Hence, educators who share anecdotal evidence from their day-to-day practice might seem more trustworthy to teachers compared to scientists. The preference for anecdotal evidence coming from other teachers could then act as a barrier to evidence-informed school practice, and, in the worst case, as Rosman and Merk (2021) argue, lead to dysfunctional practices when decisions are, for example, built on passed-on misconceptions like the prominent so-called neuromyths that can appear at an early career stage and are thus already common among student teachers (Krammer et al., 2019, 2021). There are many neuromyths with the learning style myth as a well-known example that has been debunked years ago (e.g., Pashler et al., 2008), but is still very popular among practitioners (Krammer et al., 2019, 2021). If (student) teachers do not consult scientific evidence, in light of this myth "they may waste time developing teaching materials tailored to individual students' learning styles" (Rosman and Merk, 2021, p. 1). Hence, we will analyze whether teachers trust knowledge claims by other teachers and scientific studies differently. To test this research question, we formulate the following first hypothesis:

> *H*1: *When teachers are confronted with knowledge claims regarding specific topics from educational science, they show more trust in claims if these are allegedly from another teacher (anecdotal evidence) than from a scientific study (scientific evidence).*

2.2. Trust in (educational) science

In the context of teachers' evaluation of scientific information, epistemic beliefs also play a pivotal role (Bendixen and Feucht, 2010; Fives and Buehl, 2010). Epistemic beliefs are defined as "individuals' beliefs about the nature of knowledge and the process of knowing" (Muis et al., 2016, p. 331). These beliefs can be simultaneously domain-general and domain-specific, whereby beliefs relating to different domains can influence each other (e.g., Buehl and Alexander, 2001; Muis, 2004). Hence, Muis et al. (2006) proposed the Theory of Integrated Domains in Epistemology (TIDE) that refers to the interplay of general epistemic beliefs, academic epistemic beliefs as well as domain-specific epistemic beliefs, which, in turn, are influenced by different contextual factors (i.e., the socio-cultural, academic, and instructional context). In their framework, the authors define general epistemic beliefs as "beliefs about knowledge and knowing that develop in nonacademic contexts such as the home environment, in interactions with peers, in work-related environments, and in any other nonacademic environments" (Muis et al., 2006, p. 33). Academic epistemic beliefs, on the other hand, encompass "beliefs about knowledge and knowing that begin to develop once individuals enter an educational system" (Muis et al., 2006, p. 35). Furthermore, these two belief dimensions can be differentiated from domain-specific epistemic beliefs—"beliefs about knowledge and knowing that can be articulated in reference to any domain to which students have been exposed" (Muis et al., 2006, p. 36). In 2018, Merk et al. (2018) extended the framework by adding topic-specific beliefs, i.e., beliefs regarding specific topics or theories, as a further dimension of epistemic beliefs. By analyzing student teachers' epistemic beliefs according to different educational topics, they found (at least on a correlational level) empirical support for the predictions of their framework that (1) topic-specificity is a feature of epistemic beliefs and that (2) domain-specific beliefs and topic-specific beliefs influence each other reciprocally.

Transferring the predictions of the TIDE framework to teachers' evaluation of trustworthiness, we posit that low trust in educational science and in science in general could be a barrier to evidence-informed practice via its effects on topic-specific trust. In fact, following the framework's assumption of different levels of beliefs reciprocally influencing each other, scientific findings about specific educational topics would also be deemed as less trustworthy and, therefore, less relevant for teaching practice in teachers with low trust in educational science and in science in general. In other words, we investigate whether there is an interplay between teachers' trust in specific knowledge claims from educational science, their trust in educational science, and their global trust in science. More specifically, we formulate the following second hypothesis:

H2: Trust in specific knowledge claims from educational science can be predicted by (domain-specific) trust in educational science and global trust in science.

2.3. Confirmation bias

The last factor considered in the present article involves teachers' prior beliefs about the corresponding knowledge claims. An extensive body of studies has shown that individuals' prior beliefs on a specific topic (e.g., on psychological or political issues) influence the search for and interpretation of information on this topic. In the literature, this phenomenon is labeled with different terminologies such as prior attitude effect (Druckman and McGrath, 2019), biased assimilation (Lord et al., 1979; Lord et al., 1984), or congeniality bias (Hart et al., 2009), but the most prominent one is *confirmation bias* (e.g., Lord et al., 1979; Nickerson, 1998; Oswald and Grosjean, 2004). Confirmation bias can be divided into two subcomponents: selective exposure and selective judgment. Selective exposure comes into play while seeking information on the respective topic, and manifests itself in preferring belief-consistent and ignoring belief-inconsistent information (Hart et al., 2009; Stroud, 2017). Selective judgment, on the other hand, is defined as the process of interpreting information in a way that—irrespective of the veracity of that information—this information is preferred if it is consistent with one's prior belief. By contrast, if this information is in conflict with one's prior belief, it is either quickly discounted or analyzed thoroughly to identify errors in it (e.g., Lord et al., 1979; Nickerson, 1998; Jonas et al., 2001; Oswald and Grosjean, 2004; Stroud, 2017). However, these two subcomponents are not always clearly differentiated and sometimes confirmation bias is referred to even when only one of the two subcomponents is considered (e.g., Butzer, 2020).

If teachers are subject to confirmation bias, this can act as a barrier to evidence-informed practice: On the one hand, teachers may completely distrust scientific evidence on a specific educational topic if the evidence is contradictory to their previous beliefs, resulting in ignoring the scientific evidence in their practical actions. On the other hand, they may selectively trust scientific evidence that is in line with their beliefs, and thus, selectively use scientific evidence. Even though confirmation bias is already discussed as a barrier to evidence-informed practice (e.g., Katz and Dack, 2014; Andersen, 2020), it has been less systematically analyzed in this context so far. While there are a few qualitative studies that draw attention to the existence of confirmation bias among teachers in the context of data-based decision making (e.g., Van Lommel et al., 2017; Andersen, 2020), only Masnick and Zimmerman (2009) have explicitly analyzed whether confirmation bias—or more precisely selective judgment—influences, among others, student teachers' evaluation of scientific evidence. In line with selective judgment, participants perceived the arrangement of the study as more appropriate as well as the results as more important and interesting when the results were in line with their beliefs compared to when they contradicted them.

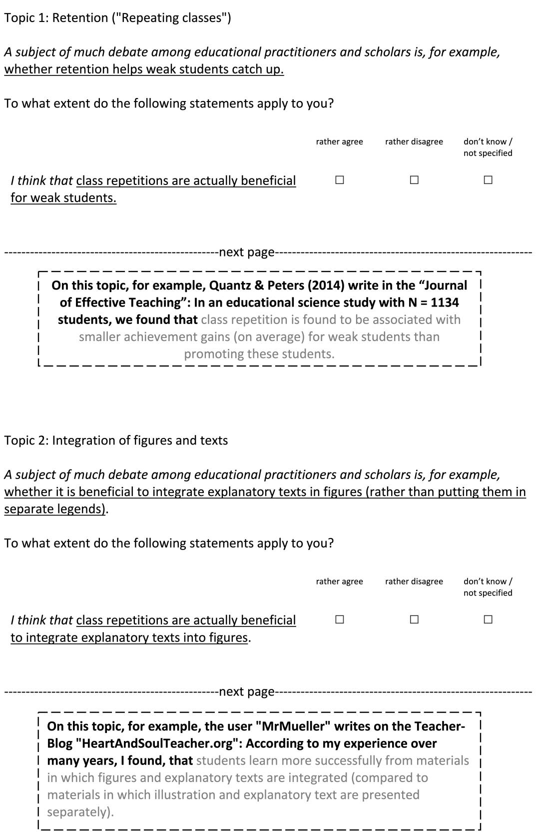
It should be noted that Masnick and Zimmerman (2009)'s dependent variables conformed to a first-hand evaluation of the study in question (e.g., evaluating the appropriateness of a study's design). However, despite the importance of second-hand evaluation in teachers' dealing with scientific evidence (see section 1), the influence of confirmation bias on such second-hand evaluations has, to the best of our knowledge, not been investigated so far. Consequently, we focus on the subcomponent selective judgment by examining whether teachers' prior topic-specific beliefs influence trust ratings in the sense of a confirmation bias. In doing so, we test the last hypothesis:

H3: When teachers are confronted with evidence for specific knowledge claims of educational science, they show more trust in these claims if these are belief-consistent.

3. Materials and methods

3.1. Experimental design

All hypotheses were preregistered (Schmidt et al., 2022). To test the hypotheses, we designed a 2 × 2 within-person experiment with the two independent variables *source of evidence* (scientific evidence from a published scientific study vs. anecdotal evidence from another teacher on a teacher blog) and *belief-consistency of evidence* (belief-consistent vs. belief-inconsistent claims). We thereby constructed texts about four topics from educational science (effects of retention, gender differences in grades, text-picture integration, signaling) for each source of evidence (i.e., scientific vs. anecdotal evidence), and with two variations of the claims made in texts (e.g., integrating text into pictures positively affects learning vs. does not positively affect learning) to allow us to manipulate belief-consistency. This resulted in 4 * 2 * 2 = 16 texts in total. These texts were mostly comparable in structure and wording and included a specific knowledge claim regarding the respective educational topic. For each topic, the only information that varied in the texts were the source of evidence and the belief-consistency of the presented knowledge claim (see Figure 1 for epitomes). To prevent respondent fatigue (Lavrakas, 2008) and unintentional unblinding, the participants, however, received only two out of these 16 texts (see below for details). All texts and all other study materials are publicly available (Schmidt et al., 2022).



Topic 1: Retention ("Repeating classes")

A subject of much debate among educational practitioners and scholars is, for example, whether retention helps weak students catch up.

To what extent do the following statements apply to you?

	rather agree	rather disagree	don't know / not specified
I think that class repetitions are actually beneficial for weak students.	☐	☐	☐

---next page---

On this topic, for example, Quantz & Peters (2014) write in the "Journal of Effective Teaching": In an educational science study with N = 1134 students, we found that class repetition is found to be associated with smaller achievement gains (on average) for weak students than promoting these students.

Topic 2: Integration of figures and texts

A subject of much debate among educational practitioners and scholars is, for example, whether it is beneficial to integrate explanatory texts in figures (rather than putting them in separate legends).

To what extent do the following statements apply to you?

	rather agree	rather disagree	don't know / not specified
I think that class repetitions are actually beneficial to integrate explanatory texts into figures.	☐	☐	☐

---next page---

On this topic, for example, the user "MrMueller" writes on the Teacher-Blog "HeartAndSoulTeacher.org": According to my experience over many years, I found, that students learn more successfully from materials in which figures and explanatory texts are integrated (compared to materials in which illustration and explanatory text are presented separately).

FIGURE 1
Epitomes of the 16 texts. Italic: Same for all sources, beliefs, and topics; Underscored: Same for both sources and beliefs within each topic; Bold: Same for every topic and both beliefs within both sources; Gray: Same for both sources within every topic and belief combination.

3.2. Procedure and measurements

At the beginning of the experiment, trust in science and trust in educational science (independent variable *H*2) were measured by self-assessment using items from the science barometer (Weißkopf et al., 2019). This large trend study uses the following item to assess trust in science: *"How much do you trust in science and research?"* along with a five-point Likert-scale (1 = *trust completely* to 5 = *don't trust*; with a *don't-know/not specified*-option). Additionally, we used the same item stem to assess trust in "educational science and educational research" (independent variable *H*2). To ensure that all participants conceptualize educational science and educational research in a similar way, we provided a brief explanation of educational science and educational research by defining it as an area *"[...] that deals with the theory and practice of education."* We further provided examples of subdisciplines of educational science and educational research, which *"are, among others, educational science, educational psychology, economics of education as well as sociology of education."* Subsequently, participants were introduced to a first (randomly chosen) topic (e.g., text-picture integration), and their belief toward this topic (auxiliary variable to construct the independent variable for *H*3) was assessed using a trichotomous item (e.g., *"I think it is actually beneficial to integrate text into pictures"*; with the answer options *"rather agree," "rather disagree,"* and *"don't know/not specified"*). Thereafter, they were presented with the randomly chosen text that included a knowledge claim referring to this topic using evidence from a randomly chosen source (independent variable *H*1; e.g., scientific evidence from a published scientific study) and of randomly chosen belief-consistency (e.g., consistent; independent variable *H*3). Referring to the latter, the knowledge claim presented was dependent on the previously stated belief of the participants and thus not *per se* an inconsistent or consistent claim (e.g., if the participant believes that integrating text into pictures is beneficial, the consistent knowledge claim informs about the benefit of integrating text into pictures). Finally, respondents were prompted to rate their trust in the respective knowledge claim (dependent variable for all three hypotheses) using an adapted item from the science barometer (Bromme et al., 2022; e.g., *"How much do you trust the claims of the educational scientists Quantz & Peters on the topic of text-picture integration?"*) on a five-point Likert-scale again ranging from 1 = *trust completely* to 5 = *don't trust* with a *don't-know/not specified*-option. Subsequently, this procedure was repeated for the second randomly chosen topic. Thereby, participants were subjected to—compared to the first round—the opposite condition of source and belief-consistency of evidence. For example, if scientific evidence including a consistent knowledge claim was presented in the first round (by random drawing, see above), the second knowledge claim referred to anecdotal evidence that was contradictory to the participants stated belief.

3.3. Sample

To achieve external validity, we commissioned a service provider to recruit a representative sample of in-service teachers in Germany. This was achieved using both random digit dialing (Wolter et al., 2009) and *post hoc* inverse-probability weighting (Mansournia and Altman, 2016). To avoid Type II errors, we conducted power analyses assuming random intercept regression models with level-1 dummy variables (*H*1 and *H*3) or a continuous level-2 predictor (*H*2) and small to moderate effect sizes (see Preregistration, Schmidt et al., 2022). Specifying a sample size of $N = 400$ participants, this resulted in good power estimates (>90%) for both model types. Correspondingly, the field provider stopped sending invitations and reminders after a sample size of $N = 400$ was reached, thus resulting in $N = 414$ participants. The distribution of our weighted sample and the corresponding population are given in Table 1.

TABLE 1 Sample characteristics in the population [Statistisches Bundesamt (Destatis), 2019] and the weighted sample in percentage.

	Population	Empirical
School type		
Elementary School (the German Grundschule)	29.3	29.3
Lower Secondary School (the German Hauptschule)	4.1	4.1
Intermediate Secondary School (the German Realschule)	8.0	8.0
Upper Secondary School (the German Gymnasium)	25.8	25.8
Comprehensive School (the German Gesamtschule)	19.9	20.0
Special School (the German Förderschule)	10.0	10.0
Age		
Age < 30	7.3	7.4
30 ≤ Age < 59	80.0	80.8
60 ≤ Age	12.4	11.8
Gender		
Men	26.9	26.9
Women	73.1	73.1

3.4. Statistical analysis

As preregistered, we used multi-level linear models (Gelman and Hill, 2007) to investigate the effects of the two experimentally manipulated independent variables (*source of evidence* and the *belief-consistency of evidence*; see Schmidt et al., 2022). As we did not preregister a detailed analysis plan, we did not include information on how to handle missing values. However, as our data contained a nontrivial amount of missing data after recoding the answer options "don't know/not specified" as missings (Lüdtke et al., 2007; 2% overall and 15% in the dependent variable), we decided to multiply impute these missing values using chained equations (van Buuren, 2018) and handle the multi-level structure of the data within these imputations using a dummy indicator approach (Lüdtke et al., 2017).

After carefully checking the imputation chains and the distributions of the imputed values, we estimated Bayesian random intercept models with flat priors for the regressions weights using the R package brms (Bürkner, 2017), which is based on the probabilistic programming language Stan (Stan Development Team, 2017). This package allows to incorporate survey weights by different contributions of data points to the likelihood, and also has built-in capacity for dealing with multiply imputed data: Distinct models are fitted for each imputed data set, resulting in as many models as imputations. While combining these models (model pooling) is a complex task in classical statistics (Rubin, 1976), it is straightforward after Bayesian estimation: One has just to join the posterior draws of the submodels (Zhou and Reiter, 2010). Furthermore, for a better interpretation, we have standardized the dependent variables and continuous predictors.

To evaluate not only the predictors but also the whole regression models, we estimated *Conditional* R^2 (Gelman et al., 2019) and compared, in cases where highest density intervals (HDIs) of the predictors indicate evidence for the null-hypothesis or negligible to small effects, the predictive performance of the models using Bayes factors based on bridge sampling (Gronau et al., 2017).

4. Results

All detailed results can be retrieved from the publicly available Reproducible Documentation of Analysis (Schmidt et al., 2022). To gain first insights into the results of our experiment, we plotted weighted means and standard deviations of the dependent variable trust by source of evidence, belief-consistency of evidence, and topic in Figure 2. These descriptive statistics imply (descriptive) evidence against $H1$, namely that teachers show more trust in claims regarding specific topics if these are allegedly from another teacher (anecdotal evidence) than from a scientific study (scientific evidence). In fact, all means of the trust variables were higher for the scientific study source, regardless of the respective topic and belief-consistency combination. Overall, this effect showed a large Cohen's $d = -0.81$, which varied substantially over the topics (gender differences in grades: $d = -1.25$, text-picture integration: $d = -1.04$, effects on retention: $d = -0.26$, signaling: $d = -0.86$).

Hypothesis 2 can be evaluated descriptively using Figure 3. As the bubble sizes in Figure 3 are proportional to the number of observations for each plot, one can see that there are some relations between trust in knowledge claims from educational science and trust in educational science (domain specific trust) as well as between the latter and trust in science (global trust). In contrast, trust in science seems to be uncorrelated with trust in knowledge claims from educational science. When only educational knowledge claims stemming from scientists, i.e., scientific evidence, are considered trust in science seems to be correlated with trust in knowledge claims from educational science, too. Furthermore, this visual impression is also reflected by descriptive correlation measures (see Pearson's r and Kendall's τ in Figure 3).

Finally, Figure 2 points toward a verification of Hypothesis 3, which posits that teachers trust claims from educational science more if these are belief-consistent. In fact, for each topic and source combination, participants consistently reported greater trust in the source if the respective claim was belief-consistent (overall: $d = 0.40$, gender differences in grades: $d = 0.43$, text-picture integration: $d = 0.28$, effects on retention:

$d = 0.64$, signaling: $d = 0.25$). Details on participants' beliefs about each topic can be found in Table 2.

To back up these descriptive results with inferential statistics, we estimated a series of Bayesian random intercept models. We started with a model including only the random intercept and dummy variables indicating the topic of the texts (see Table 3). These two predictors resulted in a *Conditional* $R^2 = 0.23$, whereas the addition of a dichotomous coded variable indicating the source (Model 2; referring to *H*1) resulted in an unstandardized coefficient of $b = 0.59$ (95%HDI: [0.48; 0.70]) and an increase of *Conditional* R^2 to 0.38. In the next two models, we consecutively included the continuous predictors trust in educational science and trust in science to test *H*2. The HDI of the standardized slope of trust in educational science did not contain zero (Model 3), whereas the HDI of the standardized slope for trust in science did so (Model 4). Both predictors explained comparatively little additional variance (*Conditional* $R^2 = 0.39$ for Model 3 and *Conditional* $R^2 = 0.39$ for Model 4). Exploratory (deviating from the preregistration) computed Bayes factors comparing Model 2 with Model 3 and Model 3 with Model 4 revealed some further evidence for this interpretation ($BF_{23} = 0.001$ and $BF_{34} = 4.86$). To test whether the findings in Model 4 on trust in science are potentially confounded by trust in educational science, we specified an additional model (Model 3a) that only included trust in science as a predictor. This exploratory (deviating from the preregistation) analysis revealed that including only trust in science as a predictor again explained comparatively little additional variance (*Conditional* $R^2 = 0.39$). The HDI of the standardized coefficient for trust in science was narrow (95%HDI: [0.00; 0.16]), which can be interpreted as evidence for a negligible effect. In Model 5, we included the third experimentally varied dichotomous variable, belief-consistency, as a dichotomous indicator (referring to *H*3). The

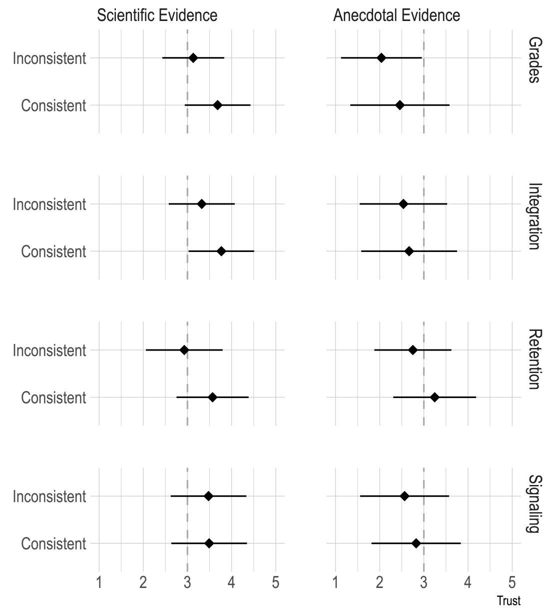


FIGURE 2
Weighted means and standard deviations of trust by belief-consistency of evidence, source of evidence, and topic.

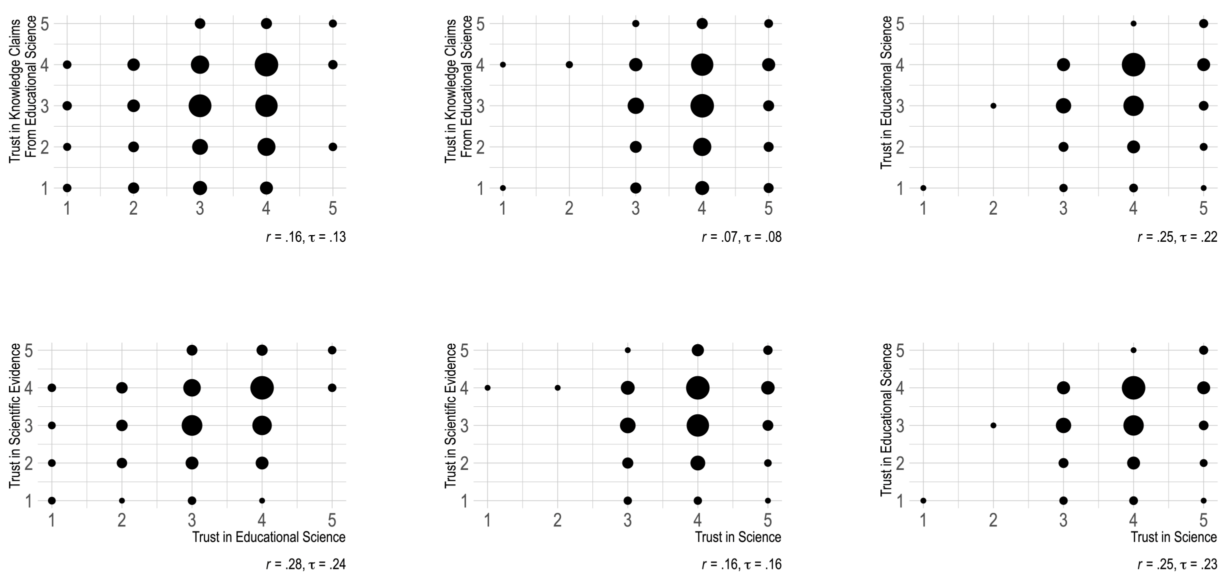


FIGURE 3
Associations of trust in knowledge claims from educational science respectively scientific evidence, educational science and science in general. Bubble plots. The expression *trust in knowledge claims from educational science* refers to trust in the knowledge claims across both sources of evidence (anecdotal and scientific evidence), whereas *trust in scientific evidence* only refers to the source scientific evidence. The bubble areas are proportional to the number of observation for each subplot. *r*, Pearson's *r*; *τ*, Kendall's *τ*.

corresponding unstandardized regression coefficient was $b = -0.35$ (95%HDI: $[-0.47; -0.23]$) and *Conditional* R^2 increased again to 0.41. In Model 5a, we analyzed the effect of belief-consistency specifically on scientific evidence, which resulted in an unstandardized regression coefficient of $b = -0.38$ (95%HDI: $[-0.55; -0.21]$) and an *Conditional* R^2 of 0.91. To further explore (a further deviation from the preregistration) the influence of trust in educational science, trust in science, and belief-consistency specifically on scientific evidence, we estimated an interaction model (Model 6) to see whether the source of evidence moderates the association between trust in educational science, respectively, trust in science and trust in knowledge claims from educational science. The HDI of the standardized coefficient of the interaction term of trust in educational science and source ($\beta = 0.11$; 95%HDI: $[-0.00; 0.22]$) as well as the HDI of the standardized coefficient for the interaction term of trust in science and source of evidence ($\beta = 0.11$; 95%HDI: $[0.00; 0.22]$) were adjacent to zero which can be interpreted as evidence for a small moderation effect of the source. However, Bayes factors comparing Model 5 with Model 6 revealed evidence in favor of Model 5 ($BF_{56} = 0.33$).

5. Discussion

In the present study, we focused on trust-related barriers teachers might face when it comes to realizing evidence-informed practice. As teachers often evaluate scientific evidence from educational science by evaluating whether the person or body who conveys the evidence appears trustworthy, we analyzed factors that might influence such second-hand evaluations. We thereby examined (1) whether teachers trust knowledge claims by other teachers and scientists differently, (2) whether there is an interplay between teachers' trust in specific knowledge claims from educational science, their trust in educational science, and their global trust in science, as well as (3) whether teachers' prior topic-specific beliefs influence their trust ratings in the sense of a confirmation bias.

5.1. Summary and discussion of main results

A key finding of our study is that teachers consider knowledge claims made by educational scientists (on average) as more trustworthy than those made by other teachers. This result is particularly surprising because it contradicts numerous previous research results on the preference of anecdotal over scientific evidence (e.g., Landrum et al., 2002; Merk and Rosman, 2019; Hendriks et al., 2021; Rosman and Merk, 2021). In the following, we want to discuss possible reasons for this deviation:

One explanation could be attributed to the fact that previous studies mainly used student teacher samples (Landrum et al., 2002; Merk and Rosman, 2019; Hendriks et al., 2021). As we have surveyed in-service teachers, the deviation might be (partially) due to the different samples' characteristics such as differences in working or practical experiences. By providing evidence that college experiences influence student teachers' beliefs about knowledge sources, a study by Perry (1999) supports this idea.

Differences in the operationalizations of dependent and independent variables between previous studies and our study can offer further explanations. Referring to the dependent variable "trust," previous studies primarily measured dimensions of trust (expertise, integrity, and benevolence) rather than an overall assessment of (dis-)trust in science. In these studies, trust ratings regarding science and scientists in general were lower than those regarding teachers—especially in terms of integrity and benevolence—but globally still rather high (Merk and Rosman, 2019; Hendriks et al., 2021; Rosman and Merk, 2021). The operationalization of the independent variable "source" also deviates from previous operationalizations. While previous studies focused on trust in the source itself without referring to any content-related information (Landrum et al., 2002; Hendriks et al., 2021; Rosman and Merk, 2021), we focused on trust in specific *claims made by a specific source*. The inconsistencies of our findings regarding prior research might thus be caused by the fact that we kept all knowledge claims mostly constant over the different sources, whereas, in other studies, there is a potential confounding between source and content. Asking participants about different sources without referring to specific topics may introduce bias since certain knowledge sources are usually associated with specific topics. To give an example, if teachers are asked about their trust in the expertise of teaching practitioners, they might think of expertise in classroom teaching, whereas, when asked the same question with reference to educational scientists, they might consider researchers' expertise in conducting empirical studies (Merk and Rosman, 2019). In this case, there is thus no consistent criterion for comparison. Hence, the finding, in previous studies, that teachers are often reluctant to use scientific knowledge for their day-to-day practice may not result from the fact that they generally doubt that science is able to deliver robust and trustworthy knowledge, but possibly due to the fact that scientific evidence is often associated with more abstract and theoretical information (e.g., Buehl and Fives, 2009; Bråten and Ferguson, 2015; Groß Ophoff and Cramer, 2022).

Furthermore, our results provide early evidence against an influence of an in-group bias (e.g., Mullen et al., 1992) on teachers' trust. The knowledge claims provided by other teachers, i.e., by an in-group member, were not perceived as more trustworthy than those by scientists—out-group members. However, this statement should be taken with caution. Although we kept the knowledge claims and wording mostly constant, the source "teacher" referred to some other (unknown) teacher who published own experiences on a teacher blog and the source "scientists" to scientists who published their findings

TABLE 2 Beliefs per topic in absolute frequencies.

Topic: Statement	Frequencies		
	Rather agree	Rather disagree	Don't know/Not specified
Gender differences in grades: "I think girls actually get better grades."	92	107	12
Text-picture-integration: "I think it is actually beneficial to integrate text into pictures."	139	39	27
Effects of retention: "I think class retention is actually beneficial for weak students."	151	58	11
Signaling: "I think it is actually beneficial to signal central information in texts."	138	43	11

in a scientific journal. Hence, the publication body (teacher blog vs. scientific journal) might also have influenced teachers' trust ratings. As Bråten and Ferguson (2015) found that student teachers believed least in knowledge stemming from social and popular media compared to anecdotal and formalized knowledge, blog posts, as a form of social media, could have decreased teachers' trust in the provided claims by other teachers. One possible explanation for the lower trust in anecdotal evidence on teacher blogs could be that anyone could write such posts, which can reduce its seriousness and therefore its trust in it. Another factor that might influence teachers' trust ratings in anecdotal evidence could be related to the familiarity of the source. Teachers might perceive a claim from a trusted colleague as more trustworthy compared to an anonymous blog-poster. Given that most previous studies investigating (student) teachers' trust in different sources also did not explicitly take source familiarity into account, but rather used generic terms such as an "experienced teacher" (Landrum et al., 2002, p. 44), a "practitioner" (Merk and Rosman, 2019, p. 4), or "a teacher who has taught at a school for a number of years" (Hendriks et al., 2021, p. 170), source familiarity might be an interesting independent variable for further research.

All in all, however, this finding is good news as a general lack of trust in statements from educational science does not seem to be a barrier to evidence-informed action in schools. But at the same time, and in line with previous research, we found strong support for a substantial confirmation bias in teachers' trust ratings. Confirmation bias cannot only come into play while evaluating scientific information first-hand (e.g., Masnick and Zimmerman, 2009), but can also distort, as demonstrated in the present study, teachers' second-hand evaluations. Thus, it can be assumed that teachers evaluate knowledge claims to a large extent in such a way that they confirm their own prior beliefs. Consequently, they might be highly selective in choosing the evidence they refer to. The general finding that teachers have traditionally been rather reluctant to turn to scientific evidence and rely heavily on their professional autonomy in making decisions (e.g., Landrum et al., 2002; Buehl and Fives, 2009; Bråten and Ferguson, 2015; Groß Ophoff and Cramer, 2022) could, thus, be viewed in a differentiated way: It is not a general lack of trust in claims from educational science that might hinder teachers from engaging with scientific evidence, but the question of which filters come into play to evaluate the evidence. The present study revealed that confirmation bias might work as one such filter and, thus, as a potential barrier for evidence-informed practice as teachers might not easily change their practice in light of scientific evidence that does not fit their beliefs. Given that many students already enter teacher education with a specific set of misconceptions such as neuromyths (Krammer et al., 2019, 2021), our findings on the existence of a confirmation bias are particularly worrying, since confirmation bias may lead to a further strengthening of such misconceptions. Furthermore, if teachers trust evidence that is consistent with their beliefs more and rather distrust evidence that is contradictory, they will continue to teach as before. This seems rather unproblematic as long as the teachers' practice is tried and tested. In addition, it is quite unrealistic to expect teachers to inform every practical decision and action with evidence. However, it is, on the one hand, problematic when it comes to concepts that are scientifically untenable but continue to persist in practice, and, on the other hand, when evidence is used to develop new approaches or to overcome hitherto unsolved problems as this makes it difficult to stimulate new avenues. Therefore, approaches must be identified to specifically motivate teachers to change their practice when scientific evidence contradicts their beliefs and existing practice.

Furthermore, of course not all teachers trust claims from educational science in general equally. This can be explained partially by referring to some of the predictions of the TIDE framework (Muis et al., 2006). Admittedly, we found evidence for a negligible influence of global trust in science on topic-specific trust, a finding that might be caused by different associations between science and educational science. For example, the former might be associated with the mixing of different chemicals in a laboratory, whereas the latter might be perceived more as a process of developing abstract theories. Nevertheless, in line with the TIDE-framework, domain-specific trust in educational science predicted teachers' topic specific trust. As a consequence, those teachers who perceive educational science in general as less trustworthy also report less trust in specific claims from educational science, and thus might ignore scientific evidence as a source of information for their practical actions. Therefore, when striving to foster teachers'

10.3389/feduc.2022.976556

TABLE 3 Results of the Bayesian multi-level models.

	Trust															
	Model 1		Model 2		Model 3		Model 3a		Model 4		Model 5		Model 5a		Model 6	
	Est.	HDI (95%)	Est.	HDI (95%)	Est.	HDI (95%)	Est.	HDI (95%)	Est.	HDI (95%)	Est.	HDI (95%)	Est.	HDI (95%)	Est.	HDI (95%)
Intercept	−0.15	−0.29 to −0.02	−0.44	−0.58 to −0.31	−0.45	−0.59 to −0.31	−0.44	−0.58 to −0.30	−0.45	−0.58 to −0.31	−0.27	−0.41 to −0.12	0.46	0.27 to 0.65	−0.27	−0.41 to −0.12
$I_{Topic}^{Integration}$	0.20	0.02 to 0.38	0.19	0.03 to 0.36	0.20	0.03 to 0.36	0.19	0.03 to 0.36	0.20	0.03 to 0.36	0.18	0.01 to 0.34	0.19	−0.05 to 0.43	0.17	0.01 to 0.34
$I_{Topic}^{Retention}$	0.24	0.06 to 0.42	0.21	0.04 to 0.37	0.21	0.05 to 0.37	0.20	0.04 to 0.37	0.21	0.04 to 0.37	0.20	0.04 to 0.36	−0.09	−0.32 to 0.14	0.20	0.04 to 0.36
$I_{Topic}^{Signaling}$	0.12	−0.06 to 0.31	0.15	−0.02 to 0.32	0.16	−0.01 to 0.33	0.15	−0.02 to 0.32	0.16	−0.01 to 0.33	0.16	−0.01 to 0.32	0.01	−0.23 to 0.26	0.16	−0.01 to 0.32
$I_{Source}^{ScientificEvidence}$			0.59	0.48 to 0.70	0.59	0.49 to 0.70	0.59	0.48 to 0.70	0.59	0.48 to 0.70	0.60	0.49 to 0.71			0.60	0.49 to 0.70
Trust in Educational Science					0.15	0.08 to 0.23			0.14	0.06 to 0.22	0.15	0.07 to 0.22	0.21	0.12 to 0.29	0.09	−0.00 to 0.19
Trust in Science							0.08	0.00 to 0.16	0.05	−0.03 to 0.12	0.04	−0.04 to 0.12	0.09	−0.00 to 0.17	−0.02	−0.11 to 0.08
$I_{Belief\text{-}Consistency}^{Inconsistent}$											−0.35	−0.47 to −0.23	−0.38	−0.55 to −0.21	−0.35	−0.47 to −0.22
Trust in Educational Science*$I_{Source}^{ScientificEvidence}$															0.11	−0.00 to 0.22
$I_{Source}^{ScientificEvidence}$*Trust in Science															0.11	0.00 to 0.22
Random effects																
σ^2	0.78		0.62		0.62		0.62		0.62		0.60		0.64		0.59	
τ_{00}	0.22		0.29		0.28		0.29		0.28		0.26		0.19		0.27	
ICC	0.22		0.32		0.31		0.32		0.31		0.30		0.77		0.31	
Clusters	412		412		411		412		411		411		411		411	
Observations	824		824		822		824		822		822		411		822	
Marginal R^2/ *Conditional* R^2	0.01/0.23		0.10/0.38		0.12/0.39		0.10/0.39		0.13/0.39		0.16/0.41		0.14/0.91		0.16/0.42	

The dependent variables [Trust in Knowledge Claims From Educational Science (Model 1–Model 5 and Model 6) and Trust in Scientific Evidence (Model 5a)] and continuous predictors (Trust in Educational Science, Trust in Science) are standardized. Est., estimate of the regression coefficient; HDI (95%), 95% highest density interval; σ^2, Level-1 residual variance; τ_{00}, Level-2 residual variance; ICC, intraclass correlation coefficient; *Marginal* R^2, variance that is explained by fixed factors; *Conditional* R^2, variance that is explained by fixed and random factors.

evidence-informed practice, one promising way may either focus on increasing trust in educational science as a whole, or on increasing trust in specific knowledge claims from that domain. However, so far little is know about how to actually do this. But even if an effective way to increase trust can be identified, we want to point out that trust in science is only a predictor in the sense of a necessary condition for acting in an evidence-informed manner, but it does not automatically imply evidence-informed actions. The same also applies to first-hand evaluation: Even a competent first-hand evaluation of the veracity of scientific evidence (e.g., enabled by comprehensive teacher training including a fundamental training in methodology and well-designed science communication) does not automatically imply engagement with evidence. This is also illustrated by the implementation steps of evidence-informed practice: after evaluating scientific evidence, teachers still need to link the evidence to their own prior knowledge and then need to find ways to concretely use the evidence in their practice (Brown et al., 2022). Nevertheless, it can be assumed that a successful first-hand evaluation supports correct receptions of scientific evidence, which in turn is a central basis for an adequate transfer of scientific evidence into practice.

In addition, we want to highlight that processes in schools and classrooms are complex, characterized by interpersonal interaction, and exposed to uncertainty in the field of educational action, which can be reduced but never completely resolved by recourse to scientific evidence (Cochran-Smith et al., 2014). Thus, even substantial engagement with evidence does not grasp the action situation in its whole complexity. In other words, the assumption of simple and fitting evidence for certain or even all conceivable questions of school practice is neither tenable nor scientifically justifiable (Renkl, 2022). Evidence in itself can also be problematic if low-quality evidence is referred to, such as findings that cannot be replicated (e.g., Makel and Plucker, 2014; Gough, 2021). Even within science, there are critical voices concerning the informativeness of common research methods (e.g., randomized controlled trials) for educational practice, which is, after all, highly context-specific (e.g., Berliner, 2002). However, we argue that understanding evidence-informed practice as an educational practice where scientific evidence is reinterpreted against the background of one's own experience and the context at hand (Brown et al., 2017) may counteract this criticism. Consequently, teachers do not only need to be able to evaluate, understand, and deal with scientific evidence, but also to reflect it in light of other information sources, for example, their own practical experiences, contextual knowledge, and local school data (e.g., Bauer et al., 2015; Brown et al., 2017). Only then can scientific evidence unfold its potential to contribute to a broadening of perspectives on the pedagogical field of action.

5.2. Methodological limitations and future research

Of course, our study is not without limitations. Some have already been mentioned before. In the following, we will briefly repeat these limitations and add further ones, as well as derive some implications for future studies.

The results of the present study indicate—and this is contrary to previous findings—that teachers perceive scientific evidence as more trustworthy than anecdotal evidence from other teachers. As mentioned in the section above, the operationalization of trust either as a multi-dimensional construct (with the dimensions expertise, integrity, and benevolence) or as an overall rating (e.g., used in the science barometer, Weißkopf et al., 2019) as well as the operationalization of the source (with or without content-related information) could explain the difference. With regard to the latter, we used claims regarding specific topics from educational science to increase internal validity. This approach avoids that participants make different, and thus not comparable, associations when thinking about the expertise of practicing teachers or educational scientists (see section above). In contrast, the external validity of our study might be curtailed through this approach, which is why our results should not simply be generalized to other educational science topics.

Furthermore, as outlined in the section above, the additional information about the publication body (teacher blog vs. scientific journal) may have confounded teachers' trust ratings in different sources. The same could also apply to the names of the persons who provide the information (Quantz and Peters, 2014 vs. Mr. Mueller) as well as the name of the publisher (Journal of Effective Teaching vs. HeartAndSoulTeacher.org), given that textual features of "scientificness" have been shown to affect the processing and evaluation of textual information (Thomm and Bromme, 2012). However, to increase the external validity of our material, we tried to create scenarios which are as realistic as possible. As teachers often do not meet scientists in person, we decided to use a written format of scientific evidence. To increase internal validity, we kept our study materials as parallel as possible across the two different sources (i.e., scientific and anecdotal evidence), which is why we also chose a written format for anecdotal evidence. In this context, blogs are typical for informing others about personal experiences in a written manner (e.g., Ray and Hocutt, 2006; Deng and Yuen, 2011). However, our approach reflects the challenge of designing internally valid study materials quite well: The more information is given in study materials, the less abstract the information appears and the more varying associations of the participants can be prevented. At the same time, it cannot be ruled out that one of these pieces of contextual information (e.g., publication body) confounds the variables that are actually relevant for the study and introduces bias on the outcomes. Therefore, it might be reasonable to systematically vary the publication bodies in further studies.

Based on our findings, one might assume that teachers indeed have a high general trust in claims from educational science, which, however, might decrease when explicitly referring to concrete teaching practice. In other words, teachers might take the attitude that what scientists say is true and trustworthy, but has not much to do with their own teaching practice or with the issues they actually encounter in the classroom (e.g., Gitlin et al., 1999). This is a limitation of our study since we did not directly relate our study materials to school practice (or other contexts). However, we argue that these concerns are mitigated by the fact that our scientific and anecdotal evidence indirectly referred to school practice by presenting claims about an educational topic published in a teacher journal respectively on a teacher blog. Nevertheless, in future studies, the context in which the evidence is intended to be used could also be explicitly considered because Hendriks et al. (2021) found that this can lead to differences in trust ratings. In fact, in their study, student teachers perceived educational psychology researchers as more trustworthy than teachers when searching for theoretical explanations, but, in contrast, the source “teachers” was trusted more when it came to practical recommendations.

As a final limitation regarding our results on Hypothesis 1, we cannot completely exclude a Hawthorne effect or social desirability bias, meaning that trust in research could have been rated higher because participants were asked by researchers. Considering previous research, however, we see this influence as rather minor, as previous research studies have found greater trust in anecdotal evidence although their participants were asked by researchers, too.

Our results also show that selective trust in evidence may be fuelled by a confirmation bias. Recent studies illustrate that prior beliefs can even lead to the conclusion that when one is confronted with belief-inconsistent scientific evidence, certain topics cannot be scientifically investigated at all (Rosman et al., 2021; Thomm et al., 2021a). In addition, previous studies on confirmation bias indicate that the strength of prior beliefs moderates the influence of confirmation bias on searching and interpreting information (e.g., Taber and Lodge, 2006). Hence, in future studies, it would be reasonable to additionally collect data on the strength of teachers’ prior beliefs, given that our results might underestimate the influence of confirmation bias on trust in teachers with strong prior beliefs and overestimate the influence in teachers with less entrenched prior beliefs. In addition to prior beliefs, other individual characteristics of teachers, but also characteristics of the scientific evidence can activate or act as filters (Fives and Buehl, 2012). With respect to our study, teachers’ varying degrees of trust in the domain of educational science can be regarded as one such individual characteristic. However, factors like the individual degree of educational research literacy or epistemic beliefs could be influential as well. Referring to the former, teachers with higher educational research literacy might generally indicate a higher trust in educational science, and possibly show less of a confirmation bias. With regard to epistemic beliefs, teachers with high multiplistic epistemic beliefs (scientific knowledge as subjective “opinions”) might rate educational science as less trustworthy and be more inclined toward confirming their prior beliefs (Hofer and Pintrich, 1997). Referring to the characteristics of scientific evidence, future research should also examine (1) whether teachers trust certain research paradigms more than others (e.g., experimental vs. observational research), (2) if they trust scientific evidence more if it is proximal to their actual teaching practice, or (3) if they place higher trust into evidence that is close to their teaching subjects (e.g., because it is more familiar).

Finally, with regard to future research, we would like to emphasize that we have not measured teachers’ engagement with evidence or use of evidence. Even though trust is a necessary (but not sufficient) predictor for engaging with evidence, it would be reasonable to additionally focus on actions based on trust (i.e., behavioral variables). In this regard, however, the conceptualization of objective measures that are intended to go beyond self-reporting (e.g., How likely is it that you will incorporate the research findings into your own practice?) is quite complex.

6. Conclusion

Taken together, our findings allow a more differentiated view of teachers’ trust in educational science and, thus, of trust-related barriers teachers face when realizing evidence-informed practice: It is not a general lack of trust in science that might hinder teachers from engaging with (educational) scientific evidence, it is more about the filter function of beliefs that come into play to evaluate evidence that is problematic for an adequate realization of evidence-informed practice. Thereby, the present study revealed teachers’ prior topic-specific beliefs as one such filter since they trusted more in evidence consistent with their prior beliefs than belief-inconsistent evidence (confirmation bias). Such selective trust can be dangerous since teachers may—if at all—inform their actions with empirical evidence, but just one-sidedly and not in its full scope. Moreover, we argue that second-hand evaluation as well as first-hand evaluation are necessary conditions to engage with evidence, which is why both needs to be fostered systematically. Nevertheless, a successful first-hand combined with a positive second-hand evaluation is an important foundation for evidence-informed practice but does not automatically imply evidence-informed actions and even if professional actions are informed by evidence, scientific evidence can only unfold its potential when it is reflected in light of information from other sources.

Data availability statement

The datasets generated for this study can be found in the Open Science Framework-Repository-Teachers Trust Educational Science - Especially if it Confirms Their Beliefs at https://osf.io/jm4tx.

Ethics statement

The study involving human participants was reviewed and approved by University of Tübingen Institutional Review Board (Protocol Number: A2.5.4-094_aa). The participants provided their written informed consent to participate in this study.

Author contributions

KS and SM: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, supervision, validation, visualization, writing—original draft, review, and editing. TR: conceptualization, investigation, writing—original draft, review, and editing. CC: funding acquisition, writing—original draft, review, and editing. K-SB: writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and the Arts, and the Karlsruhe University of Education in the Funding Programme Open Access Publishing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andersen, I. G. (2020). What went wrong? Examining teachers' data use and instructional decision making through a bottom-up data intervention in Denmark. *Int. J. Educ. Res.* 102, 101585. doi: 10.1016/j.ijer.2020.101585

Bauer, J., and Prenzel, M. (2012). European teacher training reforms. *Science* 336, 1642–1643. doi: 10.1126/science.1218387

Bauer, J., Prenzel, M., and Renkl, A. (2015). Evidenzbasierte Praxis – im Lehrerberuf?! Einführung in den Thementeil (Evidence-based practice in teaching?! Introduction to the special issue). *Unterrichtswissenschaft* 43, 188–192. doi: 10.3262/UW1503188

Bendixen, L. D., and Feucht, F. C. (2010). "Personal epistemology in the classroom: What does research and theory tell us and where do we need to go next?," in *Personal Epistemology in the Classroom, 1st Edn.*, eds L. D. Bendixen and F. C. Feucht (Cambridge, UK: Cambridge University Press), 555–586. doi: 10.1017/CBO9780511691904.017

Berliner, D. C. (2002). Comment: Educational research: the hardest science of all. *Educ. Res.* 31, 18–20. doi: 10.3102/0013189X031008018

Besa, K.-S., Gesang, J., and Hinzke, J.-H. (2020). Zum Verhältnis von Forschungskompetenz und Unterrichtsplanung (On the relationship between research literacy and instructional design). *J. Allgem. Didak.* 10, 91–107.

Black, P., and Wiliam, D. (2003). 'In praise of educational research': formative assessment. *Brit. Educ. Res. J.* 29, 623–637. doi: 10.1080/0141192032000133721

Blöbaum, B. (2016). "Key factors in the process of trust. On the analysis of trust under digital conditions," in *Trust and Communication in a Digitized World*, ed B. Blöbaum (Cham: Springer International Publishing), 3–25. doi: 10.1007/978-3-319-28059-2_1

Bråten, I., and Ferguson, L. E. (2015). Beliefs about sources of knowledge predict motivation for learning in teacher education. *Teach. Teach. Educ.* 50, 13–23. doi: 10.1016/j.tate.2015.04.003

Bromme, R., and Goldman, S. R. (2014). The public's bounded understanding of science. *Educ. Psychol.* 49, 59–69. doi: 10.1080/00461520.2014.921572

Bromme, R., Kienhues, D., and Porsch, T. (2010). "Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) to be attained from others," in *Personal Epistemology in the Classroom*, eds L. D. Bendixen and F. C. Feucht (Cambridge: Cambridge University Press), 163–194. doi: 10.1017/CBO9780511691904.006

Bromme, R., Mede, N. G., Thomm, E., Kremer, B., and Ziegler, R. (2022). An anchor in troubled times: trust in science before and within the COVID-19 pandemic. *PLoS ONE* 17, e0262823. doi: 10.1371/journal.pone.0262823

Bromme, R., Thomm, E., and Wolf, V. (2015). From understanding to deference: laypersons' and medical students' views on conflicts within medicine. *Int. J. Sci. Educ. B* 5, 68–91. doi: 10.1080/21548455.2013.849017

Brown, C. (2020). *The Networked School Leader: How to Improve Teaching and Student Outcomes Using Learning Networks*. Bingley, UK: Emerald Publishing. doi: 10.1108/9781838677190

Brown, C., MacGregor, S., Flood, J., and Malin, J. (2022). Facilitating research-informed educational practice for inclusion. Survey findings from 147 teachers and school leaders in England. *Front. Educ.* 7, 890832. doi: 10.3389/feduc.2022.890832

Brown, C., Schildkamp, K., and Hubers, M. D. (2017). Combining the best of two worlds: a conceptual proposal for evidence-informed school improvement. *Educ. Res.* 59, 154–172. doi: 10.1080/00131881.2017.1304327

Buehl, M. M., and Alexander, P. A. (2001). Beliefs about academic knowledge. *Educ. Psychol. Rev.* 13, 385–418. doi: 10.1023/A:1011917914756

Buehl, M. M., and Fives, H. (2009). Exploring teachers' beliefs about teaching knowledge: Where does it come from? Does it change? *J. Exp. Educ.* 77, 367–407. doi: 10.3200/JEXE.77.4.367-408

Bürkner, P.-C. (2017). Brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Butzer, B. (2020). Bias in the evaluation of psychology studies: a comparison of parapsychology versus neuroscience. *EXPLORE* 16, 382–391. doi: 10.1016/j.explore.2019.12.010

Cain, T. (2015). Teachers' engagement with published research: addressing the knowledge problem. *Curric. J.* 26, 488–509. doi: 10.1080/09585176.2015.1020820

Cochran-Smith, M., Ell, F., Ludlow, L., Grudnoff, L., and Aitken, G. (2014). The challenge and promise of complexity theory for teacher education research. *Teach. Coll. Rec.* 116, 1–38. doi: 10.1177/016146811411600407

Cramer, C. (2013). Beurteilung des bildungswissenschaftlichen Studiums durch Lehramtsstudierende in der ersten Ausbildungsphase im Längsschnitt (The assessment of educational scientific study courses as given by students in their first phase of teacher training in longitudinal section). *Z. Pädagog.* 59, 66–82. doi: 10.25656/01:11927

Deng, L., and Yuen, A. H. K. (2011). Towards a framework for educational affordances of blogs. *Comput. Educ.* 56, 441–451. doi: 10.1016/j.compedu.2010.09.005

Dietz, G., and Den Hartog, D. N. (2006). Measuring trust inside organisations. *Pers. Rev.* 35, 557–588. doi: 10.1108/00483480610682299

Doabler, C. T., Nelson, N. J., Kosty, D. B., Fien, H., Baker, S. K., Smolkowski, K., et al. (2014). Examining teachers' use of evidence-based practices during core mathematics instruction. *Assess. Effect. Intervent.* 39, 99–111. doi: 10.1177/1534508413511848

Druckman, J. N., and McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nat. Clim. Change* 9, 111–119. doi: 10.1038/s41558-018-0360-1

European Commission (2007). *Towards More Knowledge-Based Policy and Practice in Education and Training.* Available online at: https://op.europa.eu/en/publication-detail/-/publication/962e3b89-c546-4680-ac84-777f8f10c590 (accessed November 11, 2022).

Fives, H., and Buehl, M. M. (2010). "Teachers' articulation of beliefs about teaching knowledge: Conceptualizing a belief framework," in *Personal Epistemology in the Classroom*, eds L. D. Bendixen and F. C. Feucht (Cambridge, UK: Cambridge University Press), 470–515. doi: 10.1017/CBO9780511691904.015

Fives, H., and Buehl, M. M. (2012). "Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us?," in *APA Educational Psychology Handbook, Vol 2: Individual Differences and Cultural and Contextual Factors*, eds K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, and M. Zeidner (Washington, DC: American Psychological Association), 471–499. doi: 10.1037/13274-019

Gelman, A., Goodrich, B., Gabry, J., and Vehtari, A. (2019). R-squared for Bayesian regression models. *Amer. Stat.* 73, 307–309. doi: 10.1080/00031305.2018.1549100

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Vol. 1. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511790942

Gitlin, A., Barlow, L., Burbank, M. D., Kauchak, D., and Stevens, T. (1999). Pre-service teachers' thinking on research: implications for inquiry oriented teacher education. *Teach. Teach. Educ.* 15, 753–769. doi: 10.1016/S0742-051X(99)00015-3

Gough, D. (2021). Appraising evidence claims. *Rev. Res. Educ.* 45, 1–26. doi: 10.3102/0091732X20985072

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., et al. (2017). A tutorial on bridge sampling. *J. Math. Psychol.* 81, 80–97. doi: 10.1016/j.jmp.2017.09.005

Groß Ophoff, J., and Cramer, C. (2022). "The engagement of teachers and school leaders with data, evidence, and research in Germany," in *The Emerald International Handbook of Evidence-Informed Practice in Education*, eds C. Brown and J. Malin (Bingley, UK: Emerald Publishing Limited), 175–195. doi: 10.1108/978-1-80043-141-620221026

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychol. Bull.* 135, 555–588. doi: 10.1037/a0015701

Hendriks, F., Kienhues, D., and Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: the Muenster Epistemic Trustworthiness Inventory (METI). *PLoS ONE* 10, e0139309. doi: 10.1371/journal.pone.0139309

Hendriks, F., Kienhues, D., and Bromme, R. (2016). "Trust in science and the science of trust," in *Trust and Communication in a Digitized World*, ed B. Blöbaum (Cham: Springer International Publishing), 143–159. doi: 10.1007/978-3-319-28059-2_8

Hendriks, F., Seifried, E., and Menz, C. (2021). Unraveling the "smart but evil" stereotype: pre-service teachers' evaluations of educational psychology researchers versus teachers as sources of information. *Z. Pädagog. Psychol.* 35, 157–171. doi: 10.1024/1010-0652/a000300

Hinzke, J.-H., Gesang, J., and Besa, K.-S. (2020). Zur Erschließung der Nutzung von Forschungsergebnissen durch Lehrpersonen. Forschungsrelevanz zwischen Theorie und Praxis (Exploration of the use of research results by teachers. Relevance of research between theory and practice). *Z. Erziehungswiss.* 23, 1303–1323. doi: 10.1007/s11618-020-00982-6

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Jonas, E., Schulz-Hardt, S., Frey, D., and Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *J. Pers. Soc. Psychol.* 80, 557–571. doi: 10.1037/0022-3514.80.4.557

Katz, S., and Dack, L. A. (2014). Towards a culture of inquiry for data use in schools: breaking down professional learning barriers through intentional interruption. *Stud. Educ. Eval.* 42, 35–40. doi: 10.1016/j.stueduc.2013.10.006

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Z. Pädagog. Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

Krammer, G., Vogel, S. E., and Grabner, R. H. (2021). Believing in neuromyths makes neither a bad nor good student-teacher: the relationship between neuromyths and academic achievement in teacher education. *Mind Brain Educ.* 15, 54–60. doi: 10.1111/mbe.12266

Krammer, G., Vogel, S. E., Yardimci, T., and Grabner, R. H. (2019). Neuromythen sind zu Beginn des Lehramtsstudiums prävalent und unabhängig vom Wissen über das menschliche Gehirn. *Z. Bildungsfors.* 9, 221–246. doi: 10.1007/s35834-019-00238-2

Kultusministerkonferenz (2014). *Standards für die Lehrerbildung: Bildungswissenschaften.* Available online at: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf (accessed November 11, 2022).

Landrum, T. J., Cook, B. G., Tankersley, M., and Fitzgerald, S. (2002). Teacher perceptions of the trustworthiness, usability, and accessibility of information from different sources. *Remed. Spec. Educ.* 23, 42–48. doi: 10.1177/074193250202300106

Lavrakas, P. (2008). *Respondent Fatigue*. Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781412963947.n480

Lord, C. G., Lepper, M. R., and Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *J. Pers. Soc. Psychol.* 47, 1231–1243. doi: 10.1037/0022-3514.47.6.1231

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* 37, 2098–2109. doi: 10.1037/0022-3514.37.11.2098

Lüdtke, O., Robitzsch, A., and Grund, S. (2017). Multiple imputation of missing data in multilevel designs: a comparison of different strategies. *Psychol. Methods* 22, 141–165. doi: 10.1037/met0000096

Lüdtke, O., Robitzsch, A., Trautwein, U., and Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen (Handling of missing data in psychological research. Problems and solutions). *Psychol. Rund.* 58, 103–117. doi: 10.1026/0033-3042.58.2.103

Makel, M. C., and Plucker, J. A. (2014). Facts are more important than novelty: replication in the education sciences. *Educ. Res.* 43, 304–316. doi: 10.3102/0013189X14545513

Mansournia, M. A., and Altman, D. G. (2016). Inverse probability weighting. *BMJ* 352, i189. doi: 10.1136/bmj.i189

Masnick, A. M., and Zimmerman, C. (2009). Evaluating scientific research in the context of prior belief: hindsight bias or confirmation bias? *J. Psychol. Sci. Technol.* 2, 29–36. doi: 10.1891/1939-7054.2.1.29

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

Menz, C., Spinath, B., and Seifried, E. (2021). Where do pre-service teachers' educational psychological misconceptions come from? *Z. Pädag. Psychol.* 35, 143–156. doi: 10.1024/1010-0652/a000299

Merk, S., and Rosman, T. (2019). Smart but evil? Student-Teachers' perception of educational researchers' epistemic trustworthiness. *AERA Open* 5, 233285841986815. doi: 10.1177/2332858419868158

Merk, S., Rosman, T., Muis, K. R., Kelava, A., and Bohl, T. (2018). Topic specific epistemic beliefs: extending the theory of integrated domains in personal epistemology. *Learn. Instruct.* 56, 84–97. doi: 10.1016/j.learninstruc.2018.04.008

Muis, K. R. (2004). Personal epistemology and mathematics: a critical review and synthesis of research. *Rev. Educ. Res.* 74, 317–377. doi: 10.3102/00346543074003317

Muis, K. R., Bendixen, L. D., and Haerle, F. C. (2006). Domain-generality and domain-specificity in personal epistemology research: philosophical and empirical reflections in the development of a theoretical framework. *Educ. Psychol. Rev.* 18, 3–54. doi: 10.1007/s10648-006-9003-6

Muis, K. R., Trevors, G., and Chevrier, M. (2016). "Epistemic climate for epistemic change," in *Handbook of Epistemic Cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (Abingdon: Routledge), 213–225.

Mullen, B., Brown, R., and Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: an integration. *Eur. J. Soc. Psychol.* 22, 103–122. doi: 10.1002/ejsp.2420220202

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220. doi: 10.1037/1089-2680.2.2.175

Ostinelli, G. (2009). Teacher education in Italy, Germany, England, Sweden and Finland. *Eur. J. Educ.* 44, 291–308. doi: 10.1111/j.1465-3435.2009.01383.x

Oswald, M. E., and Grosjean, S. (2004). "Confirmation bias," in *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, ed R. F. Pohl (Hove, NY: Psychology Press), 79–96.

Parr, J. M., and Timperley, H. S. (2008). Teachers, schools and using evidence: considerations of preparedness. *Assess. Educ. Princip. Pol. Pract.* 15, 57–71. doi: 10.1080/09695940701876151

Pashler, H., McDaniel, M., Rohrer, D., and Bjork, R. (2008). Learning styles: concepts and evidence. *Psychol. Sci. Publ. Int.* 9, 105–119. doi: 10.1111/j.1539-6053.2009.01038.x

Perry, W. G. (1999). *Forms of Intellectual and Ethical Development in the College Years: A Scheme. Jossey-Bass Higher and Adult Education Series*. Hoboken, NJ: Jossey-Bass.

Ray, B. B., and Hocutt, M. M. (2006). Teacher-created, teacher-centered weblogs: perceptions and practices. *J. Comput. Teach. Educ.* 23, 11–18. doi: 10.1080/10402454.2006.10784555

Renkl, A. (2022). Meta-analyses as a privileged information source for informing teachers' practice? A plea for theories as primus inter pares. *Z. Pädagog. Psychol.* 36, 217–231. doi: 10.1024/1010-0652/a000345

Rosman, T., Kerwer, M., Chasiotis, A., and Wedderhoff, O. (2021). Person- and situation-specific factors in discounting science via scientific impotence excuses. *Eur. J. Psychol.* 17, 288–305. doi: 10.5964/ejop.3735

Rosman, T., and Merk, S. (2021). Teacher's reasons for trust and distrust in scientific evidence: Reflecting a "smart but evil" stereotype? *AERA Open.* 7:23328584211028599. doi: 10.1177/23328584211028599

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581–592. doi: 10.1093/biomet/63.3.581

Schmidt, K., Rosman, T., Cramer, C., Besa, K.-S., and Merk, S. (2022). *Teachers Trust Educational Science - Especially if it Confirms their Beliefs*. OSF. Available online at: https://osf.io/jm4tx/ (accessed November 11, 2022).

Shank, G., and Brown, L. (2007). *Exploring Educational Research Literacy*. New York, NY: Routledge.

Stan Development Team (2017). *Stan Modeling Language: User's Guide and Reference Manual*.

Stark, R. (2017). Probleme evidenzbasierter bzw.-orientierter pädagogischer Praxis [Problems of evidence-based or rather evidence-oriented educational practice]. *Z. Pädagog. Psychol.* 31, 99–110. doi: 10.1024/1010-0652/a000201

Statistisches Bundesamt (Destatis) (2019). Bildung, Forschung und Kultur (Education, research and culture). *Genesis*. (Online).

Stroud, N. J. (2017). "Understanding and overcoming selective exposure and judgment when communicating about science," in *The Oxford Handbook of the Science of Science Communication*, Vol. 1, eds K. H. Jamieson, D. M. Kahan and D. A. Scheufele (Oxford: Oxford University Press), 377–388. doi: 10.1093/oxfordhb/9780190497620.013.41

Taber, C. S., and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *Amer. J. Polit. Sci.* 50, 755–769. doi: 10.1111/j.1540-5907.2006.00214.x

Thiem, J., Preetz, R., and Haberstroh, S. (2020). 'Warum soll ich forschen?' - Wirkungen forschenden Lernens bei Lehramtsstudierenden ('Why should I do research?': effects of research-based learning on student teachers). *Z. Hochschulentw.* 15, 187–207. doi: 10.3217/zfhe-15-02/10

Thomm, E., and Bromme, R. (2012). "It should at least seem scientific!" Textual features of "scientificness" and their impact on lay assessments of online information. *Sci. Educ.* 96, 187–211. doi: 10.1002/sce.20480

Thomm, E., Gold, B., Betsch, T., and Bauer, J. (2021a). When preservice teachers' prior beliefs contradict evidence from educational research. *Brit. J. Educ. Psychol.* 91, e12407. doi: 10.1111/bjep.12407

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021b). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Z. Pädagog. Psychol.* 35, 173–184. doi: 10.1024/1010-0652/a000301

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Interdisciplinary Statistics Series. Boca Raton, FL: Chapman and Hall/CRC Press, Taylor and Francis Group.

Van Lommel, K., Van Gasse, R., Van Hoof, J., and Van Petegem, P. (2017). Teachers' decision-making: data based or intuition driven? *Int. J. Educ. Re.*, 83, 75–83. doi: 10.1016/j.ijer.2017.02.013

van Schaik, P., Volman, M., Admiraal, W., and Schenke, W. (2018). Barriers and conditions for teachers' utilisation of academic knowledge. *Int. J. Educ. Res.* 90, 50–63. doi: 10.1016/j.ijer.2018.05.003

Weißkopf, M., Ziegler, R., and Kremer, B. (2019). *Science Barometer 2018*. Cologne: GESIS Data Archive.

Williams, D., and Coles, L. (2007). Teachers' approaches to finding and using research evidence: an information literacy perspective. *Educ. Res.* 49, 185–206. doi: 10.1080/00131880701369719

Wolter, K., Chowdhury, S., and Kelly, J. (2009). "Chapter 7 - Design, conduct, and analysis of random-digit dialing surveys," in *Handbook of Statistics, Vol. 29*, ed C. R. Rao (Amsterdam: Elsevier), 125–154. doi: 10.1016/S0169-7161(08)00007-2

Zeuch, N., and Souvignier, E. (2016). "Wissenschaftliches Denken bei Lehramts- und Psychologiestudierenden (Scientific thinking in teacher and psychology students)," in *Psychologiedidaktik und Evaluation XI (Psychology Didactics and Evaluation XI)*, eds M. Krämer, S. Preiser, and K. Brusdeylins (Aachen: Shaker Verlag), 175–183.

Zhou, X., and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *Amer. Statist.* 64, 159–163. doi: 10.1198/tast.2010.09109

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 December 2022
DOI 10.3389/feduc.2022.1006079

Check for updates

OPEN ACCESS

EDITED BY
Ingo Kollar,
University of Augsburg,
Germany

REVIEWED BY
Freydis Vogel,
University of Nottingham,
United Kingdom
Jana Groß Ophoff,
Pädagogische Hochschule Vorarlberg,
Austria

*CORRESPONDENCE
Michael Rochnia
rochnia@uni-wuppertal.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 28 July 2022
ACCEPTED 25 November 2022
PUBLISHED 19 December 2022

CITATION
Rochnia M and Gräsel C (2022) Can the utility value of educational sciences be induced based on a reflection example or empirical findings—Or just somehow? *Front. Educ.* 7:1006079.
doi: 10.3389/feduc.2022.1006079

Can the utility value of educational sciences be induced based on a reflection example or empirical findings—Or just somehow?

Michael Rochnia* and Cornelia Gräsel

Institute of Educational Research in the School of Education, University of Wuppertal, Wuppertal, Germany

Educational sciences are a major component of German teacher education. However, student teachers often do not consider educational sciences in university courses (a profession-specific combination of educational psychology, pedagogy and sociology) as helpful for the practice of teaching. To prepare future teachers for evidence-based practice, this is a disadvantageous motivational starting point, because educational sciences offer a large amount of current and relevant findings that can have a positive impact on educational practice. Thus, it would be beneficial for student teachers to see the utility value of educational sciences. The present study attempts to encourage student teachers to perceive the utility value of educational sciences with a utility value short intervention. Utility value interventions contribute to connecting the learning content with one's own life to foster the motivation to use scientific knowledge. A 2×2 quasi-experiment was conducted. Two of the four groups received a utility value short intervention about educational sciences (Factor 1). In addition, a second factor was analyzed that takes up two patterns of educational reasoning in teacher education (Factor 2): Reasoning was either exemplified with an instruction to reflect on the usefulness of educational sciences (like in reflection-oriented educational reasoning) or with exemplary empirical findings from educational sciences (like in evidence-based educational reasoning). These two kinds of reasoning are objectives of teacher education and therefore could influence the effect of a utility value short intervention. Since epistemic goals influence engagement with educational sciences, they are also taken into account. The results showed that all four variants of the treatment increased the students' assessment of the utility value of educational sciences; the utility value intervention had no additional effect. This is discussed with recourse to motivational theories and concepts of teacher education.

KEYWORDS

utility value, evidence, reflection, educational sciences, teacher education

Introduction

In teacher education, there is a broad consensus that university courses in educational sciences—a profession-specific combination of educational psychology, pedagogy and sociology—are useful for evidence-based pedagogical practice (Slavin, 2008; Ferguson, 2021; Fischer, 2021; Renkl, 2022). Our understanding of evidence-based pedagogical practice follows Stark's (2017) broad conceptualization of evidence: Empirical findings and theories are pedagogical knowledge from educational sciences and contribute to ground pedagogical practice on a scientific basis. Educational sciences can therefore be understood as one source of evidence-based practice. This theoretical approach also underpins the fact that there is no direct linear application of evidence to practice. Bromme et al. (2014) emphasized that applying educational sciences is not a linear transfer of rules of action. Rather, applying educational sciences is about using it as a resource to interpret and reflect practice (e.g., *theoretical goggles* at Neuweg, 2013, p. 305; Neuweg, 2022, p. 45).

The scientific consensus on the usefulness of educational sciences is supported by findings regarding teachers' pedagogical knowledge. Teachers' pedagogical knowledge acquired in university courses of educational sciences is correlated positively with teaching quality, self-efficacy and student learning outcomes (Voss et al., 2011; König and Pflanzl, 2016). This means that when teachers have more pedagogical knowledge about the theories and empirical findings of educational sciences, they are more confident in teaching successfully under critical conditions. Furthermore, they realize better teaching and learning. This outlines the usefulness of educational sciences from an evidence-based point of view. However, evidence from educational sciences needs interpretation and can be understood as a step toward applicable knowledge (Groß Ophoff and Cramer, 2022, Figure 1). Such engagement with evidence from educational sciences is influenced by affective-motivational variables (Groß Ophoff and Cramer, 2022). Accordingly, these variables, such as usefulness, can support evidence use. This is underpinned by the effect of teachers' instrumental attitude on the use of data ($\beta = 0.25$; Prenger and Schildkamp, 2018). This finding suggests that when teachers believe in the improvement potential of data and evidence, they are more likely to engage with it. Prenger and Schildkamp (2018) pointed out that it is necessary to demonstrate the importance of research. Since evidence stems from educational sciences, the need to demonstrate its importance also counts for educational sciences.

German student teachers are prepared during teacher education to acquire knowledge from educational sciences and to learn how to apply it in teaching (Stark, 2017; Gogolin et al., 2020). In this context, applying educational sciences means that theoretical and empirical knowledge from educational sciences is used to analyze and cope with situations and requirements of classroom teaching (Bauer et al., 2015). An example with a focus on theory: A student teacher wants to make her lessons in an internship less disruptive. To achieve this, she draws on Kounin (2006) thoughts on disruption prevention and strives for teacher wittiness to prevent classroom disruptions. This process of using educational sciences when confronted with practical requirements can be understood as evidence-based reflection on school practice, which is a preparation for evidence-based pedagogical practice (Cramer et al., 2019; Hartmann et al., 2021). For this process, motivation to engage with educational sciences is required (Bauer et al., 2017).

However, findings show that student teachers lack motivation to engage in educational sciences. Diery et al. (2020) underpin that student teachers are more skeptical about the benefits of educational sciences for practice than teacher educators. A recent study by Voss (2022) pointed in the same direction. This can be interpreted as an indicator that student teachers are not fully convinced of the usefulness of educational sciences. Bråten and Ferguson (2015) also showed that student teachers favor knowledge from practitioners over scientific knowledge. This means that student teachers rate the opinions of teachers as more important for coping with school demands than scientific knowledge of educational sciences. Some studies have shown that the perceived usefulness of educational sciences depends on epistemic aims (Merk et al., 2017; Hendriks et al., 2021; Kiemer and Kollar, 2021). Student teachers generally find knowledge of educational sciences trustworthy for explanations and use it for academic requirements but consider it less relevant for practical requirements or fail to transfer scientific knowledge when dealing with practical educational tasks, such as preparing lessons (Merk et al., 2017; Hendriks et al., 2021; Kiemer and Kollar, 2021). This illustrates that the perceived usefulness of educational sciences depends on what aims the user pursues with the knowledge. Educational sciences might be rated as useful for one task and useless for another. Thus, epistemic aims shape engagement in educational sciences. In accordance, the current state of research illustrates that student teachers have problems perceiving the usefulness of educational sciences. This is an unfavorable motivational disposition for evidence-based reflection on school practice as preparation for evidence-based pedagogical practice. However, utility value intervention is a way to address this inappropriate motivational situation that is derived from expectancy-value theory (Wigfield and Eccles, 2000). In utility value interventions, learners perceive the usefulness of a learning content (Hulleman and Harackiewicz, 2021). Meta-analyses of utility value interventions show an effect size of $d = 0.24$ (Hulleman and Harackiewicz, 2021). According to expectancy-value theory (Wigfield and Eccles, 2000), value has three sources: intrinsic value, attainment value and utility value (Rosenzweig et al., 2019). These sources can be understood as potential starting points for interventions, with the goal of increasing motivation. The different sources have in common that they are important for new ways of seeing, doing, talking, knowing and thinking (Borg, 2010). We chose perceived utility value for designing the intervention in our study, because perceived utility value is more flexible than other sources of value. This flexibility is substantiated by the autonomy-supportive approach of utility value (Hulleman and

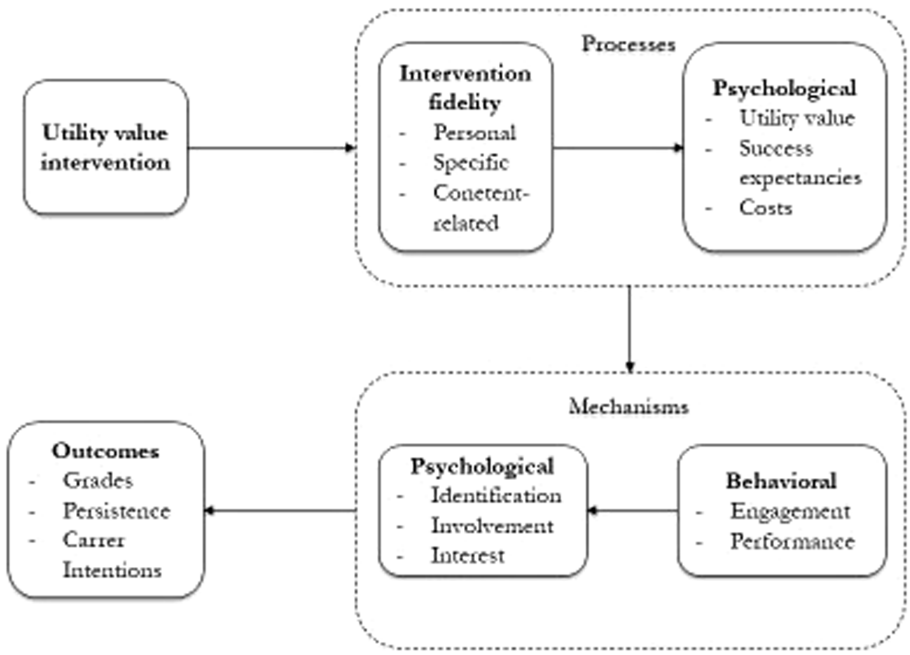


FIGURE 1
Utility value interventions logic model (Hulleman and Harackiewicz, 2021).

Harackiewicz, 2021): Student teachers are free to make their own connections between educational sciences and their lives but are still tied to the goals of teacher education to some degree. Another advantage of utility value interventions is the variability in intervention duration and content. There are short interventions that last about 5 min (Rosenzweig et al., 2019), medium ones that last 20 min (Kosovich et al., 2019), and long approaches that last around 90 min (Gaspard et al., 2015). Similarly, utility value can be fostered with different contents. Students can be asked to write a short essay or letter about the connection between a certain topic and their lives or to rate quotations for relevance and interest (Hulleman and Harackiewicz, 2021).

Recently, some studies have tried to encourage students to engage in educational sciences when dealing with practical educational tasks. Zeeb et al. (2019) promoted the integration of pedagogical knowledge and pedagogical content knowledge by emphasizing the relevance of different knowledge sources among student teachers. The researchers used examples that illustrated the relevance of integrated knowledge. Even though relevance and usefulness are different constructs, the effectiveness of the relevance instruction can serve as a theoretical orientation for usefulness. This is because interventions of usefulness and relevance work similarly (Hulleman and Harackiewicz, 2021). The results show that relevance instructions promote the use of evidence with a medium effect size, $\eta 2 = 0.10$ (Zeeb et al., 2019). A study by Lorentzen et al. (2019) followed a similar path. In that study, the authors successfully promoted the professional relevance of subject-related study content. Participation in the intervention correlated with the perceived relevance of the course as a whole ($\beta = 0.29$; Lorentzen et al., 2019).

All of these approaches have in common that they instruct student teachers on how to use educational sciences rather than letting them perceive the usefulness of educational sciences for themselves. From the perspective of utility value interventions, such approaches are less motivating because messages from external actors will not be internalized to the same degree as personally perceived utility value (Hulleman and Harackiewicz, 2021). Instead, student teachers should be encouraged to perceive the utility value of educational sciences by themselves. The authors point out that inducting utility value needs to take personal relations into account to trigger a mechanism of identification, involvement and interest. This means student teachers need an explanation for why educational sciences are useful but also need to think about the perceived usefulness of educational sciences for themselves and find opportunities to express the perceived usefulness. The present study aims to develop and investigate such an intervention according to the idea of utility value intervention. Due to Covid-19, this is more challenging than before (e.g., Hasselhorn and Gogolin, 2021). Online teaching became the status quo during the pandemic, and contact between student teachers and academic staff was reduced. Therefore, we decided to conduct an online utility value short intervention in educational sciences to avoid further burdening the tense teaching situation. This online short intervention can be implemented in teacher education seminars and lectures in a time-saving way. In the present study, we empirically tested a utility value short intervention.

Utility value as a predictor of the use of educational sciences

Increasing perceived utility value is about motivating students. One main reason for the lack of motivation is that students do not see why they should learn something about a certain topic (Hulleman and Harackiewicz, 2021). This is an adverse motivational disposition for the acquisition (and transfer) of knowledge. University education is also affected by this problem. For example, students find it difficult to see the need for statistics and correlations and often show little motivation to learn these contents (Jang, 2008; Hulleman and Harackiewicz, 2021). In contrast, when students consider a topic valuable, they expend more learning effort and achieve better learning outcomes (Wigfield et al., 2017). This means that when students ascribe value to a topic, they have a more favorable motivational disposition. Increasing perceived value can therefore be understood as promoting motivation to engage in a particular topic.

Here, utility value interventions come into play: Utility value interventions should enable learners to generate their own personal connections between the learning content and their lifeworld (Canning and Harackiewicz, 2015; Durik et al., 2015). How utility value interventions work is shown in Figure 1. To be effective, such connections should be (1) personal, (2) specific, and (3) relevant to the content (Hulleman and Harackiewicz, 2021). From the perspective of utility value theory, this means that connections should be made by the students themselves and should relate to the content as precisely as possible. Perceived utility value means stressing the benefits of certain knowledge, making it clear that knowledge is useful now or in the future (Hulleman and Harackiewicz, 2021). Interventions help clarify the utility and relevance of (scientific) knowledge for students. The strength of perceived utility value is that it connects knowledge acquisition to people's real lives. For example, students were asked to collect statistics (results and graphics) in popular magazines that seemed important to them. These were then discussed in the seminar (Hulleman and Harackiewicz, 2021). This helped the students see the relevance of statistics.

This state of research on utility value interventions can also be applied to teacher education: A utility value short intervention in educational sciences could foster student teachers' experience of the utility value of educational sciences. Empirical findings show that teachers who rate educational sciences as useful use them more often, with a medium effect size, $r = 0.44$ (Rochnia and Trempler, 2019). The situation is similar for student teachers. The rating of the perceived utility value of different sources of knowledge corresponds to the intended processing goals (Viehauser, 2021; Figures 6, 7). This means that the perceived utility value that student teachers associate with educational sciences predicts their use of educational sciences.

A utility value intervention about the use of educational sciences for student teachers and hypotheses

According to interventions on the theoretical basis of utility value, student teachers should perceive the utility value of educational sciences for themselves (Hulleman and Harackiewicz, 2021). Gaspard et al. (2021) proposed that this can be done *via* a combination of two factors: a communication of utility value in an essay-reading task and a short essay-writing task in which students write about their perceived utility value. This shows that utility value interventions consist of two steps: The interventions starts with explaining the utility value of educational sciences to student teachers using an example. This is rather passive and is followed by a more active step—student teachers write about their perceived utility value of educational sciences. This step is the main part of the intervention. Gaspard et al. (2021) suggested that student teachers should read about the utility value of educational sciences to set the stage for thinking about the perceived utility value of educational sciences for themselves in an essay task. The aim of the essay-writing task is to create a link between educational sciences and the student teachers' everyday lives—this is the first factor of our study. A similar approach can also be found in a study by Nickl et al. (2022).

The second factor provides an example of educational reasoning used in teacher education to the student teachers (e.g., Csanadi et al., 2021). Two classical methods of educational reasoning are used. (1) Evidence-based educational reasoning: The importance of educational sciences for teaching quality is shown to students by empirical findings. These findings should foster the perceived utility value of educational sciences. (2) Reflection: This is a common type of educational reasoning in German teacher education (Neuweg, 2021). Students are encouraged to think about educational situations or tasks; mostly, no evidence is provided or used (Hartung-Beck and Schlag, 2020). Classical reflection tasks use a cycle of reflection (e.g., Volmer, 2022, Table 1). Reflection is an established activity in the teaching profession (Schön, 1983; Hargreaves, 2000) and is currently frequently encouraged (Cramer et al., 2019). The findings are perceived by student teachers as less relevant for practice but trustworthy for explanations (Merk et al., 2017;

TABLE 1 Treatment of the present study.

Factor 1: Utility value short intervention (with/without)	EG 1, $n = 36$ with utility value intervention empirical findings	CG 1, $n = 60$ without utility value intervention empirical findings
Factor 2: Educational reasoning (with empirical findings/ or reflection example)	EG 2, $n = 35$ with utility value intervention reflection example	CG 2, $n = 48$ Without utility value intervention reflection example

Hendriks et al., 2021; Kiemer and Kollar, 2021). It is unclear which option of educational reasoning would work best together with a utility value intervention – evidence-based educational reasoning or reflection. Both options can be interpreted as a form of educational reasoning and as a starting point for perceiving the utility value of educational sciences. This means that we distinguish between two ways of reasoning about educational phenomena: an evidence-oriented way based on empirical findings and a reflexive way. The distinction between the two forms of reasoning is rooted in what Hinzke et al. (2020) call the habitus of the design of teaching. The authors distinguish between scientific and praxeological approaches to educational reasoning. The scientific approach is more open to empirical findings than the praxeological approach, which focusses on reflecting educational matters right out from the situation.

Therefore, it is unclear whether student teachers are more likely to draw their own references to the perceived utility value of educational sciences based on empirical findings or reflection as an example. With regard to Hargreaves (2000) and Schön (1983), reflection might be more effective than empirical findings. On the other hand, from the perspective of evidence-based education it is desirable that, empirical findings can be the basis for professional decision making and should encourage student teachers to value the importance of educational sciences. Furthermore, findings show that student teachers judge empirical findings as trustworthy. Thus, there are different theoretical perspectives, and whether "encouraging student teachers to reflect" or "presenting findings to student teachers" has better effects in combination with the utility value short intervention depends on viewpoint. Since no studies have been conducted on this topic, we formulated the research question but no specific hypothesis on the effect of factor 2.

Against this backdrop, we formulated our research questions and tested the hypotheses under control of the epistemic aims, persuasion and comprehensibility of the treatment:

> Research question 1: To what extend does teacher students' perceived utility value of educational sciences improve after being engaged with the treatment? H1: The perceived utility value of educational sciences is higher after the treatment for all groups (effect of time, post-test vs. pre-test).
>
> Research question 2: How does a utility value short intervention affect the utility rating of evidence? H2: The two experimental groups with the utility value short intervention show higher perceived utility values for educational sciences than the two control groups without the utility value short intervention (main effect factor 1, utility value short intervention vs. no utility value short intervention).
>
> Research question 3: How does educational reasoning () influence utility value rating of evidence (main effect factor 2, empirical findings vs. reflection)? Because no direction can be derived from theory, we formulate no hypothesis here.
>
> Research question 4: How does educational reasoning (factor 2) interact with the utility value short intervention (factor 1)? With regard to this research question, we do not formulate a specific hypothesis, because it is a theoretically and empirically open question.

Materials and methods

Participants and design

One hundred and seventy-nine student teachers from six German universities participated in this study (77% female; $M_{age} = 27.47$, $SD_{age} = 4.99$). Recruitment took place *via* university courses and social media. All student teachers were enrolled in a Master of Education program. Participation in the survey was voluntary and not part of a course. Participants were randomly assigned to one of four groups in an experimental 2×2 factorial between-subjects design (see Table 1).

Procedure

An overview of the procedure is presented in Figure 2. The study was conducted online, with a link leading to the experiment. The link was sent to the student teachers. In the first step, participants agreed to participate in the study and then some demographic information was collected. Next, the student teachers rated the perceived utility value of educational sciences (see "Measures"). Subsequently, student teachers were randomly assigned to one of the four groups, as shown in Table 1. The material used is shown in "Materials". Experimental group 1 received the utility value intervention, and the usefulness of the evidence was clarified with empirical results. Experimental group 2 also received the utility value intervention, but the usefulness of the evidence was exemplified by reflection. An exemplary reflection cycle was outlined for this purpose. Control group 1 received no utility intervention—only an explanation of the usefulness of evidence backed by empirical results. Control group 2 received only an example for reflection. Afterwards, the participants again rated the perceived utility value of educational sciences. At the end of the survey, the student teachers were thanked for their participation.

In summary, EG 1 read a text about the usefulness of evidence from educational sciences, justifying its usefulness with empirical findings. After that, they were encouraged to write a short essay about what makes evidence from educational sciences useful in their opinion. This is the explicit utility value intervention. EG 2 engaged in the same short essay task, but before this, they read a different text about the usefulness of evidence from educational sciences. In this text, the usefulness of evidence from educational sciences was backed up with a reflection example. CG 1 and CG 2 were not engaged in short essay writing; both groups just read about the usefulness of evidence from educational sciences. CG 1

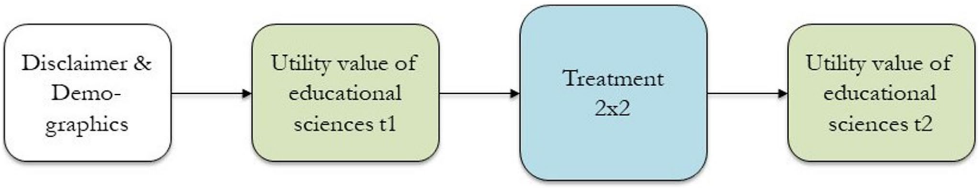


FIGURE 2
Procedure.

TABLE 2 Scales used for the perceived utility value of educational sciences *t*1 and *t*2.

Johnson and Sinatra (2013)		α t1/t2
1.	What I learn in educational sciences I can apply in the classroom.	0.93/0.95
2.	Educational sciences are useful for teaching.	
3.	I think that studies from educational sciences are useful for teaching.	
4.	I can apply my knowledge from educational sciences in school situations.	
5.	Knowledge of theories from educational sciences will be helpful in school.	
6.	It is useful for teachers to know what educational science says about teaching.	
7.	Knowledge of empirical results from educational sciences will be helpful in school.	

Translated from German.

read the same text as EG 1. Likewise, CG 2 received the same text as EG 2.

Materials

The study used four self-generated text materials to create the four groups (see Table 1); all texts were in the German language.

For each text, we calculated the length, Flesch reading ease score, and readability index LIX. All materials are very similar in terms of these indicators of text comprehensibility. The material for the experimental groups was longer because the utility value short intervention was inserted. Materials 1 and 2 started and ended with a utility value short intervention. The first section recognized the role of the student teachers and concluded with an essay assignment in which personal connections were to be drawn between the student teachers' lifeworld and educational sciences. The short essays produced had an average length of 87.75 words. Exemplary statements about the usefulness of educational science were as follows: *(1) The theories taught within the educational sciences provide a kind of framework or foundation for practice. (2) Evidence-based findings from educational science are indispensable at the macro level for teaching–learning processes. (3) I think that many aspects of the educational sciences are useful for later work as a teacher*. Materials 1 and 3 used empirical findings as an example of the utility value of educational sciences. In materials 2 and 4, the example of the utility value of educational sciences was based on reflection. The question about the central message of the text for both control groups served as an implementation check so that the study website was not skipped. All four materials are shown in the Appendix.

Measures

The measures used in this study can be divided into four categories: (1) demographics, (2) perceived utility value of educational sciences, (3) epistemic aims and (4) persuasion and comprehensibility of the treatment. The items in categories 2, 3 and 4 were presented in random order.

1. Demographics: Student teachers were asked for their gender, age, course of study, number of semesters, previously acquired credits and teaching experience.
2. Perceived utility value of educational sciences: We administered a 7-point Likert scale developed by Johnson and Sinatra (2013) before and after the treatment. The measurement of perceived utility value of educational sciences after the treatment (*t*2) was used as the dependent variable. Items and Cronbach's *α* are shown in Table 2. Johnson and Sinatra's (2013) instrument highlights the use of educational research in practice.
3. Epistemic aims: We used two scales with three items each to measure the students' epistemic aims (see Table 3). The 7-point Likert scales were administered before and after the treatment. This measure at *t*2 was used as a control variable.
4. Persuasion and comprehensibility of the treatment: We used two items from Richter (2007) to measure how

TABLE 3 Scales used for epistemic aims *t*1 and *t*2.

Hendriks et al. (2021)	*α t*1/*t*2
Epistemic aim: understanding	0.87/0.89
In educational sciences, my goal is …	
1. … to achieve as much of a comprehensive overview about the state of evidence regarding the topics addressed as possible.	
2. … to understand the addressed topics as thoroughly as possible.	
3. … to deal with as many current scientific findings as possible.	
Epistemic aim: practical knowledge	0.91/0.93
1. … to achieve an overview of possible applications of the topics addressed for school contexts.	
2. … to internalize applications of the topics addressed for the school context.	
3. … to deal with as many possible applications for everyday school life as possible.	

Translated from German.

TABLE 4 Items for persuasion and comprehensibility of the treatment.

Richter (2007)	
1.	How persuasive did you find the passage you just read?
2.	How comprehensible did you find the passage you just read?

Translated from German.

convincing and understandable the treatment was (see Table 4). Both items were administered with a 4-point Likert scale after the treatment. This measure was used as a control variable.

Data analyses

We analyzed the data according to the hypotheses. The alpha error level was set at 0.05, and we used Cohen's (1988) effect size measures.

As the first step, we checked whether the experimental groups and the control groups differed in their ratings of the perceived utility value of educational sciences before the treatment. This was not the case. The ANOVA results were calculated for the group differences in the perceived utility value of educational sciences: $F(3, 175) = 0.635$, $p = 0.593$.

Next, we tested whether the results differed for persuasion and comprehensibility. The ANOVA result for the group difference in persuasion was $F(3, 175) = 0.599$, $p = 0.616$. For comprehensibility, the ANOVA result was close to significance; therefore, we conducted *post hoc* tests ($F(3, 175) = 2.356$, $p = 0.074$). The *post hoc* tests showed that EG 1 rated their material—material 1, as more comprehensible than the other groups rated theirs.

We found no significant differences in epistemic aims between the groups: $F(3, 173) = 0.289$, $p = 0.834$ for the epistemic aim of understanding and $F(3, 173) = 0.435$, $p = 0.728$ for the epistemic aim of gathering practical knowledge.

As the next step, we tested the hypotheses with an ANCOVA with repeated measures and the grouping variable as a between-subject factor. As control variables, we used epistemic aims and the persuasion and comprehensibility of the treatment. We report the Greenhouse–Geisser correction of the ANOVAs and ANCOVAs.

We also analyzed dropout data. We compared the subjects who dropped out of the study with those who completed the study. Forty-seven student teachers (68% female; $M_{age} = 27.47$ s, $SD_{age} = 4.53$) finished the first two parts of the study (see Figure 2) but did not complete the treatment and the post-measures. In EG 1, 17 participants dropped out and 20 student teachers did not finish EG 2. Three subjects left CG 1 and 7 participants dropped out of CG 2. However, t1 data about the perceived utility value of educational sciences and epistemic aims is available for them. This data can be used to investigate differences in important variables of the study between the dropouts and subjects that stayed in the study. We conducted *t*-tests on the utility value of educational sciences and epistemic aims between the 47 student teachers who dropped out and the 179 student teachers who stayed in the study. There were no significant differences in the perceived utility value of educational sciences and epistemic aims between both groups ($p > 0.05$). This can be interpreted as an indication that the dropouts do not differ in key variables of the study from the subjects who remained in the study.

Results

To test hypothesis 1, we checked whether the rating of the perceived utility value of educational sciences after the treatment was higher than before. The ANCOVA result showed a significant effect of time: ($F(1,169) = 13.488$, $p < 0.001$, $\eta 2 = 0.074$). All four treatments had a positive impact on the rating of the perceived utility value of educational sciences (after treatment: $M = 4.58$, $SD = 1.27$; before treatment: $M = 4.39$, $SD = 1.23$), with a small effect size ($d = 0.15$; see Table 5). This result was held under the control of the epistemic aims and the persuasion and comprehensibility of the treatment as covariates. Therefore, the results are in line with H1.

We tested hypothesis 2 to further investigate the differences between the four treatment groups. There was no significant effect of grouping ($F(3,169) = 0.451$, $p < 0.717$, $\eta 2 = 0.008$) or interaction with time ($F(3,169) = 0.586$, $p < 0.625$, $\eta 2 = 0.010$). Combining this with the effect of time, it becomes clear that all variants of the intervention were similarly effective. It was expected that the two groups with utility value intervention would have a positive effect

TABLE 5 Perceived utility value of educational sciences pre- and post-differences.

Group	*n*	t1 *M (SD)*	t2 *M (SD)*
EG1	36	4.30 (1.41)	4.57 (1.48)
EG2	35	4.38 (1.18)	4.56 (1.22)
CG1	60	4.48 (1.27)	4.72 (1.32)
CG2	48	4.34 (1.08)	4.44 (1.10)
All	179	4.39 (1.23)	4.58 (1.27)

on the rating of the perceived utility value of educational sciences. The results led to the rejection of this hypothesis 2 and have relevance for research question 3 and 4. This result also means that the kind of educational reasoning had no effect, neither a main effect (RQ 3) nor an interaction effect (RQ 4), on the utility value intervention.

However, we found a remarkable interaction between treatment and the epistemic aim of understanding: ($F(3,169) = 4.308$, $p = 0.039$, $\eta 2 = 0.025$). Further analyses revealed that student teachers with the epistemic aim of understanding phenomena rated the perceived utility value of educational sciences higher after the treatment. In general, student teachers with the epistemic aim of understanding rated the perceived utility value of educational sciences higher. The result was supported by a main effect: $F(3,169) = 40.908$, $p < 0.001$, $\eta 2 = 0.195$.

Discussion

The findings can be summarized as follows: (1) University students in all treatment groups rated the perceived utility value of educational sciences higher after the treatments. (2) The utility value short intervention had no additional effect. There was no significant difference between the experimental and control groups. (3) Student teachers with the epistemic aim of understanding rated the perceived utility value of educational sciences higher. This means that our utility value short intervention did not have an additional positive effect.

The utility value short intervention gave student teachers the opportunity to make specific personal connections between educational sciences and their lives. According to the literature (Hulleman and Harackiewicz, 2021), the utility value short intervention should have an additional effect next to both types of educational reasoning on the perceived utility value of educational sciences. However, in our study this expected effect did not occur. This might indicate that the connections the student teachers made were not internalized (Hulleman and Harackiewicz, 2021). This means that the student teachers might not really identify with their own connections, and thus, the connections did not change the perceived utility value. Therefore, our study did not ensure that student teachers want to be reflective in terms of educational sciences (e.g., Brown et al., 2021). Furthermore, the connections could be experienced as forced to a certain degree. This would explain why the student teachers did not really identify with the connections they made (Hulleman and Harackiewicz, 2021). However, the connections might not have been specific enough. Perhaps student teachers need help drawing concrete connections between theory and practice—the examples we provided might not have been sufficient. Perhaps a concrete situation in which scientific knowledge clarifies a problem or corrects a mistake would have been better used in the intervention than our explanation without educational content. Our experimental situation did not offer a concrete problem to work on—probably this was not concrete enough to make personal connections. Neither the exemplary reflection nor the exemplary findings (Factor 2) set the stage for an additional effect of the utility value short intervention.

The student teachers rated the empirical findings and the reflection examples as equally convincing. This is surprising, as reflection is a central concern of German teacher education and should therefore be more familiar to student teachers (e.g., Hartung-Beck and Schlag, 2020; Neuweg, 2021). However, the finding that empirical findings are as effective as the reflection example also means that student teachers rate educational sciences as trustworthy (e.g., Hendriks et al., 2021; Kiemer and Kollar, 2021). Reflection may be a guiding principle of teacher education, but student teachers can evidently be convinced by the value of evidence, too, as, according to our results, reflection is not more persuading than empirical findings.

In summary, all groups expressed higher perceived utility values for educational sciences after the treatment. The effect size was smaller than the average effect of utility value interventions reported by Hulleman and Harackiewicz (2021) and smaller than effects in the studies by Zeeb et al. (2019) and Lorentzen et al. (2019). It is conceivable that a short intervention, such as this study, cannot evoke larger effects. Accordingly, a utility value short intervention would not provide any additional value in addition to the examples. Likewise, examples of the usefulness of educational sciences may have overwritten a potential effect of the utility value intervention itself. Therefore, the short essay-writing task had no additional effect because the exemplary texts already worked well in inducing the perceived utility value of educational sciences. Another explanation for the unspecific effects of our intervention might lie in the nature of the short essay task we gave to EG 1 and EG 2. This short essay task could be interpreted as some kind of reflexive writing (e.g., Spalding and Wilson, 2002). Therefore, maybe the short essay task triggered a reflection process like the examples about the usefulness of educational sciences already did—with no additional effect.

The relationship between the utility rating of educational sciences and the epistemic aim of understanding shows that when student teachers aim to understand pedagogical phenomena, they tend to perceive educational sciences as more useful than with an epistemic aim of gaining practical knowledge. This finding is in line with the current state of the research and sheds light on a problem in German teacher education: Student teachers might be disappointed by educational sciences when they expect practical knowledge (e.g., Hendriks et al., 2021; Kiemer and Kollar, 2021). Student teachers who want to understand pedagogical phenomena might find educational sciences more useful. Perhaps they do not expect any practical advice from educational sciences—understanding pedagogical matters is already a sufficient gain for them. This points out a direction for further studies: Epistemic aims might shape the perceived usefulness of educational sciences. Thus, the perceived usefulness of educational sciences depends on what one expects from educational sciences. Our findings underline the fact that there might be epistemic aims that foster (or hinder) engagement with educational sciences. This is in line with conceptions of inquiry learning. If we apply considerations of inquiry learning to engagement with educational science, it becomes clear that the understanding of pedagogical phenomena is in the foreground (e.g., Huber and Reinmann, 2019). Inquiry learning might therefore be an option to drive student teachers' epistemic aims toward understanding.

Therefore, two main directions for improving the intervention are emerging: Should student teachers change their epistemic aims or should educational sciences become more practical? The former illustrates that there are more or less favorable epistemic aims for the study of educational sciences. Those who expect practical tips will probably be chronically disappointed by abstract educational sciences. Future directions to improve the intervention ought to actively address epistemic aims. This means providing student teachers with realistic epistemic aims, i.e., making it clear that understanding a pedagogical issue is already a value that could be useful in future practice. From this perspective, the manipulation of epistemic aims may be a lever to foster the perceived usefulness of educational sciences. On the other hand, it is the task of educational sciences to generate knowledge for solving educational problems. It would be best to work on both possibilities.

Ferguson (2021) offers a useful guiding idea for this kind of teacher education: evidence-informed teaching and practice-informed research. Maybe students should have realistic epistemic aims with regard to educational sciences, i.e. perhaps they should not expect instructions on how to teach successfully. As Neuweg (2013) and Cramer et al. (2019) point out, theoretically permeating a pedagogical matter already brings a profit for practice. Following the perspective of Brown (2017), it could be that this is a challenge that is more likely to be met in research learning communities. In these communities, a group of experienced teachers engage with educational sciences to enhance practice. These communities might be close to what Hulleman and Harackiewicz (2021) understand as an identification with the connection between educational sciences and practice. Additionally, educational sciences should remain open to practical formats of evidence communication (e.g., Seidel et al., 2017). This means that getting educational sciences into practice is a task that can be approached from two sides: fostering the perceived utility value side of the user of educational sciences and providing the practical evidence side of the producer of educational sciences.

Limitations

This study has several limitations. The utility value measurement was somewhat abstract. No case study or similar format was used in which evidence was actually presented. Zeuch et al. (2017; Figure 1), for example, developed vignettes for the assessment of data literacy. Adapting such vignettes to illustrate the usefulness of educational sciences would be a viable avenue for future studies. Instead, our study was based on the intention to use educational sciences under certain conditions and no real pedagogical behavior. This means that our study is not about the usefulness of educational sciences' findings *per se* but about the perceived usefulness by student teachers. Another limitation is the selection of participants. An attempt was made to recruit students from several universities, but the majority of the participants came from one university.

Another weakness of the study relates to the implementation of the utility value intervention. Due to the Covid-19 pandemic, a short online-based intervention was conducted. It is conceivable that a longer face-to-face intervention would have been longer and therefore more effective. We also expect that a face-to-face intervention would suffer less from dropouts (Van Selm and Jankowski, 2006), because poor online response rates are a common issue (Nayak and Narayan, 2019).

Another limitation is the differences in dropouts among the four groups. The participants were unequally distributed among the survey groups, although the survey software randomly assigned the participants to one of the four groups. This means that more participants were not motivated to write down the personal connections between educational sciences and their lifeworld in the essay task of the utility value short intervention and thus finished their participation in the study. This indicates that utility induction itself was not motivating for the students. One could even speak of demotivation, because the student teachers were willing to participate in the study but dropped out in the essay task quite often. The control group (CG 1), in which the utility value of educational sciences was illustrated with empirical results, contained the largest number of participants and therefore experienced the lowest number of dropouts. This can be interpreted as an indication of the motivational effect of this variant. This suggests that the exemplary findings were most likely to be interpreted as interesting and convincing. Furthermore, it could be possible that student teachers have become a little weary of reflection and therefore dropped out. Another limitation of the study is that reflection and empirical findings are not combined. In a further study, the two factors of educational reasoning could be connected. Perhaps an evidence-based reflection might be even more convincing about the utility value of educational science than our present attempts.

Conclusion

This study highlights two points. (1) The utility value short intervention did not have the expected effect. When student teachers wrote about personal connections between their lives and educational sciences, they did not rate the perceived utility value of educational sciences as higher than the other groups. (2) However, the perceived utility value of educational sciences can be fostered. All four intervention conditions raised the rating of the perceived utility value of educational sciences to a small degree. Student teachers with the epistemic aim of understanding were most likely to benefit from all forms of intervention.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study involving human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants in accordance with the national legislation and the institutional requirements.

Author contributions

MR and CG developed the theoretical conception of the study and wrote parts of the manuscript. MR performed all analyses, designed the figures and tables, collected the data, and wrote the main part of the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.1006079/full#supplementary-material

References

Bauer, J., Berthold, K., Hefter, M. H., Prenzel, M., and Renkl, A. (2017). Wie können Lehrkräfte und ihre Schülerinnen und Schüler lernen, fragile Evidenz zu verstehen und zu nutzen? *Psychol. Rundsch.* 68, 188–192. doi: 10.1026/0033-3042/a000363

Bauer, J., Prenzel, M., and Renkl, A. (2015). Evidenzbasierte Praxis—im Lehrerberuf?! Einführung in den Thementeil. *Unterrichtswissenschaft* 43, 188–192.

Borg, S. (2010). Language teacher research engagement. *Lang. Teach.* 43, 391–429. doi: 10.1017/S0261444810000170

Bråten, I., and Ferguson, L. E. (2015). Beliefs about sources of knowledge predict motivation for learning in teacher education. *Teach. Teach. Educ.* 50, 13–23. doi: 10.1016/j.tate.2015.04.003

Bromme, R., Prenzel, J., and Jäger, M. (2014). Empirische bildungsforschung und evidenzbasierte bildungspolitik. *Z. für Erzieh.* 17, 3–54.

Brown, C. (2017). Research learning communities: how the RLC approach enables teachers to use research to improve their practice and the benefits for students that occur as a result. *Res. All* 1, 387–405. doi: 10.18546/RFA.01.2.14

Brown, C., Poortman, C., Gray, H., Gross-Ophoff, J., and Wharf, M. (2021). Facilitating collaborative reflective inquiry amongst teachers: what do we currently know? *Int. J. Educ. Res.* 105:101695. doi: 10.1016/j.ijer.2020.101695

Canning, E. A., and Harackiewicz, J. M. (2015). Teach it, don't preach it: the differential effects of directly-communicated and self-generated utility-value information. *Motiv. Sci.* 1, 47–71. doi: 10.1037/mot0000015

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cramer, C., Harant, M., Merk, S., Drahmann, M., and Emmerich, M. (2019). Meta-Reflexivität und Professionalität im Lehrerinnen-und Lehrerberuf. *Zeitschrift für Pädagogik* 65, 401–423. doi: 10.25656/01:23949

Csanadi, A., Kollar, I., and Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Diery, A., Vogel, F., Knogler, M., and Seidel, T. (2020). Evidence-based practice in higher education: teacher educators' attitudes, challenges, and uses. *Front. Educ.* 5:62. doi: 10.3389/feduc.2020.00062

Durik, A. M., Shechter, O., Noh, M. S., Rozek, C. R., and Harackiewicz, J. M. (2015). What if I can't? Perceived competence as a moderator of the effects of utility value information on situational interest and performance. *Motiv. Emot.* 39, 104–118. doi: 10.1007/s11031-014-9419-0

Ferguson, L. E. (2021). Evidence-informed teaching and practice-informed research. *Zeitschrift für Pädagogische Psychologie* 35, 199–208. doi: 10.1024/1010-0652/a000310

Fischer, F. (2021). Some reasons why evidence from educational research is not particularly popular among (pre-service) teachers: a discussion. *Zeitschrift für Pädagogische Psychologie* 35, 209–214. doi: 10.1024/1010-0652/a000311

Gaspard, H., Dicke, A.-L., Flunger, B., Brisson, B. M., Häfner, I., Nagengast, B., et al. (2015). Fostering adoledcents' value beliefs for mathematics with a relevance intervention in the classroom. *Dev. Psychol.* 51, 1226–1240. doi: 10.1037/dev0000028

Gaspard, H., Parrisius, C., Piesch, H., Kleinhansl, M., Wille, E., Nagengast, B., et al. (2021). The potential of relevance interventions for scaling up: a cluster-randomized trial testing the effectiveness of a relevance intervention in math classrooms. *J. Educ. Psychol.* 113, 1507–1528. doi: 10.1037/edu0000663

Gogolin, I., Hannover, B., and Scheunpflug, A. (2020). "Evidenzbasierung als leitendes Prinzip in der Ausbildung von Lehrerinnen und Lehrern–Editorial" in *Evidenzbasierung in der Lehrkräftebildung. Edition ZfE* (Wiesbaden: Springer), 1–9.

Groß Ophoff, J., and Cramer, C. (2022). "The engagement of teachers and school leaders with data, evidence and research in Germany" in *The Emerald Handbook of*

Evidence-informed Practice in Education. eds. C. Brown and J. R. Malin (Bingley: Emerald Publishing Limited), 175–195.

Hargreaves, D. H. (2000). "The production, mediation and use of professional knowledge among teachers and doctors: a comparative analysis" in *Knowledge Management in the Learning Society* (Paris: OECD Publishing), 219–238.

Hartmann, U., Kindlinger, M., and Trempler, K. (2021). Integrating information from multiple texts relates to pre-service teachers' epistemic products for reflective teaching practice. *Teach. Teach. Educ.* 97:103205. doi: 10.1016/j.tate.2020.103205

Hartung-Beck, V., and Schlag, S. (2020). Lerntagebücher als Reflexionsinstrument imPraxissemester. *HLZ* 3, 75–90. doi: 10.4119/hlz-2492

Hasselhorn, M., and Gogolin, I. (2021). Editoral: Bildung in Corona-Zeiten. *Z. Erzieh.* 24, 233–236. doi: 10.1007/s11618-021-01019-2

Hendriks, F., Seifried, E., and Menz, C. (2021). Unraveling the "smart but evil" stereotype: pre- service teachers' evaluation of educational psychology researchers versus teachers as sources of information. *Zeitschrift für Pädagogische Psychologie* 35, 157–171. doi: 10.1024/1010-0652/a000300

Hinzke, J.-H., Gesang, J., and Besa, K.-S. (2020). Zur Erschließung der Nutzung von Forschungsergebnissen durch Lehrpersonen. Forschungsrelevanz zwischen Theorie und Praxis. *Z. Erzieh.* 23, 1303–1323. doi: 10.1007/s11618-020-00982-6

Huber, L., and Reinmann, G. (2019). *Vom forschungsnahen zum forschenden Lernen an Hochschulen*. Wiesbaden: Springer.

Hulleman, C. S., and Harackiewicz, J. M. (2021). "The utility-value intervention" in *Handbook of Wise Interventions*. eds. G. M. Walton and A. J. Crum (New York, NY: Guilford Press), 100–125.

Jang, H. (2008). Supporting students' motivation, engagement, and learning during an uninteresting activity. *J. Educ. Psychol.* 100, 798–811. doi: 10.1037/a0012841

Johnson, M. L., and Sinatra, G. M. (2013). Use of task-value instruction inductions for facilitating engagement and conceptual change. *Contemp. Educ. Psychol.* 38, 51–63. doi: 10.1016/j.cedpsych.2012.09.003

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence- informed teaching. *Zeitschrift für Pädagogische Psychologie.* 2021, 127–141. doi: 10.1024/1010-0652/a000302

König, J., and Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *Eur. J. Teach. Educ.* 39, 419–436. doi: 10.1080/02619768.2016.1214128

Kosovich, J. J., Hulleman, C. S., Phelps, J., and Lee, M. (2019). Improving algebra success with a utility-value intervention. *J. Dev. Educ.* 42

Kounin, J. S. (2006). *Techniken der Klassenführung*. Münster: Waxmann.

Lorentzen, J., Friedrichs, G., Ropohl, M., and Steffensky, M. (2019). Förderung der wahrgenommenen Relevanz von fachlichen Studieninhalten: Evaluation einer Intervention im Lehramtsstudium Chemie. *Unterrichtswissenschaft* 47, 29–49. doi: 10.1007/s42010-018-00036-1

Merk, S., Rosman, T., Rueß, J., Syring, M., and Schneider, J. (2017). Pre-service teachers' percieved value of general pedagogical knowledge for practice: relations with epistemic beliefs and source beliefs. *PLoS One* 12:e0184971. doi: 10.1371/journal.pone.0184971

Nayak, M. S. D. P., and Narayan, K. A. (2019). Strengths and weaknesses of online surveys. *J. Humanit. Soc. Sci.* 24, 31–38. doi: 10.9790/0837-2405053138

Neuweg, G. H. (2013). Lehrerinnen- und Lehrerbildung durch Wissenschaft: Zur Vielschichtigkeit einer zeitgenössischen Einigungsformel. *Beiträge zur Lehrerbildung* 31, 301–309. doi: 10.25656/01:13851

Neuweg, G. H. (2021). Reflexivität. *Z. für Bild.* 11, 459–474. doi: 10.1007/s35834-021-00320-8

Neuweg, G. H. (2022). *Lehrerbildung. Zwölf Denkfiguren im Spannungsfeld von Wissen und Können*. Münster: Waxmann.

Nickl, M., Sommerhoff, D., Böheim, R., Ufer, S., and Seidel, T. (2022). *Fostering pre-service teachers' assessment skills in simulations: Effects of conceptual prompts and a utility value intervention*. Presentation at EARLI SIG 6 & 7. Zollikofen, Switzerland.

Prenger, R., and Schildkamp, K. (2018). Data-based decision making for teacher and student learning: a psychological perspective on the role of the teacher. *Educ. Psychol.* 38, 734–752. doi: 10.1080/01443410.2018.1426834

Renkl, A. (2022). Meta-analyses as a privileged information source for informing teachers' practice? *Zeitschrift für Pädagogische Psychologie* 2022, 1–15. doi: 10.1024/1010-0652/a000345

Richter, T. (2007). *Epistemologische Einschätzungen beim Textverstehen*. Lengerich: Pabst Science Publishers.

Rochnia, M., and Trempler, K. (2019). Welche externen Wissensquellen bevorzugen Lehrkräfte für ihr professionelles Handeln? *Lehrerbildung auf dem Prüfstand* 12, 125–142.

Rosenzweig, E. Q., Wigfield, A., and Eccles, J. S. (2019). "Expectancy-value theory and its relevance for student motivation and learning" in *The Cambridge Handbook of Motivation and Learning*. eds. K. A. Renninger and S. E. Hidi (Cambridge: Cambridge University Press), 617–644.

Rost, D. H., Sparfeldt, J. R., and Buch, S. R. (Eds.) (2018). *Handwörterbuch Pädagogische Psychologie*. Weinheim: Beltz.

Schön, D. A. (1983). *The Reflective Practitioner. How Professionals Think in Action*. New York, NY: Basic Books.

Seidel, T., Knogler, M., Mok, S. Y., Hetmanek, A., Vogel, F., Bannert, M., et al. (2017). Forschung fördert Bildung. Das Clearing House Unterricht. *J. für Lehrer Innenbildung* 3, 23–28. doi: 10.25656/01:20532

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educ. Res.* 37, 5–14. doi: 10.3102/0013189X08314117

Spalding, E., and Wilson, A. (2002). Demystifying reflection: a study of pedagogical strategies that encourage reflective journal writing. *Teach. Coll. Rec.* 104, 1393–1421. doi: 10.1111/1467-9620.00208

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer Praxis. *Zeitschrift für Pädagogische Psychologie* 31, 99–110. doi: 10.1024/1010-0652/a000201

Van Selm, M., and Jankowski, N. W. (2006). Conducting online surveys. *Qual. Quant.* 40, 435–456. doi: 10.1007/s11135-005-8081-8

Viehauser, S. (2021). *Nutzung studienrelevanter Informationen. Häufigkeitsmuster und motivationale Bedingungen*. Dissertation: Regensburg.

Volmer, M. (2022). Entwicklung der Reflexionsfähigkeit und der Forschenden Haltung von Sachunterrichtsstudierenden im Praxissemester: Konzeption und Evaluation eines Begleitseminars. Dissertation, Westfälische Wilhelms-Universität Münster.

Voss, T. (2022). Not useful to inform teaching practice? Student teachers hold skeptical beliefs about evidence from education science. *Front. Educ.* doi: 10.3389/feduc.2022.976791

Voss, T., Kunter, M., and Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: test construction and validation. *J. Educ. Psychol.* 103, 952–969. doi: 10.1037/a0025125

Wigfield, A., and Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemp. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015

Wigfield, A., Rosenzweig, E. Q., and Eccles, J. S. (2017). "Achievement values" in *Handbook of Competence and Motivation: Theory and Application*. eds. A. Elliot, C. Dweck and D. Yeager (New York, NY: Guilford Press), 116–134.

Zeeb, H., Biwer, F., Brunner, G., Leuders, T., and Renkl, A. (2019). Make it relevant! How prior instructions foster the integration of teacher knowledge. *Instr. Sci.* 47, 711–739. doi: 10.1007/s11251-019-09497-y

Zeuch, N., Förster, N., and Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessments: Results form tests and interviews. *Learn. Disabil. Res. Pract.* 32, 61–70. doi: 10.1037/a0025125

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/feduc.2022.1011241

Check for updates

OPEN ACCESS

EDITED BY
Ingo Kollar,
University of Augsburg,
Germany

REVIEWED BY
Elisabeth Mayweg-Paus,
Humboldt University of Berlin, Germany
Robin Stark,
Saarland University,
Germany

*CORRESPONDENCE
Jana Groß Ophoff
jana.grossophoff@ph-vorarlberg.ac.at

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 04 August 2022
ACCEPTED 14 November 2022
PUBLISHED 04 January 2023

CITATION
Groß Ophoff J, Brown C and Helm C (2023) Do pupils at research-informed schools actually perform better? Findings from a study at English schools. *Front. Educ.* 7:1011241.
doi: 10.3389/feduc.2022.1011241

Do pupils at research-informed schools actually perform better? Findings from a study at English schools

Jana Groß Ophoff[1]*, Chris Brown[2] and Christoph Helm[3]

[1]Institute for Secondary Education, University College of Teacher Education Vorarlberg, Feldkirch, Austria, [2]Department of Education Studies, University of Warwick, Coventry, United Kingdom, [3]Linz School of Education, Johannes Kepler University, Linz, Austria

Introduction: Across the globe, many national, state, and district level governments are increasingly seeking to bring about school "self improvement" *via* the fostering of change, which, at best, is based on or informed by research, evidence, and data. According to the conceptualization of research-informed education as inquiry cycle, it is reasoned that there is value in combining the approaches of data-based decision-making and evidence-informed education. The originality of this paper lies in challenging common claims that teachers' engagement with research supports development processes at schools and pupil performance.

Methods: To put this assumption to test, a data-set based on 1,457 staff members from 73 English primary schools (school year 2014/2015) was (re-)analyzed in this paper. Not only survey information about trust among colleagues, organizational learning and the research use climate was used (*cf*. Brown et al., 2016), but also the results from the most recent school inspections and the results from standardized assessment at the end of primary school. Of particular interest was, as to whether the perceived research use climate mediates the association between organizational learning and trust at school on the one hand and the average pupil performance on the other, and whether schools that were rated as "outstanding," "good," or "requires improvement" in their most recent school inspection differ in that regard. Data was analyzed based on multi-level structural equation modelling.

Results: Our findings indicate that schools with a higher average value of trust among colleagues report more organizational and research informed activities, but also demonstrate better results in the average pupil performance assessment at the end of the school year. This was particularly true for schools rated as "good" in previous school inspections. In contrast, both "outstanding" schools and schools that "require improvement" appeared to engage more with research evidence, even though the former seemed not to profit from it.

Discussion: The conclusion is drawn that a comprehensive model of research-informed education can contribute to more conceptual clarity in future research, and based on that, to theoretical development.

KEYWORDS

research-informed practice, evidence-informed teacher education, data-based decision-making, organisational learning, trust, school inspection, english education system, pupil performance assessment

Introduction

Across the globe, many national, state, and district level governments are increasingly seeking to bring about school "self improvement," *via* the fostering of so-called "bottom-up" change (Brown et al., 2017; Brown, 2020; Malin et al., 2020): change undertaken by school staff to address their needs and which, when optimal, is based on, or informed by research, evidence, and data (Brown and Malin, 2022). The focus of this paper is on the English education system: one that previously has been characterized as *hierachist* (Coldwell, 2022), where rather autonomous local authorities, schools and professionals are seeking to "self-improve" while simultaneously situated within a system of strong central regulation and marketisation of state schooling which serves to influence behaviour (Helgøy et al., 2007). For example, the results in regularly administered standardized national tests or school inspection ratings are publicly accessible, and can be used by parents for their choice of school which, in turn affecting levels of school funding (*cf.*, Coldwell, 2022).

Against this backdrop, this paper aims at investigating the link between the use of data feedback (e.g., school inspection results), the research-use climate at schools, which is framed by organizational conditions like trust among colleagues, or organizational learning in general. This is in turn expected to promote pupil performance in standardized assessments. In the following, research-informed education is introduced as umbrella term for the conceptual link between data-based decision-making and evidence-informed educational practice. As the data set analyzed here is based on a sample of educators at several English primary schools, the specific contextual conditions of the English education system are then introduced. This is followed by an overview of the state of research about the use of research by educational practitioners and schools, and of the specific effects of school inspections. Based on this, the research questions are specified, the methods are explained, and, in conclusion, the results are discussed with reference to the theoretical background of this paper.

Theoretical framework: Research-informed educational practice at English primary schools

As noted above, school "self improvement" is ideally achieved through teachers' engagement with research, evidence, and data (Brown and Malin, 2022). Of course, how this engagement occurs, is as important as that it actually occurs in the first place. This is described in conceptual models of data-based decision-making and evidence-informed education. Even though the two approaches differ in their focus, they share the same assumptions about the sequence of phases nonetheless. The sequence is comparable to the steps of a research process (e.g., Teddlie and Tashakkori, 2006), but also to conceptions of learning as discovery or inquiry process (e.g., Bruner, 1961).

Conceptual models of data-based decision-making and evidence-informed education

Many commentators suggest that an "ideal sequence" of teachers' and schools' engagement with research, evidence or data is that described in conceptual frameworks, which model the process of evidence and data use as a complex, cognitive, knowledge-based problem-solving or inquiry cycle with consecutive phases, that are not ensued in a linear, but rather iterative fashion (Mark and Henry, 2004; Mandinach et al., 2008; Schildkamp and Kuiper, 2010; Coburn and Turner, 2011; Marsh, 2012; Schildkamp and Poortman, 2015; Schratz et al., 2018; Groß Ophoff and Cramer, 2022). This is outlined in Figure 1: In the course of the (1) conceptual phase, a question or problem is identified or a specific goal is set. Subsequently, data is collected and analyzed as part of the empirical phase (2). One point worthy of note here is, that only by way of systematic organization, processing, and analysis under consideration of theories, methods and context knowledge, can *evidence* evolve from (raw) *data* (Bromme et al., 2014; Groß Ophoff and Cramer, 2022). During the inferential phase (3), the available evidence allows focussing attention, provides new insights, challenges beliefs or reframes thinking, without immediate effect on decision-making. This was identified as conceptual research use by Weiss (1998). But evidence can also be used to identify or develop concrete measures to be taken (instrumental use), or even as justification or support of existing positions or established procedures (symbolic use). Based on the inferences drawn, possible change measures are identified, put to practice (4) and at best evaluated (5), too.

This inquiry cycle can be addressed on different levels of the educational system, that is, with regard to the characteristics of data or evidence (inner layer in Figure 1; e.g., graph types, *cf.* Merk et al., in press), of the users (middle layer in Figure 1; e.g., perceived usefulness, *cf.* Prenger and Schildkamp, 2018), or of the context (outer layer in Figure 1, e.g., trust among colleagues, *cf.* Brown et al., 2022). The current study mainly focusses on the latter, namely the schools' culture of trust and organizational learning that can support (or hinder), how open to research schools are perceived by its staff.

Furthermore, the research process phases apply to both an *engagement in research* and an *engagement with research* (Borg, 2010). The former corresponds with being able to independently pass through the full research process, which requires advanced research-methodological competencies (Brown et al., 2017; Voss et al., 2020). The latter refers to the reflection on evidence for professionalization or development purposes, without the necessity to gather forms of evidence, such as research. The *engagement with research* is conceptually related to notions of *data use* and *data-based decision-making,* where approaches in this field

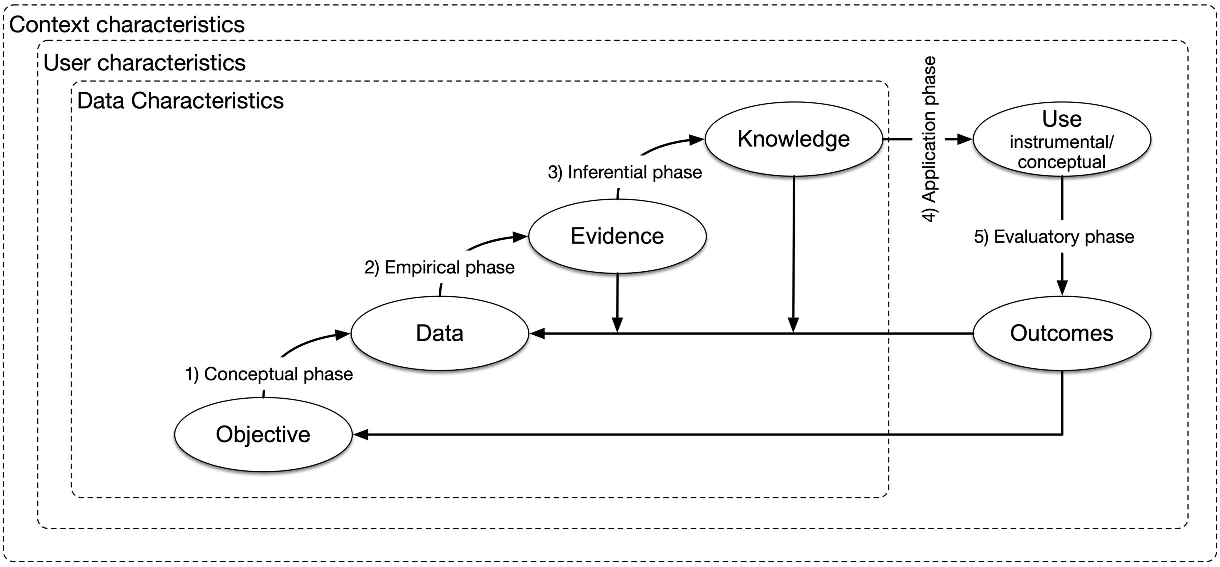


FIGURE 1
Conceptual framework of research informed-educational practice (adapted, a.o. Mark and Henry, 2004; Mandinach et al., 2008; Marsh, 2012; Schildkamp and Poortman, 2015; Schratz et al., 2018; Groß Ophoff and Cramer, 2022).

increasingly take different kinds of data (e.g., from surveys, observations or conversations) into consideration that can be used for educational decisions and development processes (*cf.* Schildkamp et al., 2013; Mandinach and Gummer, 2016; Mandinach and Schildkamp, 2021; Groß Ophoff and Cramer, 2022). Accordingly, data available to schools can be further subdivided into *internal school data* (e.g., student feedback, collegial observations) and *external* sources (e.g., school inspection, regular mandatory pupil performance assessment, central exams), from more "generic" scientific *research* evidence (academic and professional literature, *cf.* Demski, 2017). Wiesner and Schreiner (2019) conceptualize the term "evidence" as a continuum that spans from evidence in the wider sense (e.g., school-internal data) to evidence in the stricter, that is, more scientific sense. The latter is derived as part of an engagement in research and serves as the foundation of theory development (Kiemer and Kollar, 2021; Renkl, 2022), but is usually viewed as rather abstract, impractical information by teachers (Harper et al., 2003; Hammersley, 2004; Zeuch et al., 2017). Less conclusive is the characterization of so-called referential data, e.g., from centrally purported school inspections or pupil performance assessments, as it can be characterized both as a) data and as b) evidence according to the above introduced problem-solving cycle (Groß Ophoff and Cramer, 2022). This ambiguity is used in the following to highlight two different lines of research in this field:

On the one hand (a), data feedback from school inspections or performance assessment might be perceived as "raw" data because it still requires sense-making (2: empirical phase) to being able to develop and implement of concrete instructional or school development measures. This perspective is typical for approaches in the field of *data-based decision-making* (Schildkamp and Kuiper, 2010; Mandinach, 2012; Schildkamp, 2019). Accordingly, Mandinach and Schildkamp (2021) describe the underlying conceptual models of data use as "theories of action" that conceptualize informed educational decision-making as a process of collecting and analysing different forms of data.

On the other hand (b), referential data shows features of research evidence in the stricter sense, as data collection, analysis, and result processing are based on research-methodological and conceptual scientific knowledge. Therefore, data feedback itself can provide teachers and school with scientific knowledge that might be useful in supporting or even supplementing processes during the inferential phase (3), and even might indicate that there is a need for more knowledge (in the sense of informal learning, e.g., Evers et al., 2016; Mandinach and Gummer, 2016; *cf.* Brown et al., 2017). So departing from (a), the (3) inferential phase of the problem-solving cycle is more paramount in this case – even though research evidence can be, of course, useful throughout the whole process (Huguet et al., 2014). Whether the need for more information or a deeper understanding is recognized at all, and which information sources are chosen is meaningful for the depth, scope and direction of subsequent use processes (Brown and Rogers, 2015; Vanlommel et al., 2017; Dunn et al., 2019; Kiemer and Kollar, 2021). Such an engagement with research evidence requires some basic understanding of the underlying scientific concepts or, if necessary, the willingness to acquire the germane knowledge (Rickinson et al., 2020). However, conceptual models of *data-based decision-making* stay rather vague in that regard: For example Schildkamp et al. (2018), describe as late as for the application phase (4) of the Data Team Procedure (a school intervention based on the approach of data-based decision-making), that in order to gather ideas about possible measures, different sources can be used like "the knowledge of team members and colleagues in the school, networks such as the teacher's union, practitioner journals, scientific literature, the internet and the experiences of other schools" (p. 38). In contrast, approaches in the field of *evidence-informed teacher education* (with focus on pre-service teachers, e.g., Kiemer and Kollar, 2021; Greisel et al., 2022) or *research-informed educational practice* (with focus on in-service teachers, e.g., Cain, 2015; Brown et al., 2016, 2017) concentrate on the process of an engagement with evidence (in the sense of acquiring scientific knowledge) and can therefore be described as taking in the perspective of a "theory of learning." Accordingly, Kiemer and Kollar (2021) identify as grounds for evidence-informed education, that in case certain "educational problems come up repeatedly, […] teachers should be able to seek out, obtain and potentially apply what scientific research on teaching and learning has to offer to pave the way for competent action (p. 128), which is why it is "a core task for preservice teacher education to equip future teachers with the skills and abilities necessary to engage in competent, evidence informed teaching" (p. 129). With focus on practicing teachers, but still in the same line of reasoning, Brown et al. (2017) describe the process of research-informed educational practice as the use of "existing research evidence for designing and implementing actions to achieve change" (p. 158).

Despite the different foci of these approaches to data, evidence, or research use in education, Brown et al. (2017) point out that there is value in a comprehensive approach "to educational decision-making that critically appraises different forms of evidence before key improvement decisions are made" (Brown et al., 2017, p. 154) by combining "the best of two worlds" (Brown et al., 2017). Take this example: A certain school is rated "requires improvement" during school inspection particularly in the key judgment category *quality of education* (a.o. based on lesson observations, *cf.* Office for Standards in Education, 2022). According to the inspection report, progress in mathematics is inconsistent because there are missed opportunities for pupils to extend their mathematical knowledge and skills in other subjects. Based on the available recommendations, the school staff aims at providing more challenging and motivating learning activities in mathematics lessons (= data-based decision making). This is why the teachers involved search deliberately for relevant evidence (here for example: cognitive activation in mathematics, e.g., Neubrand et al., 2013), appraise and discuss the available information with their colleagues, but might also obtain advice by a school-based coordinator. The transition between the empirical (2) and the inferential phase (3) (= research-informed educational practice) is probably particularly sensitive to organizational conditions: If organizational learning is a matter of course and staff members trust each other, improvement measures are more easily implemented than in a school climate that is distrustful and adverse to change. Provided that supportive contextual conditions are given, staff at the exemplary school might not only draw the conclusion that there is a need to revise, extend or swap learning materials, but actually decide to put these changes into practice as conclusion of the (3) inferential phase. After the implementation of measures considered suitable (4, application phase), the school staff is interested in evaluating the impact of the concrete change measures (5), for example, based on the pupils' performance in central assessment tests and exams or their learning progress in lessons. Another possibility here is to investigate, whether the collaboration between staff has considerably improved. If the results are not satisfactory or new questions emerge, this in turn can represent the starting point for another cycle of inquiry.

So as the meaning "research" envelopes the full inquiry cycle, and both data and evidence are part of that process (Groß Ophoff and Cramer, 2022), we propose to use the umbrella term "research-informed education," where data-based decision-making (a) and evidence-informed teacher education/educational practice (b) are conceptualized as part of the same comprehensive process. To put this theoretical assumption to test, a data-set based on 1,457 staff members from 73 English primary schools (2014/2015) is (re-)analyzed in this paper. For this sample, not only survey information about trust among colleagues, organizational learning and the research use climate was available (originally used in Brown et al., 2016), but also the results from the most recent school inspections (before the school staff survey) and the results from standardized assessment at the end of primary school (after the survey). As such, the English educational system and its highly regulated, "hierachist" accountability regime (Coldwell, 2022) is explained in more detail below. It serves as the contextual framework of this study (see outer layer of conceptual model, Figure 1), where schools and teachers operate under so-called high-stakes conditions – even though rather the performance of pupils than organizational (learning) processes are affected by such conditions (e.g., Lorenz et al., 2016).

Characteristics of the English educational (accountability) system

The education system in England could, in modern times, be most accurately described as "self-improving" (Greany, 2017). Here accountability systems "combine quasi-market pressures – such as parental choice of school coupled with funding following the learner – with central regulation and control" (Greany and Earley, 2018). A key aspect of this system is the regular school inspections process undertaken by Ofsted (England's school inspection agency). Ofsted inspections are highlighted by many school leaders as a key driver of their behaviour (Chapman, 2001; Greany, 2017). As a result of an inspection, for which there is typically less than 24 h' notice, schools are placed into one of four hierarchical categories of grades. The top grade: "outstanding," typically results in the school becoming more attractive for parents: thus more students apply, and more funding is directed toward the school. Conversely, schools with lower ratings like "requires improvement" find it more challenging to attract families and the funding attached to student applications. In addition, up until 2019, schools rated "outstanding" were exempt for immediate subsequent inspections, meaning that the pressures of accountability are subsequently considerably lessened (with the converse applying to those which require improvement). Given this, it is possible to suggest a theory of action for why inspections might drive school improvement, with school inspection serving the function of (i) gaining and reporting information about school's educational quality, (ii) ensuring accountability with regard to educational standards, (iii) contributing to school development and improvement, and (iv) enforcing an adherence to educational standards and criteria (Landwehr, 2011; Hofer et al., 2020; Ali et al., 2021).

Another characteristic of the "rigorous system of quality control" (Baxter and Clarke, 2013, p. 714) in English education is the regular implementation of standardized assessment tests (SAT). For example, at the end of key stage 2 (KS2: sixth and final year in primary education) statutory external tests in English (reading, writing) and mathematics have to be carried out alongside regular teacher assessment (*cf.*, Isaacs, 2010). The average student at the end of key stage 2 is expected to reach level 4. By way of context, in 2014, 78% of pupils, and in 2015, 80 percent achieved level 4 or above in reading, writing and mathematics combined (Department for Education, 2015). In a recent white paper, the UK Government (2022) announced the ambition (as one of the "levelling up missions") for 90 percent of KS2 pupils to reach the expected standards by 2030. Overall, a number of different variables, student outcomes (e.g., KS2 SAT results), and inspection ratings are publicly available as government produced annual "league tables" of schools under https://www.compare-school-performance.service.gov.uk/, which enables schools to be ranked according to a number of different variables and student outcomes. As a result, it is acknowledged that England's accountability framework both focuses the minds of – and places pressure – on educators to focus on very specific forms of school improvement. More specifically such improvement tends, in the main, on ensuring pupils achieve well in progress tests in key subject areas (e.g., English literacy and mathematics, Ehren, 2018).

Thus, at the centre of this contribution is the *research-informed educational practice* (RIEP) of a sample of primary schools, which is framed by two vital aspects of accountability. In order to being able to hypothesize possible effects, the next chapter gives an overview of the state of research on verifiable effects of research-informed education on teachers, schools and pupil learning.

Research use and school improvement – Is there a link?

Connected with the concept of research-informed education is the expectation that up-to-date (= evidence-informed), hence professional teachers are able to engage with the multitude of data available to them and maybe identify the results as occasion for (more) professional development or instructional and school development. This in turn is expected to improve the quality of teaching, and mediated by that support pupil performance (e.g., Davies, 1999; Slavin, 2002). In other words, research-informed education is about "making the study and improvement of teaching more systematic and "less happen-stance" and relying on evidence to solve local problems of practice" (Ermeling, 2010, p. 378). So, the question is whether this claim is valid. Below an overview of the current state of research is given.

Effects on schools and teachers

On *school and teacher level* (Outcomes, see Figure 1), the benefits thought to accrue from RIEP include improvements in pedagogic knowledge and skills, greater job satisfaction and greater teacher retention, and evidently support a changed perspective on problems, greater teacher confidence or self-efficacy and improved critical faculties, as well as the ability to make autonomous professional decisions (e.g., Lankshear and Knobel, 2004; Boelhauve, 2005; Bell et al., 2010; Mincu, 2014; Godfrey, 2016; König and Pflanzl, 2016). But instead of using data or evidence, research findings indicate that school leaders and practitioners reportedly prefer to rely on intuition during data-based decision-making (Vanlommel et al., 2017), which is prone to confirmation bias and mistakes (Fullan, 2007; Dunn et al., 2019). Moreover, even though teachers and school leaders report to read professional literature regularly (VanLeirsburg and Johns, 1994; Lankshear and Knobel, 2004; Broemmel et al., 2019), there appears to be a strong preference for practical or guidance journals with no or only limited evidence orientation (Hetmanek et al., 2015; Rochnia and Gräsel, submitted), This inclination appears to be rather stable, as it is already observed in initial teacher education (Muñoz and Valenzuela, 2020; Kiemer and Kollar, 2021). Furthermore, Coldwell (2022) draws the conclusion that research use among English teachers is rather low; a result that is comparable to reports from educational systems all over the world (Malin et al., 2020). Regarding the effects of school inspections, "the evidence base [...] is scattered" (Malin et al., 2020, p. 4), but has increased in recent years (a.o., de Wolf and Janssens, 2007; Gärtner and Pant, 2011; Husfeldt, 2011; Penninckx and Vanhoof, 2015). In their recent systematic review, Hofer et al. (2020) identify positive inspections effects on school evaluation activities, probably due to the goal to prepare for future inspections. For example, English schools rated as "requires improvement" are faced with shorter inspection cycles (*cf.* Chapman, 2002). By contrast, negative and non-significant inspection effects were found in general for school or instructional processes (e.g., Gärtner et al., 2014; Ehren et al., 2016).

Effects on pupils

When it comes to *pupils* (Outcomes, see Figure 1), there is nascent, but still inconclusive evidence linking the use of data or evidence to learning or performance (see Figure 1; e.g., Mincu, 2014; Cain, 2015; Cordingley, 2015; Godfrey, 2016; Rose et al., 2017; Crain-Dorough and Elder, 2021). Extant studies have tended to be small scale and qualitative, so providing limited causal pathways linking research use by teachers and improved pupil performance. A similar picture emerges for data-based classroom and school development measures with findings suggesting their effect on pupil performance in the medium and long term is ambiguous (e.g., Hellrung and Hartig, 2013; Richter et al., 2014; Kemethofer et al., 2015; Lai and McNaughton, 2016; Van Geel et al., 2016). A different picture emerges for the effects of school inspections, as a specific example of quality control in education and the ensuing data feedback: According to Hofer et al. (2020), school inspections appear mainly to accomplish the enforcement of policy in schools. Even though they report twice as much non-significant effects identified compared to positive inspection effects, the latter most consistently emerged for pupil performance in standardized achievement tests – which critical voices might trace back to a narrowing of the curriculum and unwarranted teaching to the test (Au, 2007; Collins et al., 2010; Ehren and Shackleton, 2014).

Research questions

Overall, a causal link between inspections and school improvement cannot be clearly supported from the literature. Moreover, whether the available information about the scientific foundation of data feedback, or scientific evidence are considered by educational practitioners, is influenced by data or individual characteristics (a.o., Prenger and Schildkamp, 2018), In particular, the feeling of being controlled is reported to have detrimental effects on the development processes based on data (Kuper and Hartung, 2007; Maier, 2010; Groß Ophoff, 2013). But the focus of this paper lies particularly on the contextual conditions of research informed education. In that regard, innovative and data-based school culture, evidence-oriented leadership, communication, and collaboration (as indicators of trusting relationships, *cf.* Datnow and Hubbard, 2016), but also professional development measures are deemed supportive (Diemer and Kuper, 2011; Groß Ophoff, 2013; Vanhoof et al., 2014; Van Gasse et al., 2016; Van Geel et al., 2016; Wurster, 2016; Brown and Malin, 2017; Keuning et al., 2017; Schildkamp et al., 2018).

As stated above, there is – on one hand – a push by governments toward school and teacher engagement with research, evidence, and data in order to drive school "self improvement"; on the other hand, a lack of evidence suggesting that this actually has any materially positive outcomes for students. Given the hierarchist context within which particularly the English school system operates, and the role of inspection in reporting information about school's educational quality by "signposting" which and how educational standards are to be achieved, in this paper we seek to ascertain a link between both data-based and evidence-informed school "self improvement" and student outcomes. For that purpose, data from a survey study of teachers and school leaders at English primary schools (Brown et al., 2016) is re-analyzed in combination with the schools' previous Ofsted school inspection ratings and the KS 2 SAT results of pupils in the school year 2014/2015 (in both cases referential data, *cf.* Wiesner and Schreiner, 2019). In the survey study, school staff provided information about the *research use* climate (as an indicator of Evidence, see Figure 2) at their respective schools, which corresponds with the transition from the empirical to the inferential phase. In the study presented, characteristics like the perceived trust among colleagues and the organizational learning (both Context, see Figure 2) are particularly taken into consideration, because they are expected to support the research use climate at school (Evidence, see Figure 2). Mediated by ensuing change measures this is supposed to support pupil performance in standardized assessment tests like the KS2 SAT (Outcomes, see Figure 2). Furthermore, schools with other than "outstanding" inspection ratings might particularly perceive pressure to self-improve, for example based on or informed by research. And in that respect, the annual results in centralized assessment tests, but also the implementation of school inspection ratings (Data, see Figure 2) are both feedback sources that are strongly shaped by the specific conditions of the educational system. So, in the case of this study, the English system of quality control in education serves as an example of the more general contextual conditions, under which schools are supposed to improve based on, or informed by data, evidence, or research.

Beyond providing insights into the links between data feedback (school inspection ratings and report) and evidence use at English schools, the specific contribution therefore lies in operationalizing the research use climate at schools as a possible mediating mechanism to support pupil performance on school organizational level. This points to the following research questions, that are ordered according to the steps of our analysis (see Chapter 4):

First of all, we assume that pupil performance (Outcome, see Figure 2) on the school level is positively influenced by the extent of research use at the same school (Evidence):

> RQ1: What effect do teachers' perceptions of the research use climate of their school have on the overall pupil performance in KS2 SATs?

Secondly, the research use climate (Evidence) is itself dependent on the trust and organizational climate within the school (Context, see Figure 2). Therefore, we investigate:

> RQ2: What effect do teachers' perceptions of the presence of in-school organizational learning and collegial trust have on the perceived research use climate at school, but also on the overall pupil performance in KS2 SATs?

Because of that, the research use climate (Evidence) represents a variable that supposedly mediates the relationship between trust/organizational climate (Context) and pupil performance

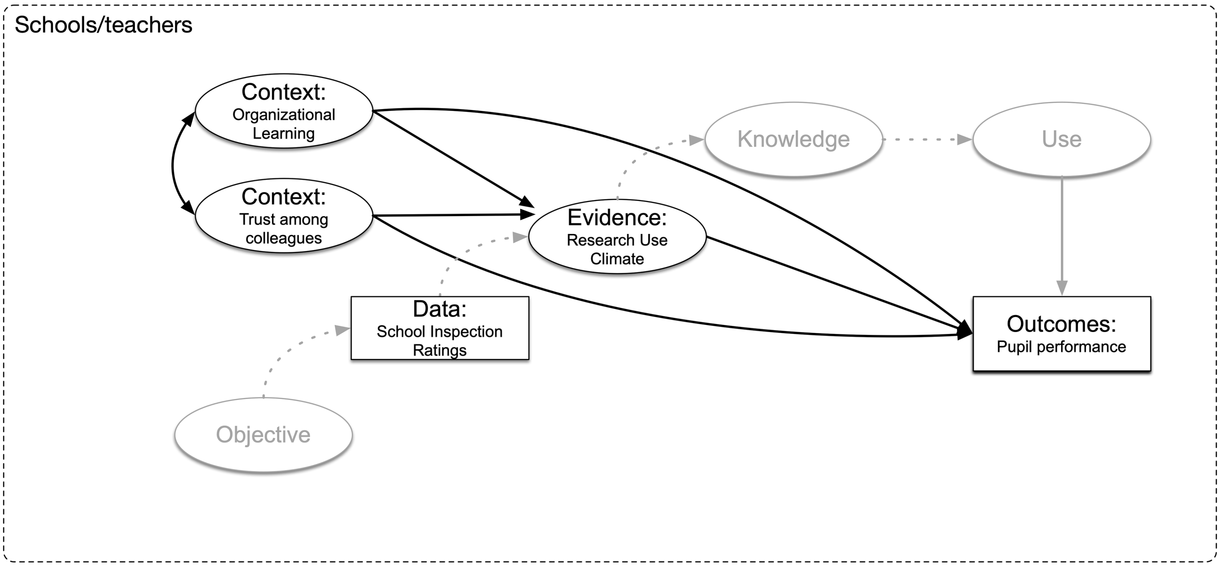


FIGURE 2
Conceptual framework of research informed-educational practice with the constructs included in the study presented in this paper.

(Outcomes). Such an effect should be reflected in statistically significant, indirect effects of trust/organizational climate on student achievement:

> RQ3: Does the perceived research use climate actually mediate the association between organizational learning and trust at school and the overall school performance?

The feedback of school inspection ratings (Data) is expected to shape development activities on school level (Context, Evidence), both regarding their extent, but also the paths between the variables considered in RQ1-RQ3:

> RQ4: Is there a difference in the research use climate between schools that were rated differently in school inspection? And how does this affect the association between trust among colleagues, organizational learning, research use, and performance in KS2 SAT?

Materials and methods

Sample

The sample for this study was gathered within a project that sought to investigate how schools can be supported in applying existing research findings to improve outcomes and narrow the gap in pupil outcomes. Funding for the project was granted by the Education Endowment Foundation in 2014. Schools were recruited by Brown and colleagues through use of Twitter, the direct contacts of the project team and *via* direct mail (e)mailing lists held by the UCL Institute of Education's London Centre for Leadership in Learning. Schools were invited to sign up to the project straight away, to discuss the project and any queries directly with the project team or to attend one of two recruitment events held in June 2014. For the analysis presented below, a sample of 1,457 staff members from 73 primary schools was available. Approximately 20 teachers per school answered the survey. In terms of their characteristics, 70 percent of the study participants had, at that time, less than four years of experience working in their current position. Further, 81 percent were female; approximately 48 percent were serving as a subject leader (e.g., math lead or coordinator); and 18 percent held a formal and senior leadership position (e.g., headteacher). The majority of participating schools were judged as "good" in their most recent school inspection (67.1%), and a smaller amount were graded as "outstanding" (26.0%) or as "requir[ing] improvement" (4.1%), while for 2.7 percent no such information was available.

Data collection and operational variables

The survey data was collected during autumn of 2014 and included self-assessment scales (see below) and demographic background variables. Furthermore, social network data was collected, for which the results have been published elsewhere (Brown et al., 2016). In this paper, additional information on school inspection ratings and pupil performance in KS2 SATs for each participating school have been included in the analysis presented below.

School inspection rating (grouping variable)

School inspection results were documented for each of the schools investigated and were available before the study was carried out. However, it should be noted that the inspections were not carried out within the same time frame, but within a four-year interval for each school. The inspection results were used as control (grouping) variable in the analysis below. Schools are classified on a four-point grading scale used for inspection judgments as outstanding (grade 1), good (grade 2), requires improvement (grade 3), and inadequate (grade 4). To be judged, for example, as "outstanding," schools must "must meet each and every good criterion" of overall school effectiveness (Office for Standards in Education, 2022), that is, with regard to the (former) four key judgment categories (i) achievement of pupils, (ii) the quality of teaching, (iii) the behaviour and safety of pupils, and (iv) leadership and management (Office for Standards in Education, 2022). None of the participating schools were rated as inadequate.

Trust among colleagues (predictor)

Instruments operationalizing trust in colleagues (Hoy and Tschannen-Moran, 2003; Finnigan and Daly, 2012) were adapted to the study sample and context. The final trust (TR) scale consisted of six items, on a five-point Likert type scale, which ranged from 1 (strongly disagree) to 5 (strongly agree), measuring teachers' perceptions as to the levels of trust within their school. For example, by asking respondents to indicate the extent to which they were in agreement with statements such as "Staff in this school trust each other" (TR1, see Table 1). For this scale, Brown et al. (2016) identified a single factor solution on individual level, explaining 52.9 percent of the variance with Cronbach's α of 0.82.

Organizational learning (predictor)

The organizational learning (OL) scale was drawn from a previously validated instrument (Garvin et al., 2008; Finnigan and Daly, 2012) and was again adapted to fit the study context. The OL scale is composed of six items on the same five-point Likert type scale, and measures schools' capacity, cultures, learning environments as well as their structures, systems, and resources. A sample item is: "This school experiments with new ways of working." (OL1, see Table 1). Based on exploratory factor analysis (EFA), on individual level a one-dimensional solution was identified (Brown et al., 2016) explaining 62.2 percent of the variance with Cronbach's α of 0.88.

Research use climate (predictor, mediator)

The research use (RU) climate scale was adapted from a previous study (see Finnigan and Daly, 2012) and is composed of seven items on the same five-point Likert type scale. The construct measures participants' perceptions as to whether school cultures are geared toward research use, both in terms of whether teachers felt encouraged to use research and evidence, and whether they perceived the improvement strategies of their schools to be grounded in research and evidence. For example, a sample item from the scale is "My school encourages me to use research findings to improve my practice." (RU3, see Table 1). Based on EFA with only three of seven items, a single factor solution was identified by Brown et al. (2016) explaining 63.3 percent of the variance with Cronbach's α of 0.71. In the current (re-)analysis, all seven items were included in the identification of the measurement model on both individual and school level.

Pupil performance (criterion)

The outcome variable in the path model analyzed below is average pupil performance operationalized by the percentage of students, who reached level 4 (L4, see Table 1) in the Key Stage 2 (KS2) Standard Assessment Test (SAT), i.e., the expected level to be achieved by the average 11 year old. This high-stakes test is carried out in the core subjects of English and Mathematics (since 2010 not anymore in Science, *cf.* Isaacs, 2010) in English state schools at the end of primary education (year 6) prior to the move to senior school (e.g., Tennent, 2021). The KS2 SAT results are published on school level for accountability and comparative purposes.[1] The data set analyzed here contains KS2 SAT-results for the school year 2014/2015.

Analysis

In the present study, the specific characteristics of the school environment and its effects on a schools' average pupil performance are of interest, which is why the appropriate level of analysis is the group level or school level. Hence, to investigate effects of the school context, the analytical approach of choice is a multilevel (or two-level) model. Accordingly, multilevel structural equation modelling was applied. All analyses were conducted in R (R Core Team, 2021) in combination with Mplus 8 (Muthén and Muthén, 2017).

As the available sample "only" comprises of 73 schools (of which only 3 schools with overall 25 teachers were judged as "requires improvement"), full multilevel latent covariate models (controlling for the measurement and sampling error = latent-latent) may not perform best. Instead, latent-manifest or manifest-latent models are superior with regard to estimation bias (Lüdtke et al., 2011; McNeish and Stapleton, 2016). This is why a sequence of different models [i.e., single level model, multilevel model (latent-latent), multilevel model (latent-manifest)] was performed to gain an understanding, whether different analytical approaches lead to differences in the coefficients of interest (i.e., the effects of trust and research use on school-average performance, see Supplementary Appendix A1). To account for the multilevel structure of the data (teachers clustered in schools; ICC of the items, see Table 1) we used TYPE = COMPLEX in all single level analyses and TYPE = TWOLEVEL in all multilevel analyses. The models were estimated by a robust maximum likelihood estimation (MLR). To determine the model fit, common cut-off criteria were used (Hu and Bentler, 1999) – Bentler's comparative fit index

1 https://www.gov.uk/school-performance-tables

TABLE 1 Overview of the items in the self-assessment scales used.

	Mean	Standard deviation	ICC
Trust scale (TR)			
TR1: Staff in this school trust each other.	4.59	0.83	0.111
TR2: When senior leadership in this school tell you something you can believe it.	4.66	0.78	0.086
TR3: People in this school are eager to share information about what does and does not work.	4.76	0.66	0.062
TR4: When middle leadership in this school tell you something you can believe it.	4.69	0.74	0.105
TR5: Staff in this school respect each other.	4.74	0.68	0.109
TR6: When teachers in this school tell you something you can believe it.	4.71	0.69	0.107
Organizational learning (OL) scale			
OL1: This school experiments with new ways of working.	4.49	1.11	0.032
OL2: This school has a formal process for evaluating programs or practices.	4.30	1.12	0.006
OL3: This school frequently discusses underlying assumptions that might affect key decisions.	4.08	1.14	0.015
OL4: In this school time is made available for education/training activities for school staff.	4.55	1.06	0.010
OL5: This school has forums for sharing information among staff.	4.22	1.18	0.016
OL6: In this school, people value new ideas.	4.38	1.14	0.031
Research Use (RU) Climate scale			
RU1: In the last year, I have discussed relevant research findings with my colleagues.	4.49	0.98	0.101
RU2: Staff at my school use research and evidence to stimulate conversation/dialogue around an issue.	4.39	0.97	0.179
RU3: My school encourages me to use research findings to improve my practice.	4.46	0.95	0.131
RU4: I have found information from research useful in applying new approaches in the classroom.	4.55	1.03	0.054
RU5: Information from research plays an important role in informing my teaching practice.	4.37	1.14	0.049
RU6: I support implementing a school-wide change without research to support it.	4.04	1.09	0.025
RU7: Research and evidence is used to inform staff here about potential improvement strategies.	4.53	0.90	0.090

(CFI ≥ 0.90), the Tucker-Lewis index (TLI ≥ 0.90), the root mean square error of approximation (RMSEA ≤0.08), and the standardized root mean square residual — at both the teacher and school levels (SRMR ≤0.08).

All of the variables used have missing values. Most of the items assessed contained between 9.6% (L4) and 53.2% (OL) missing values. To account for missing information, the Full Information Maximum Likelihood (FIML) method was used as implemented in Mplus.

TABLE 2 Descriptive statistics, reliability, and latent correlations between the study variables.

	M	*SD*	*α*	TR	OL	RU	L4
Trust (TR)	4.69	0.53	0.81	1	**0.825**	**0.436**	**0.341**
Organizational learning (OL)	4.34	0.89	0.88	**0.171**	1	0.733	0.659
Research use (RU)	4.47	0.54	0.74	**0.385**	**0.176**	1	0.262
Level 4 (L4)	0.83	0.09	–	0.158	0.066	0.168	1

Significant correlation coefficients ($\alpha = 0.05$, two-tailed testing) are indicated in bold. Lower diagonal triangle = latent correlations based on single level analysis (model fit: CFI 0.940, TLI.939, RMSEA 0.031, SRMR 0.043). Upper diagonal triangle = level 2 correlations based on manifest-latent variable modelling (model fit: CFI 1.000, TLI 1.000, RMSEA.000, SRMR L2.044).

Results

In the following, descriptive statistics and the model fit evaluation are reported. Subsequently, results related to the four research questions are presented.

Descriptive statistics

Table 2 contains the descriptive statistics, i.e., mean values, standard deviations, as well as reliability information and bivariate correlations. The mean values of the three scales all exceed the value of 4 and are thus quite high for a five-point Likert-scale, indicating ceiling effects. In line with the high mean scale values, the standard deviations are low, i.e., which means that there is high agreement among the teachers in general. Internal consistency (i.e., Cronbach's alpha) is acceptable to good. The Level 4-value of 0.83 indicates further, that at the school investigated here, 83 percent of the students reached level 4 at the KS2 SAT on average.

On the right side of Table 2, correlation coefficients are reported (lower triangle = single level correlations, upper triangle = level 2 correlations). Depending on the analysis level, the associations between the study variables vary in size: In case of single-level analysis, the latent variables correlate only weakly compared to multilevel analysis, where the variables correlate moderately to strongly. In particular, RU and OL show a high correlation ($r = 0.733$). In order to avoid problems related to multicollinearity, in the final multigroup models (model 3–5, see Table 3) a joint factor representing RU and OL was specified, while in single level analyses (model 1 and 2), the originally assumed factor structure is kept (Supplementary Appendix A2 provides a test of the factor structure of RU and OL). The differences between the single-and multilevel model are illustrated in Figures 3, 4.

Model fit evaluation

In Table 4, the fit indices of the final analyses are reported. In Models 1 and 2, data is analysed on level-1 (teachers), while in Models 3 to 5 multilevel analysis was applied (Table 3, second row from left). Furthermore, Models 1, 3 and 4 analyze the full sample, whereas Models 2 and 5 distinguish between schools that were judged as "good" vs. "outstanding" in school inspection (third row from left, multigroup analysis).

According to Table 4, models 1 to 3 show acceptable to good fit, indicating that the proposed measurement models account reasonably for the observed data. In these cases, both measurement and sampling error are controlled. Deviating from this, in Models 4 and 5, residuals were modeled latent-manifest, which means that only sampling error is controlled. Because of that, both models are saturated and no model fit indices are provided.

Results for the research questions

RQ1: What effect do teachers' perceptions of the research use climate of their school have on the overall pupil performance in KS2 SATs?

Teachers' perception of the research use climate of their school is only significantly related to overall pupil performance in KS2 SATs at the school when the data is modeled at teacher level (single level analysis). However, the corresponding effect size is low ($\beta_{RU \rightarrow L4} = 0.130$), but statistically significant ($p = 0.05$). That is, teacher's individual perception of the research use climate is higher at schools with higher pupil performance. If the data is modeled at school level (multilevel analysis), the effect size is large ($\beta_{OLRU \rightarrow L4} = 0.507$) but statistically insignificant ($p = 0.965$).

TABLE 3 Path coefficients (fully standardized) for structural models 1–5.

Model	Groups	TR→RU	OL→RU	TR→L4	OL→L4	RU→L4
Single level analyses						
1	Total	**0.290**	**0.137**	0.107	0.029	**0.130**
2[a]	Good	**0.291**	**0.201**	0.036	−0.065	0.146
	Outstanding	−0.007	−0.041	0.010	0.067	−0.161
Model	Groups		TR→OLRU	TR→L4		OLRU→L4
Multilevel analyses						
3	Total		0.457	0.331		0.028
4	Total		**0.651**	0.013		0.507
5[a]	Good		**0.741**	−0.181		0.517
	Outstanding		0.528	0.202		−0.089

TR, trust scale [see section "School inspection rating (grouping variable)"]; OL, Organizational learning scale (see section "Organizational learning (predictor)"]; RU, Research use climate scale [see section "Research use climate (predictor, mediator)"]. Significant path coefficients ($\alpha = 0.05$, two-tailed testing) are indicated in bold. [a]The number of schools rated as "requires improvement" was too small to be included in a multi-level group comparison of the path models.

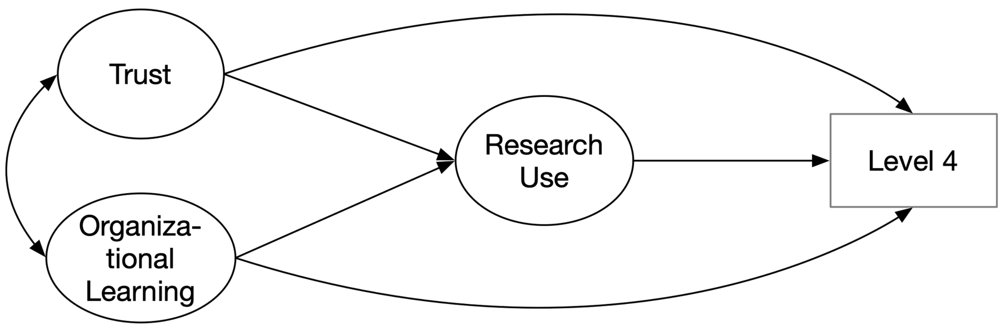


FIGURE 3
Path model on level-1 (1457 school staff members).

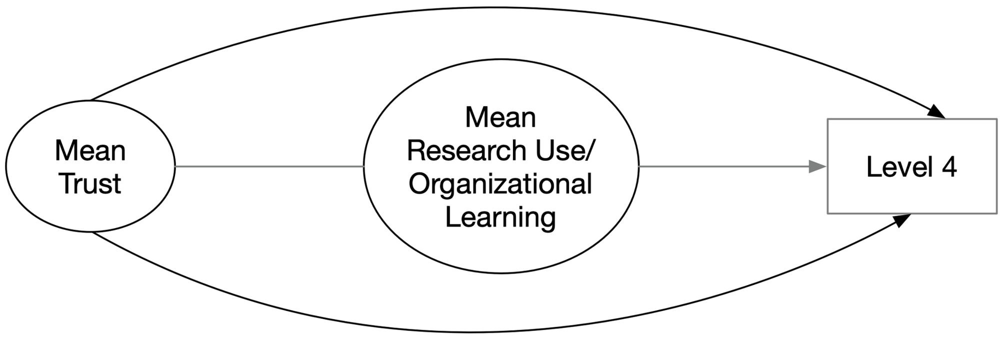


FIGURE 4
Path model on level-2 (73 schools).

TABLE 4 Model fit indices for the structural models on single-level and multilevel for both single group and multigroup analysis and variants of error modelling.

Model	Levels	# groups	Error modelling	Path	# par	Chi²	*df*	Chi²/*df*	CFI	TLI	RMSEA	SRMR
1	SLA	1	Latent	TR,OL → RU → L4	68	401.52	162	2.5	0.932	0.920	0.032	0.046
2	SLA	2	Latent	TR,OL → RU → L4	98	706.44	362	2.0	0.901	0.896	0.038	0.186
3	MLA	1	Latent-latent	TR → OLRU → L4	229	308.76	151	2.0	0.967	0.929	0.027	0.291
4	MLA	1	Latent-manifest	TR → OLRU → L4	12	0	0	0	1.000	1.000	0.000	0.001
5	MLA	2	Latent-manifest	TR → OLRU → L4	24	0	0	0	1.000	1.000	0.000	0.003

SLA, Single-level analysis; MLA, Multilevel analysis; # groups, number of groups (2, multigroup analysis: good vs. outstanding schools); TR, trust scale [see section "Trust among colleagues (predictor)"]; OL, Organizational learning scale [see section "Organizational learning (predictor)"; RU, Research use climate scale [see section "Research use climate (predictor, mediator)"]; # par, number of parameters.

RQ2: What effect do teachers' perceptions of the presence of in-school organizational learning and collegial trust have on the perceived research use climate at school, but also on the overall pupil performance in KS2 SATs?

As expected, teachers' perceptions of the presence of in-school organizational learning and collegial trust significantly predict the perceived research use climate at school. The effect size of in-school organisational learning is $\beta_{TR \rightarrow RU} = 0.137$ ($p = 0.006$) when modeled at teacher level. The effect size of trust is $\beta_{TR \rightarrow RU} = 0.290$ ($p < 0.001$) when modeled at teacher level; and even higher when modeled at school level (latent-manifest: $\beta_{TR \rightarrow RUOL} = 0.457$, $p = 0.060$; latent-latent: $\beta_{TR \rightarrow RUOL} = 0.651$, $p = 0.001$). Contrary to our assumptions, neither teachers' perceptions of the presence of in-school organizational learning nor collegial trust is significantly positively related to pupil performance in the five estimated models. Effect sizes range from $\beta_{OL \rightarrow L4} = -0.181$ to.517. However, they do not reach statistical significance.

RQ3: Does the perceived research use climate actually mediate the association between organizational learning and trust at school and the school performance?

When modeled at teacher level, the indirect effects of organizational learning and trust at school on school performance *via* research use climate at school (= mediation) are weak ($\beta_{OL\_ind} = 0.018$, $\beta_{TR\_ind} = 0.038$) but statistically significant (*95%-CI*$_{OL\_ind}$: [0.006, 0.038], *95%-CI*$_{TR\_ind}$: [0.014, 0.079]). When modeled at the school level, the indirect effect of trust at school on school performance *via* the newly formed joint factor "research use climate and organizational learning at school" (= mediation) is comparably high in effect size ($\beta_{TR\_ind} = 0.330$), but – due to the small sample on school level – does not reach statistical significance.

RQ4: Is there a difference in the research use climate between schools that were rated differently in school inspection? And how does this affect the association between trust among colleagues, organizational learning, research use, and performance in KS2 SAT?

In preparation of the analyses related to RQ4, measurement invariance analyses were conducted to examine whether the instruments used to capture the study variables performed equally well in the three groups with different school inspection ratings (i.e., requires improvement, good, outstanding schools). The results show that only partial measurement invariance could be established (*cf.* Supplementary Appendix A3 for details).

Group comparison results (see Table 5) indicate that particularly schools, who received an school inspection rating as

TABLE 5 Mean differences for trust, organizational learning, research use (individual level) and Level 4-results (school level).

	Scales	"Requires improvement"	"Good"	"Outstanding"	*F*	df (within/between)	*p*	η^2
M (SD)	TR	4.81 (0.28)	4.67 (0.55)	4.83 (0.41)	6.391	2/710	<0.05	0.018 (small effect)
	OL	4.61 (0.43)	4.31 (0.86)	4.49 (0.93)	3.185	2/709	<0.05	0.009 (no effect)
	RU	4.78 (0.33)	4.37 (0.77)	4.72 (0.63)	17.942	2/785	<0.05	0.044 (small effect)
	L4	80.0% (2.8%)	81.4% (8.1%)	90.8% (7.1%)	8.989	2/62	<0.05	0.225 (large effect)
95%-CI	TR	(4.66; 4.96)	(4.62; 4.72)	(4.77; 4.89)				
	OL	(4.38; 4.83)	(4.24; 4.38)	(4.35; 4.63)				
	RU	(4.61; 4.94)	(4.31; 4.43)	(4.63; 4.81)				
	L4	(78.0%; 82.0%)	(80.2%; 82.6%)	(89.0%; 92.6%)				

TR, trust scale [see section "School inspection rating (grouping variable)"]; OL, Organizational learning scale [see section "Organizational learning (predictor)"]; RU, Research use climate scale [see section "Research use climate (predictor, mediator)"]. *M*, mean; *SD*, standard deviation.

"good" (meaning that one key area was identified as "requires improvement," *cf.* Office for Standards in Education, 2022) report a significantly lower mean value for research use climate at school ($M = 4.37$) compared to schools rated as "requires improvement" ($M = 4.78$) and "outstanding" ($M = 4.72$). The same is true for the other study variables trust and organisational learning. In other words, staff at "good" schools report slightly lower trust among colleagues and are a little less active in organizational learning and research use. Furthermore, there is also a considerable difference of 10 % between pupils' performance in KS2 SATs in schools that were rated as "good" or as "requires improvement" and "outstanding" schools.

Regarding the association of the study variables, the multigroup comparison further shows, that only for schools rated as "good," the research use climate is significantly predicted by trust ($\beta_{TR \rightarrow RU} = 0.291$, $p < 0.001$) and organisational learning ($\beta_{OL \rightarrow RU} = 0.201$, $p = 0.008$), while this is not true for schools rated "outstanding" (see Table 3).

Discussion

Even though it is unusual to start the discussion of a paper with its limitations, we would like to point out some, because they are of importance for the conclusions that can be drawn from this study: First of all, the data originates from the school year 2014/2015, and is therefore of some "age." Between 2022 and 2014, certainly a lot of changes have taken place. The United Kingdom's departure from the European Union in 2020 or the world-wide COVID-19-pandemic and the ensuing school closures are two examples that come to mind. English schools were affected by school closures during the pandemic, too, and – among others – the ramifications of cancelling traditional, centralised exams highlighted that the English market-and accountability oriented educational system is prone to crisis (Ziauddeen et al., 2020; McCluskey et al., 2021). But despite the continuing voices of criticism (Jones and Tymms, 2014; Perryman et al., 2018; Grayson, 2019; Coldwell, 2022), the overall educational governance strategy has remained the same (e.g., Office for Standards in Education, 2022; UK Government, 2022). Still, a new framework for school inspection has been introduced in 2019 and was updated just recently (major changes are, a.o., new school inspection labels, end of transition period for updating school curricula, new grade descriptors, etc., *cf.* Office for Standards in Education, 2022). Another limitation of the study presented here is, that even though the school inspection ratings existed before the study was carried out, the research design is not longitudinal, but correlational. In other words, causal interpretations are not eligible here. However, part of the re-analyzed data was originally used in the article by Brown et al. (2016), but particularly the school inspection ratings and the average pupil performance in the Key Stage 2 (KS2) Standard Assessment Test (SAT) were not included then. Therefore, the correlation with school organisational characteristics has not been analyzed until now. Another limitation is that even though the research use climate reported by school staff was included as a theoretically sound mediating variable (Vanhoof et al., 2014; Van Gasse et al., 2016; Keuning et al., 2017; van Geel et al., 2017; Schildkamp, 2019), we cannot know to what extent the schools in this sample engaged with the school inspection results, or more general, what kind of research was actually used, what (if any) improvement measures were implemented, and what other influencing factors were involved. Nor do we know how apt the teachers and school leaders in the current sample were in evidence-informed reasoning, or how they could have been scaffolded in that regard. This rather requires – at best–controlled before-after studies. Some examples for such study designs can be found in the field of *evidence-informed initial teacher education,* of which some are represented in this special issue (e.g., Futterleib et al., 2022; Grimminger-Seidensticker and Seyda, 2022; Lohse-Bossenz et al., 2022; Voss, 2022). Further limitations of this study are, that the sample size at level 2, i.e., the school level, was too low to estimate complex models like the multilevel latent covariate model and multilevel multiple group models. Furthermore, the schools included in this study might be more predisposed to research engagement than the majority of England's primary schools, as they voluntarily applied for participation in a project that sought to investigate how schools can be supported RIEP.

We are aware, of course, that the average pupil performance in KS2-assessments is comparatively distal to collaborative research use processes among school staff, and the findings need to be interpreted with due caution because of that. Nonetheless, we insist that is an important contribution to this field to validate common claims, like for example, that teachers' engagement with research (including centrally administered data and research evidence) facilitates professional and school development, and mediated by that, pupil performance (e.g., Davies, 1999; Slavin, 2002). Therein lies the originality of this paper. As this study was carried out in the field of educational practice, this corroborates the external validity of the findings presented here, too. Another strength is theoretical foundation of our approach (*cf.* Chapter 2) that combines the perspective of data-based decision making (as theory of action, *cf.* Mandinach and Schildkamp, 2021) and of evidence-informed education (as theory of learning, *cf.* Groß Ophoff and Cramer, 2022) – which was already proposed by Brown et al. (2017). For example, in this study the two forms of data feedback on school level were treated as (referential) data (in the tradition of data-based decision-making). According to the phases of the inquiry cycle, the *school inspection results* stand for evaluatory data at the conceptual phase (2), during which school staff ought to appraise the feedback (aka inspection report) under consideration of local school data with the medium-to long-term goal of "self-improvement" (see Figure 1). In turn, *pupils' performance in the KS2 assessments* are located at the end of the inquiry cycle and therefore represent data that schools can use to evaluate (phase 5) the impact of hypothetical improvement measures. Both *trust* among colleagues and *organisational learning* were treated as school-contextual factors, that have been repeatedly identified as important for educational change in general (Louis, 2007; Ehren et al., 2020) and for the use of data or evidence in particular (Schildkamp et al., 2017; Gaussel et al., 2021; Brown et al., 2022). School staff's reported *research use climate* was then modeled as an indicator of evidence-informed educational practice and was included in the analysis as mediator between the school organisational context and pupil performance at the end of the school year 2014/2015. So what is to be learned from the results?

According to the first research question (RQ1), an effect of teachers' perceptions of the research use climate on the overall pupil performance in KS2 SATs was expected. But only a small effect, and only on teacher level could be identified. In other words, school staff who reported a higher research use climate worked at schools with higher pupil performance.

The second question (RQ2) aimed at investigating the effect of teachers' perceptions of the presence of in-school organizational learning and collegial trust on the perceived research use climate at school, but also on the overall pupil performance. In line with the current state of research (see above), our findings indicate that both the perception of in-school organizational learning and collegial trust significantly predict the perceived research use climate at school, but not the pupil performance in central assessments. It should be noted, that on school level the latent constructs of research use and organizational learning could not be separated psychometrically as originally proposed by Brown et al. (2016). In deviation from the original analysis, we used a larger (available) school staff sample and all seven items assigned to research use, and we evaluated the measurement models not only on teacher, but on school level, too. Notwithstanding, this finding is theoretically plausible, as both constructs refer to collaborative learning processes required for the identification, application, and evaluation of school and instructional development measures (Brown et al., 2021).

The third research question (RQ3) pursued as to whether the perceived research use climate mediates the association between organizational learning and trust at school on the one hand and the average pupil performance on the other. This could be demonstrated both on teacher and school level, with the higher effect for the latter. This means that schools with a higher average value of trust among colleagues report more organizational and research informed activities and, mediated by that, demonstrated better results in the average pupil performance assessment at the end of the school year 2014/2015.

The fourth research question (RQ4) finally asked whether there is a difference between schools that were rated as "outstanding," "good," or "requires improvement" in the means of central variables like trust, organizational learning, research use climate, and average pupil performance, and the path model (see RQ3). And in fact, differences emerged here, too:

- "Good" schools demonstrated a significant lower percentage of students, who reached level 4 in the Key Stage 2 (KS2) Standard Assessment Test (SAT) compared to "outstanding" schools, and its staff reported slightly lower trust among colleagues, and to be slightly less active in organizational learning and research use.
- "Outstanding" schools showed considerably better results in KS2 SATs, that could not be predicted by the research-informed organizational learning processes at school or the trust among colleagues. Nonetheless, school staff reported a slightly more pronounced research use climate and trust among colleagues.
- School rated as "requires improvement" showed school organisational results comparable to outstanding schools, but were less successful in pupil performance assessment.

In particular, the more staff at "good" schools reported of organizational and research informed activities, the better was the average pupil performance in centralized assessments. In contrast, both outstanding schools and schools that require improvement appeared to engage more with research evidence, even though the former seemed not to profit (or need?) it. However, no conclusion can be made about schools that required improvement because of their too small proportion (3.9%) in this sample. In comparison, in the school year 2014/2015 19 percent of British primary schools received the same rating (Office for Standards in Education, 2015).

In sum, we could replicate the findings by Brown et al. (2016), according to which trust among colleagues is of vital importance in initiating research-informed educational practice. But our analysis goes beyond, as we could show that assumptions derived from the conceptual model of research-informed educational practice stood the test. School organizational conditions like trust among colleagues and organizational learning proved to be supportive for the research use climate on teacher level, and mediated by that, for the performance of pupils in assessment tests. Particularly interesting is, that on school level, organizational learning and research use climate could not be separated psychometrically, which supports the notion of research-informed education as a learning process that can (and should) take place on school level, too (*cf.* Argyris and Schön, 1978). Furthermore, it could be substantiated, that schools differ in the extent of research-informed organizational learning and in their "paths" to pupils' performance depending on their school inspection ratings. This was particularly true for schools rated as "good" in school inspections. These schools are required to improve at least one key area of educational quality, but reported the lowest research use climate in this study. In our view, this finding is cause for optimism, as particularly such schools seem to benefit from an engagement with evidence and might be convinced to use it more, for example because of by professional development. In line with that, there is ample evidence on the importance that practitioners need to be convinced about the usefulness of data and evidence to engage with it in a meaningful way (e.g., Hellrung and Hartig, 2013; Prenger and Schildkamp, 2018; Rickinson et al., 2020).

So, what are the further implications of this study? First of all, our findings support the notion that the inquiry cycle of research-informed education (see Chapter 2) can and should combine the approaches to data-based decision-making and evidence-informed education (Brown et al., 2017). This is supported by the notion that "the strengths of each appear to mirror and compensate for the weaknesses of the other" (Brown et al., 2017, p. 156). For example, the authors identify as strength of *data-based decision-making*, that the school-specific vision and goals are considered for the problem identification (conceptual phase). Weaknesses are in turn, that ample data literacy is required for meaningful and in-depth data use processes. Data cannot provide educators with solutions that "work best." Instead, a substantive content expertise is needed to be able to identify potential causes and solutions of a problem. In turn, the strength of *evidence-informed educational practice* lies in enabling schools to identify and understand the underlying mechanisms of effective approaches to improving teaching and learning, and provides (at best) instructions how research-informed approaches might be implemented to address a given problem. The pitfalls of this approach are, that it is challenging to recognize the (re-)sources that are adequate and relevant to the problem at hand. This particularly requires research literate teachers and school leaders. Continuing Brown et al.'s (2017) line of reasoning, the field of *evidence-informed teacher education* is another promising asset to the conceptualization of research-informed education. Research in this field focusses on the necessary learning processes and effective teaching strategies in higher education (error-based learning, e.g., Klein et al., 2017; case-based learning, e.g., Syring et al., 2015; inquiry learning, e.g., Wessels et al., 2019). Even though the strong focus on intervention studies in this field is a strength from a research-methodological perspective (internal validity), it is yet unresolved, whether the resulting insights can be applied to professional development and educational practice (external and consequential validity).

A promising approach that somewhat combines the theories of action and learning, is Beck and Nunnaley's (2021) *continuum of data literacy for teaching*: Based on the works of Shulman (1987), Beck and Nunnaley distinguish four levels of expertise (novice users, developing users, developing expert users, expert users) for the use of data and apply them to the different phases of the inquiry cycle. For example, *novice users* may recognize that a problem exists, but are unable to identify relevant and appropriate data sources, and face difficulties in establishing connections between data and one's own teaching methods or even in deriving improvement measures. With increasing expertise, for example *developing users* are already able to recognize the connection between a question and different data and evidence sources, but the effects of changes measures can still only be monitored superficially, etc. Even though the authors remain quite vague as to how particular levels of expertise can be reached, they see potential in practical training that is supported by (academic) mentors for pre-service teachers. For in-service teachers, they propose coaching (e.g., Huguet et al., 2014) or establishing professional learning communities (a.o. Brown, 2017; Brown et al., 2021) based on long-term and goal-oriented engagement with research. But obviously, the current conditions in the English educational-political context are not favorable in that regard: According to Coldwell (2022), educational governance "militates against widespread research evidence use" (Coldwell, 2022, p. 63) due to budget cuts, high accountability pressures, and the (unwarranted) politicization of research use. Related to that, unintended responses and side effects of school inspection and other data-based accountability measures have been a constant, still unresolved issue in this research field (a.o., Ehren and Visscher, 2006; Bellmann et al., 2016). This illustrates further, why it is instructive to take a closer look at educational systems that, in the case of England, represent a high-stakes quality control system, particularly as educational governance tends in general to oscillate between control and autonomy (e.g., Higham and Earley, 2013; Altrichter, 2019).

For future research, a comprehensive model of research-informed education is certainly useful, but necessitates the clarification, which kind of research (data, evidence, or a combination like in the current study) is of interest in the respective study, and related to that, which inquiry phases are to be investigated, what the target group of research users is (pre-service teachers, in-service teachers, or a combination), and on which level (school, staff like teachers or school leaders, pupils, etc.) development processes are expected. In future studies, both schools that "require improvement," but also schools judged as "inadequate" need to be explicitly included, and at best oversampled, in order to gain more in-depth insights into the particular challenges for these schools (e.g., Keuning et al., 2017). But also schools rated as "good" are of interest, and so-called design-based approaches are auspicious in identifying enablers and barriers of self-improvement processes at schools (e.g., Mintrop et al., 2018; Mintrop and Zumpe, 2019). Another question for future research could be in particular, how trust as a crucial contextual condition of research-informed education can be fostered, for example by local educational leaders (e.g., meta-analysis for the economic sector: Legood et al., 2021), but also by central authorities (for the English educational system, e.g., Taysum, 2020). With regard to specifics of the study design presented here, further implications emerge: In particular, the lack of full measurement invariance for all three self-assessment scales (trust, organizational learning, and research use climate, see Data collection and operational variables and Supplementary Appendix A3) indicates that the items have different meanings across schools groups with different Ofsted ratings. This is an interesting finding, that calls for further in-depth inquiry. As mentioned above, the investigated path model relies on theoretically established (causal) assumptions, that cannot be conclusively tested based on the current cross-sectional data, but should be investigated in longitudinal studies. Furthermore, the path from organizational learning and communication processes is a long one, and other plausible mediators are not covered in our dataset, which is another mandate for future research.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data was kindly provided by Chris Brown, University of Warwick, UK. Requests to access these datasets should be directed to Chris.D.B.Brown@warwick.ac.uk.

Ethics statement

The studies involving human participants were reviewed and approved by UCL, Institute of Education. The patients/participants provided their written informed consent to participate in this study.

Author contributions

CB contributed to conception and design of the study. JG and CH organized the database and performed the statistical analysis. JG wrote the first draft of the manuscript. CB and CH wrote sections of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

Funding for the project was granted by the Education Endowment Foundation in 2014. No extra funding was awarded for the current re-analysis of the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2022.1011241/full#supplementary-material

References

Ali, N. J. M., Kadir, S. A., Krauss, S. E., Basri, R., and Abdullah, A. (2021). A proposed framework of school inspection outcomes: understanding effects of school inspection and potential ways to facilitate school improvemessnt. *Int. J. Acad. Res. Bus. Soc. Sci.* 11, 418–435. doi: 10.6007/IJARBSS/v11-i2/8451

Altrichter, H. (2019). "School autonomy policies and the changing governance of schooling" in *Geographies of Schooling* (Cham: Springer), 55–73.

Argyris, C., and Schön, D. A. (1978). *Organizational Learning: A Theory of Action Perspective*. Boston, MA: Addison Wesley.

Au, W. (2007). High-stakes testing and curricular control: a qualitative metasynthesis. *Educ. Res.* 36, 258–267. doi: 10.3102/0013189X07306523

Baxter, J., and Clarke, J. (2013). Farewell to the tick box inspector? Ofsted and the changing regime of school inspection in England. *Oxf. Rev. Educ.* 39, 702–718. doi: 10.1080/03054985.2013.846852

Beck, J. S., and Nunnaley, D. (2021). A continuum of data literacy for teaching. *Stud. Educ. Eval.* 69:100871. doi: 10.1016/j.stueduc.2020.100871

Bell, M., Cordingley, P., Isham, C., and Davis, R. (2010). *Report of professional practitioner use of research review: practitioner engagement in and/or with research*. Available at: http://www.curee.co.uk/node/2303

Bellmann, J., Dužević, D., Schweizer, S., and Thiel, C. (2016). Nebenfolgen Neuer Steuerung und die Rekonstruktion ihrer Genese. Differente Orientierungsmuster schulischer Akteure im Umgang mit neuen Steuerungsinstrumenten. *Z. Pädagog.* 62, 381–402. doi: 10.25656/01:16764

Boelhauve, U. (2005). "Forschendes Lernen–Perspektiven für erziehungswissenschaftliche Praxisstudien" in *Zentren Für Lehrerbildung–Neue Wege Im Bereich Der Praxisphasen (Paderborner Beiträge Zur Unterrichtsforschung)*, vol. *10* (Münster), 103–126.

Borg, S. (2010). Language teacher research engagement. *Lang. Teach.* 43, 391–429. doi: 10.1017/S0261444810000170

Broemmel, A. D., Evans, K. R., Lester, J. N., Rigell, A., and Lochmiller, C. R. (2019). Teacher Reading as professional development: insights from a National Survey. *Read. Horiz.* 58:2.

Bromme, R., Prenzel, M., and Jäger, M. (2014). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. *Z. Erzieh.* 17, 3–54. doi: 10.1007/s11618-014-0514-5

Brown, C. (2017). Research learning communities: how the RLC approach enables teachers to use research to improve their practice and the benefits for students that occur as a result. *Res. All* 1, 387–405. doi: 10.18546/RFA.01.2.14

Brown, C. (2020). *The Networked School Leader: How to Improve Teaching and Student Outcomes Using Learning Networks*. Bingley Emerald

Brown, C., Daly, A., and Liou, Y.-H. (2016). Improving trust, improving schools: findings from a social network analysis of 43 primary schools in England. *J. Profess. Capit. Commun.* 1, 69–91. doi: 10.1108/JPCC-09-2015-0004

Brown, C., MacGregor, S., Flood, J., and Malin, J. (2022). Facilitating research-informed educational practice for inclusion. Survey findings from 147 teachers and school leaders in England. *Front. Educ.* 7:890832. doi: 10.3389/feduc.2022.890832

Brown, C., and Malin, J. (2017). Five vital roles for school leaders in the pursuit of evidence of evidence-informed practice. *Teach. Coll. Rec.*

Brown, C., and Malin, J. R. (2022). *The Emerald Handbook Of Evidence-Informed Practice in Education: Learning from International Contexts*. Bingley Emerald Group Publishing

Brown, C., Poortman, C., Gray, H., Groß Ophoff, J., and Wharf, M. (2021). Facilitating collaborative reflective inquiry amongst teachers: what do we currently know? *Int. J. Educ. Res.* 105:101695. doi: 10.1016/j.ijer.2020.101695

Brown, C., and Rogers, S. (2015). Knowledge creation as an approach to facilitating evidence informed practice: examining ways to measure the success of using this method with early years practitioners in Camden (London). *J. Educ. Chang.* 16, 79–99. doi: 10.1007/s10833-014-9238-9

Brown, C., Schildkamp, K., and Hubers, M. D. (2017). Combining the best of two worlds: a conceptual proposal for evidence-informed school improvement. *Educ. Res.* 59, 154–172. doi: 10.1080/00131881.2017.1304327

Bruner, J. S. (1961). The act of discovery. *Harv. Educ. Rev.* 31, 21–32.

Cain, T. (2015). Teachers' engagement with research texts: beyond instrumental, conceptual or strategic use. *J. Educ. Teach.* 41, 478–492. doi: 10.1080/02607476.2015.1105536

Chapman, C. (2001). Changing classrooms through inspection. *Sch. Leadersh. Manag.* 21, 59–73. doi: 10.1080/13632430120033045

Chapman, C. (2002). Ofsted and school improvement: teachers' perceptions of the inspection process in schools facing challenging circumstances. *Sch. Leadersh. Manag.* 22, 257–272. doi: 10.1080/1363243022000020390

Coburn, C. E., and Turner, E. O. (2011). Research on data use: a framework and analysis. *Measurement* 9, 173–206. doi: 10.1080/15366367.2011.626729

Coldwell, M. (2022). "Evidence-informed teaching in England" in *The Emerald Handbook of Evidence-Informed Practice in Education*. eds. C. Brown and J. R. Malin (Bingley: Emerald Publishing Limited), 59–68.

Collins, S., Reiss, M., and Stobart, G. (2010). What happens when high-stakes testing stops? Teachers' perceptions of the impact of compulsory national testing in science of 11-year-olds in England and its abolition in Wales. *Assess. Educat.* 17, 273–286. doi: 10.1080/0969594X.2010.496205

Cordingley, P. (2015). The contribution of research to teachers' professional learning and development. *Oxf. Rev. Educ.* 41, 234–252. doi: 10.1080/03054985.2015.1020105

Crain-Dorough, M., and Elder, A. C. (2021). Absorptive capacity as a means of understanding and addressing the disconnects between research and practice. *Rev. Res. Educ.* 45, 67–100. doi: 10.3102/0091732X21990614

Datnow, A., and Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: a literature review of international research. *J. Educ. Chang.* 17, 7–28. doi: 10.1007/s10833-015-9264-2

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, 47, 108–20.

de Wolf, I. F., and Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxf. Rev. Educ.* 33, 379–396. doi: 10.1080/03054980701366207

Demski, D. (2017). *Evidenzbasierte Schulentwicklung: Empirische Analyse eines Steuerungsparadigmas*. Wiesbaden Springer.

Department for Education. (2015). National *curriculum assessments at key stage 2 in England, 2015 (revised)* (National Statistics SFR 47/2015; p. 23). Department for Education. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/483897/SFR47_2015_text.pdf

Diemer, T., and Kuper, H. (2011). Formen innerschulischer Steuerung mittels zentraler Lernstandserhebungen. *Zeitsch. Pädag.* 57, 554–571. doi: 10.25656/01:8746

Dunn, K. E., Skutnik, A., Patti, C., and Sohn, B. (2019). Disdain to acceptance: future teachers' conceptual change related to data-driven decision making. *Act. Teach. Educat.* 41, 193–211. doi: 10.1080/01626620.2019.1582116

Ehren, M. C. (2018). *Accountability and trust: two sides of the same coin? Invited talk as part of the University of Oxford's: Intelligent accountability symposium II*. Oxford.

Ehren, M. C., Honingh, M., Hooge, E., and O'Hara, J. (2016). Changing school board governance in primary education through school inspections. *Educat. Manag.* 44, 205–223. doi: 10.1177/1741143214549969

Ehren, M., Paterson, A., and Baxter, J. (2020). Accountability and trust: two sides of the same coin? *J. Educ. Chang.* 21, 183–213. doi: 10.1007/s10833-019-09352-4

Ehren, M. C., and Shackleton, N. (2014). *Impact of School Inspections on Teaching and Learning in Primary and Secondary Education in the Netherlands*, Netherlands, London: Institute of Education.

Ehren, M. C., and Visscher, A. J. (2006). Towards a theory on the impact of school inspections. *Br. J. Educ. Stud.* 54, 51–72. doi: 10.1111/j.1467-8527.2006.00333.x

Ermeling, B. A. (2010). Tracing the effects of teacher inquiry on classroom practice. *Teach. Teach. Educ.* 26, 377–388. doi: 10.1016/j.tate.2009.02.019

Evers, A. T., Kreijns, K., and Van der Heijden, B. I. (2016). The design and validation of an instrument to measure teachers' professional development at work. *Stud. Contin. Educ.* 38, 162–178. doi: 10.1080/0158037X.2015.1055465

Finnigan, K. S., and Daly, A. J. (2012). Mind the gap: organizational learning and improvement in an underperforming urban system. *Am. J. Educ.* 119, 41–71. doi: 10.1086/667700

Fullan, M. (2007). *The New Meaning of Educational Chang*, New York: Routledge.

Futterleib, H., Thomm, E., and Bauer, J. (2022). The scientific impotence excuse in education – disentangling potency and pertinence assessments of educational research. *Front. Educat.* 7:1006766. doi: 10.3389/feduc.2022.1006766

Gärtner, H., and Pant, H. A. (2011). How valid are school inspections? Problems and strategies for validating processes and results. *Stud. Educ. Eval.* 37, 85–93. doi: 10.1016/j.stueduc.2011.04.008

Gärtner, H., Wurster, S., and Pant, H. A. (2014). The effect of school inspections on school improvement. *Sch. Eff. Sch. Improv.* 25, 489–508. doi: 10.1080/09243453.2013.811089

Garvin, D., Edmondson, A., and Gino, F. (2008). Is yours a learning organization? *Harv. Bus. Rev.* 86, 109–116. PMID: 18411968

Gaussel, M., MacGregor, S., Brown, C., and Piedfer-Quêney, L. (2021). Climates of trust, innovation, and research use in fostering evidence-informed practice in French schools. *Int. J. Educ. Res.* 109:101810. doi: 10.1016/j.ijer.2021.101810

Godfrey, D. (2016). Leadership of schools as research-led organisations in the English educational environment: cultivating a research-engaged school culture. *Educat. Manag. Administr. Leader.* 44, 301–321. doi: 10.1177/1741143213508294

Grayson, H. (2019). School accountability in England: a critique. *Headteacher Update* 2019, 24–25. doi: 10.12968/htup.2019.1.24

Greany, T. (2017). Karmel oration-leading schools and school systems in times of change: a paradox and a quest. Research Conference 2017-Leadership for Improving Learning: Lessons from Research, 15–20.

Greany, T., and Earley, P. (2018). The paradox of policy and the quest for successful leadership. *Prof. Develop. Today* 19, 6–12.

Greisel, M., Wekerle, C., Wilkes, T., Stark, R., and Kollar, I. (2022). Pre-service teachers' evidence-informed reasoning: do attitudes, subjective norms, and self-efficacy facilitate the use of scientific theories to analyze teaching problems? *Psychol. Learn. Teach.*:147572572211139. doi: 10.1177/14757257221113942

Grimminger-Seidensticker, E., and Seyda, M. (2022). Enhancing attitudes and self-efficacy toward inclusive teaching in physical education pre-service teachers: results of a quasi-experimental study in physical education teacher education. *Front. Educat.* 7:909255. doi: 10.3389/feduc.2022.909255

Groß Ophoff, J. (2013). *Lernstandserhebungen: Reflexion und Nutzung* Münster: Waxmann.

Groß Ophoff, J., and Cramer, C. (2022). "The engagement of teachers and school leaders with data, evidence and research in Germany" in *The Emerald International Handbook of Evidence-Informed Practice in Education*. eds. C. Brown and J. R. Malin (Bingley: Emerald), 175–196.

Hammersley, M. (2004). "Some questions about evidence-based practice in education" in *Evidence-Based Practice in Education*. eds. G. Thomas and R. Pring, vol. *2004* (Berkshire: Open University Press), 133–149.

Harper, D., Mulvey, R., and Robinson, M. (2003). Beyond evidence based practice: rethinking the relationship between research, theory and practice. *Appl. Psychol.*, 158–171. doi: 10.4135/9781446279151.n11

Helgøy, I., Homme, A., and Gewirtz, S. (2007). "Local autonomy or state control? Exploring the effects of new forms of regulation in education" in *European Educational Research Journal*, vol. *6* (London, England: SAGE Publications Sage UK), 198–202.

Hellrung, K., and Hartig, J. (2013). Understanding and using feedback–a review of empirical studies concerning feedback from external evaluations to teachers. *Educ. Res. Rev.* 9, 174–190. doi: 10.1016/j.edurev.2012.09.001

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M. R., Gräsel, C., et al. (2015). Wozu nutzen Lehrkräfte welche Ressourcen? Eine Interviewstudie zur Schnittstelle zwischen bildungswissenschaftlicher Forschung und professionellem Handeln im

Higham, R., and Earley, P. (2013). School autonomy and government control: school leaders' views on a changing policy landscape in England. *Educat. Manag. Administr. Leader.* 41, 701–717. doi: 10.1177/1741143213494191

Hofer, S. I., Holzberger, D., and Reiss, K. (2020). Evaluating school inspection effectiveness: a systematic research synthesis on 30 years of international research. *Stud. Educ. Eval.* 65:100864. doi: 10.1016/j.stueduc.2020.100864

Hoy, W. K., and Tschannen-Moran, M. (2003). "The conceptualization and measurement of faculty trust in schools: the omnibus T-scale" in *Studies in Leading and Organizing Schools*. eds. W. K. Hoy and C. G. Miskel (Greenwich, CT: Information Age Publishing), 181–208.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* 6, 1–55. doi: 10.1080/10705519909540118

Huguet, A., Marsh, J. A., and Farrell, C. C. (2014). Building teachers' data-use capacity: insights from strong and developing coaches. *Educat. Pol. Anal. Arch.* 22, 1–28. doi: 10.14507/epaa.v22n52.2014

Husfeldt, V. (2011). Wirkungen und Wirksamkeit der externen Schulevaluation. *Z. Erzieh.* 14, 259–282. doi: 10.1007/s11618-011-0204-5

Isaacs, T. (2010). Educational assessment in England. *Assess. Educat.* 17, 315–334. doi: 10.1080/0969594X.2010.491787

Jones, K., and Tymms, P. (2014). Ofsted's role in promoting school improvement: the mechanisms of the school inspection system in England. *Oxf. Rev. Educ.* 40, 315–330. doi: 10.1080/03054985.2014.911726

Kemethofer, D., Zuber, J., Helm, C., Demski, D., and Rie, C. (2015). Effekte von Schulentwicklungsmaßnahmen auf Schüler/innenleistungen im Fach Mathematik. *SWS-Rundschau* 55, 26–47.

Keuning, T., Van Geel, M., and Visscher, A. (2017). Why a data-based decision-making intervention works in some schools and not in others. *Learn. Disabil. Res. Pract.* 32, 32–45. doi: 10.1111/ldrp.12124

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Z. Pädagog. Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

Klein, M., Wagner, K., Klopp, E., and Stark, R. (2017). Fostering of applicable educational knowledge in student teachers: effects of an error-based seminar concept and instructional support during testing on qualities of applicable knowledge. *J. Educ. Res. Online* 9, 88–104. doi: 10.25656/01:14898

König, J., and Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *Eur. J. Teach. Educ.* 39, 419–436. doi: 10.1080/02619768.2016.1214128

Kuper, H., and Hartung, V. (2007). Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. *Eine professionstheoretische Analyse. Zeitschrift Für Erziehungswissenschaft* 10, 214–229. doi: 10.1007/s11618-007-0028-5

Lai, M. K., and McNaughton, S. (2016). The impact of data use professional development on student achievement. *Teach. Teach. Educ.* 60, 434–443. doi: 10.1016/j.tate.2016.07.005

Landwehr, N. (2011). "Thesen zur Wirkung und Wirksamkeit der externen Schulevaluation [theses on the impact and effectiveness of the external evaluation of schools]" in *Wirkungen und Wirksamkeit der externen Schulevaluation [Effects and Effectiveness of External School Evaluation]*. eds. C. Quesel, N. Husfeldt, N. Landwehr and P. Steiner (Bern, Switzerland: Hep der Bildungsverlag), 35–69.

Lankshear, C., and Knobel, M. (2004). *A Handbook for Teacher Research*. UK McGraw-Hill Education.

Legood, A., van der Werff, L., Lee, A., and Den Hartog, D. (2021). A meta-analysis of the role of trust in the leadership-performance relationship. *Eur. J. Work Organ. Psy.* 30, 1–22. doi: 10.1080/1359432X.2020.1819241

Lohse-Bossenz, H., Bloss, C., and Doerfler, T. (2022). Constructing multi-theory vignettes to measure the application of knowledge in ambivalent educational situations. *Front. Educat.* 7

Lorenz, R., Eickelmann, B., and Bos, W. (2016). High-stakes testing and educational governance–insights and innovations in European school systems. Special Issue Editorial. *J. Educat. Res. Online* 8, 5–9. doi: 10.25656/01:12416

Louis, K. S. (2007). Trust and improvement in schools. *Journal of Educational Change* 8, 1–24. doi: 10.1007/s10833-006-9015-5

Lüdtke, O., Marsh, H. W., Robitzsch, A., and Trautwein, U. (2011). A 2 × 2 times taxonomy of multilevel latent contextual models: accuracy-bias trade-offs in full and partial error correction models. *Psychol. Methods* 16, 444–467. doi: 10.1037/a0024376

Maier, U. (2010). Accountability policies and teachers' acceptance and usage of school performance feedback – a comparative study. *Sch. Eff. Sch. Improv.* 21, 145–165. doi: 10.1080/09243450903354913

Malin, J. R., Brown, C., Ion, G., van Ackeren, I., Bremm, N., Luzmore, R., et al. (2020). World-wide barriers and enablers to achieving evidence-informed practice in education: what can be learnt from Spain, England, the United States, and Germany? *Humanit. Soc. Sci. Communicat.* 7, 1–14. doi: 10.1057/s41599-020-00587-8

Mandinach, E. B. (2012). A perfect time for data use: using data-driven decision making to inform practice. *Educ. Psychol.* 47, 71–85. doi: 10.1080/00461520.2012.667064

Mandinach, E. B., and Gummer, E. S. (2016). What does it mean for teachers to be data literate: laying out the skills, knowledge, and dispositions. *Teach. Teach. Educ.* 60, 366–376. doi: 10.1016/j.tate.2016.07.011

Mandinach, E. B., Honey, M., Light, D., and Brunner, C. (2008). A conceptual framework for data-driven decision making. Data-Driven School Improvement: Linking Data and Learning, 13–31.

Mandinach, E. B., and Schildkamp, K. (2021). Misconceptions about data-based decision making in education: an exploration of the literature. *Stud. Educ. Eval.* 69:100842. doi: 10.1016/j.stueduc.2020.100842

Mark, M. M., and Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation* 10, 35–57. doi: 10.1177/1356389004042326

Marsh, J. A. (2012). Interventions promoting educators' use of data: research insights and gaps. *Teach. Coll. Rec.* 114, 1–48. doi: 10.1177/016146811211401106

McCluskey, G., Fry, D., Hamilton, S., King, A., Laurie, M., McAra, L., et al. (2021). School closures, exam cancellations and isolation: the impact of Covid-19 on young people's mental health. *Emot. Behav. Diffic.* 26, 46–59. doi: 10.1080/13632752.2021.1903182

McNeish, D., and Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivar. Behav. Res.* 51, 495–518. doi: 10.1080/00273171.2016.1167008

Merk, S., Groß Ophoff, J., and Kelava, A. (in press). *Rich Data – Poor in-Formation? Teacher's Perceptions of Mean Differences in Graphical Feedback from State-Wide Tests* Learning and Instruction.

Mincu, M. (2014). *Inquiry Paper 6: Teacher Quality and School Improvement – What is the Role of Research?*. Available at: https://www.bera.ac.uk/wp-content/uploads/2014/02/BERA-RSA-Interim-Report.pdf

Mintrop, R., and Zumpe, E. (2019). Solving real-life problems of practice and education leaders' school improvement mind-set. *Am. J. Educ.* 125, 295–344. doi: 10.1086/702733

Mintrop, R., Zumpe, E., and Baral, M. (2018). *Design-Based School Improvement and Research for Education Leaders. The SAGE Handbook of School Organization*. London: Sage Publications.

Muñoz, C., and Valenzuela, J. (2020). Demotivation in academic reading during teacher training. *J. Res. Read.* 43, 41–56. doi: 10.1111/1467-9817.12288

Muthén, L. K., and Muthén, B. O. (2017). *MPlus Version 8*, Los Angeles, CA: Muthén & Muthén.

Neubrand, M., Jordan, A., Krauss, S., Blum, W., and Löwen, K. (2013). "Task analysis in COACTIV: examining the potential for cognitive activation in German mathematics classrooms" in *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project*. eds. M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss and M. Neubrand (New York: Springer US), 125–144.

Office for Standards in Education. (2022). *School Inspection Handbook*. Available at: https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook#reaching-a-judgement-of-outstanding

Office for Standards in Education (2015). *The Annual Report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2014/15*, London: HM Government.

Penninckx, M., and Vanhoof, J. (2015). Insights gained by schools and emotional consequences of school inspections. A review of evidence. *Sch. Leadersh. Manag.* 35, 477–501. doi: 10.1080/13632434.2015.1107036

Perryman, J., Maguire, M., Braun, A., and Ball, S. (2018). Surveillance, Governmentality and moving the goalposts: the influence of Ofsted on the work of schools in a post-panoptic era. *Br. J. Educ. Stud.* 66, 145–163. doi: 10.1080/00071005.2017.1372560

Prenger, R., and Schildkamp, K. (2018). Data-based decision making for teacher and student learning: a psychological perspective on the role of the teacher. *Educ. Psychol.* 38, 734–752. doi: 10.1080/01443410.2018.1426834

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna R Foundation for Statistical Computing.

Renkl, A. (2022). Meta-analyses as a privileged information source for informing teachers' practice? *Z. Pädagog. Psychol.* 36, 217–231. doi: 10.1024/1010-0652/a000345

Richter, D., Böhme, K., Becker, M., Pant, H. A., and Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten. Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern. *Z. Pädagog. Psychol.* 60, 225–244. doi: 10.3262/ZP1402225

Rickinson, M., Sharples, J., and Lovell, O. (2020). "Towards a better understanding of quality of evidence use" in *Getting Evidence into Education: Evaluating the Routes to Policy and Practice*. ed. S. Gorard (Abingdon: Routledge), 219–233.

Rochnia, M., and Gräsel, C. (submitted). Can the utility value of educational sciences be induced based on a reflection example or empirical findings—Or just somehow? 7.

Rose, J., Thomas, S., Zhang, L., Edwards, A., Augero, A., and Rooney, P. (2017). *Research learning communities evaluation report and executive summary*. Available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Research_Learning_Communities.pdf

Schildkamp, K. (2019). Data-based decision-making for school improvement: research insights and gaps. *Educ. Res.* 61, 257–273. doi: 10.1080/00131881.2019.1625716

Schildkamp, K., Handelzalts, A., Poortman, C. L., Leusink, H., Meerdink, M., Smit, M., et al. (2018). *The Data Team Procedure: A Systematic Approach to School Improvement*, Cham, Switzerland: Springer.

Schildkamp, K., and Kuiper, W. (2010). Data-informed curriculum reform: which data, what purposes, and promoting and hindering factors. *Teach. Teach. Educ.* 26, 482–496. doi: 10.1016/j.tate.2009.06.007

Schildkamp, K., Lai, M. K., and Earl, L. (2013). *Data-Based Decision Making In Education. Challenges and Opportunities*. Dordrecht: Springer

Schildkamp, K., and Poortman, C. (2015). Factors influencing the functioning of data teams. *Teach. Coll. Rec.* 117, 1–42. doi: 10.1177/016146811511700403

Schildkamp, K., Poortman, C., Luyten, H., and Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *Sch. Eff. Sch. Improv.* 28, 242–258. doi: 10.1080/09243453.2016.1256901

Schratz, M., Wiesner, C., Rößler, L., Schildkamp, K., George, A. C., Hofbauer, C., et al. (2018). Möglichkeiten und Grenzen evidenzorientierter Schulentwicklung. *Nationaler Bildungsbericht Österreich* 2, 403–454. doi: 10.17888/nbb2018-2-10

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harv. Educ. Rev.* 57, 1–23. doi: 10.17763/haer.57.1.j463w79r56455411

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher* 31, 15–21. doi: 10.3102/0013189x031007015

Syring, M., Kleinknecht, M., Bohl, T., Kuntze, S., Rehm, M., and Schneider, J. (2015). How problem-based or direct instructional case-based learning environments influence secondary school pre-service teachers' cognitive load, motivation and emotions: a quasi-experimental intervention study in teacher education. *J. Educ. Hum. Dev.* 4, 115–129. doi: 10.15640/jehd.v4n4a14

Taysum, A. (2020). An innovative strategy to re-build trust in educational governance systems after 40 years of English destabilising rapid reforms. *J. Educat. Administr. History* 52, 145–161. doi: 10.1080/00220620.2019.1605341

Teddlie, C., and Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Res. Sch.* 13, 12–28.

Tennent, W. (2021). The assessment of reading comprehension in English primary schools: investigating the validity of the key stage 2 reading standard assessment test (SAT). *Education 3–13* 49, 481–494. doi: 10.1080/03004279.2020.1735472

UK Government (2022). Levelling up the United Kingdom. Executive Summary. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1095544/Executive_Summary.pdf

Van Gasse, R., Vanlommel, K., Vanhoof, J., and Van Petegem, P. (2016). Teacher collaboration on the use of pupil learning outcome data: a rich environment for professional learning? *Teach. Teach. Educ.* 60, 387–397. doi: 10.1016/j.tate.2016.07.004

Van Geel, M., Keuning, T., Visscher, A. J., and Fox, J.-P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *Am. Educ. Res. J.* 53, 360–394. doi: 10.3102/0002831216637346

van Geel, M., Keuning, T., Visscher, A., and Fox, J.-P. (2017). Changes in educators' data literacy during a data-based decision making intervention. *Teach. Teach. Educ.* 64, 187–198. doi: 10.1016/j.tate.2017.02.015

Vanhoof, J., Vanlommel, K., Thijs, S., and Vanderlocht, H. (2014). Data use by Flemish school principals: impact of attitude, self-efficacy and external expectations. *Educ. Stud.* 40, 48–62. doi: 10.1080/03055698.2013.830245

VanLeirsburg, P., and Johns, J. L. (1994). Teachers as readers. Literacy Research Report No. 18.

Vanlommel, K., Van Gasse, R., Vanhoof, J., and Van Petegem, P. (2017). Teachers' decision-making: data based or intuition driven? *Int. J. Educ. Res.* 83, 75–83. doi: 10.1016/j.ijer.2017.02.013

Voss, T. (2022). Not useful to inform teaching practice? Student teachers hold skeptical beliefs about evidence from education science. *Front. Educat.* 7:976791. doi: 10.3389/feduc.2022.976791

Voss, T., Zeeb, H., Dehmel, A., and Fauth, B. (2020). "Forschungsmethoden in der Lehrerinnen-und Lehrerbildung" in *Handbuch Lehrerbildung*. eds. C. Cramer, J. König, M. Rothland and S. Blömeke (Klinkhardt/UTB), 667–672.

Weiss, C. H. (1998). "Improving the use of evaluations: whose job is it anyway?" in *Advances in Educational Productivity. Evaluation Research for Educational Productivity*. eds. A. J. Reynolds and H. J. Walberg, vol. 7 (Greewich, CT: JAI Press Inc.), 263–276.

Wessels, I., Gess, C., and Deicke, W. (2019). "Competence development through inquiry-based learning" in *Inquiry-Based Learning–Undergraduate Research*, ed. H. A. Mieg (Berlin: Springer), 59–69.

Wiesner, C., and Schreiner, C. (2019). "Implementation, Transfer, Progression und Transformation: Vom Wandel von Routinen zur Entwicklung von Identität. Von Interventionen zu Innovationen, die bewegen. Bausteine für ein Modell zur Schulentwicklung durch Evidenz (en)" in *Praxistransfer Schul-und Unterrichtsentwicklung*. eds. C. Schreiner, C. Wiesner, S. Breit, P. Dobbelstein, M. Heinrich and U. Steffens (Münster: Waxmann), 79–140.

Wurster, S. (2016). *Evaluationsgeschützte Schul-und Unterrichtsentwicklung: Eine multiperspektivische Analyse von Vergleichsarbeiten, zentralen Abschlussprüfungen, Schulinspekion und interner Evaluation*. Berlin Humboldt-Universität zu Berlin.

Zeuch, N., Förster, N., and Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: results from tests and interviews. *Learn. Disabil. Res. Pract.* 32, 61–70. doi: 10.1111/ldrp.12126

Ziauddeen, N., Woods-Townsend, K., Saxena, S., Gilbert, R., and Alwan, N. A. (2020). Schools and COVID-19: reopening Pandora's box? *Public Health Pract.* 1:100039. doi: 10.1016/j.puhip.2020.100039

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 20 January 2023
DOI 10.3389/feduc.2023.974696

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University,
Germany

REVIEWED BY
Christina Wekerle,
University of Augsburg,
Germany
Theresa Krause-Wichmann (Wilkes),
Saarland University,
Germany
Eric Klopp,
Saarland University,
Germany

*CORRESPONDENCE
Swantje Tannert
✉ swantje.tannert@uni-erfurt.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 21 June 2022
ACCEPTED 02 January 2023
PUBLISHED 20 January 2023

CITATION
Tannert S, Eitel A, Marder J, Seidel T,
Renkl A and Glogger-Frey I (2023) How can
signaling in authentic classroom videos support
reasoning on how to induce learning
strategies?
Front. Educ. 8:974696.
doi: 10.3389/feduc.2023.974696

How can signaling in authentic classroom videos support reasoning on how to induce learning strategies?

Swantje Tannert[1]*, Alexander Eitel[2], Johanna Marder[3], Tina Seidel[4], Alexander Renkl[5] and Inga Glogger-Frey[1]

[1]Department of Psychology, University of Erfurt, Erfurt, Germany, [2]Department of Psychology, Justus Liebig University Giessen, Giessen, Germany, [3]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany, [4]TUM School of Education, Technical University Munich, Munich, Germany, [5]Institute for Psychology, University of Freiburg, Freiburg, Germany

Classroom videos are a viable means to implement evidence-informed reasoning in teacher education in order to establish an evidence-informed teaching practice. Although learning with videos relieves pre-service teachers from acting in parallel and might reduce complexity, the material still poses higher cognitive load than written text vignettes or other traditionally used static material. In particular, the information they deliver is transient and can, therefore, easily be missed. *Signaling* can guide learners' attention to central aspects of a video, thereby reducing cognitive load and enhancing learning outcomes. In the current project, pre-service teachers acquired scientific knowledge about learning strategies and their promotion in a computer-based learning environment. We explored the effect of different arrangements of signaling in classroom video-examples on conceptual knowledge and the reasoning-component of professional vision. Therefore, we conducted a set of two studies with 100 student teachers including two signal arrangements in order to investigate how signaling can help learning to reason about classroom videos. In addition, we varied if participants received information on the use of signals in advance (informed) or not (uninformed). We measured conceptual knowledge by asking participants what they knew about self-regulation strategies. Additionally, we assessed reasoning by asking participants to notice sequences in a video where teachers induced learning strategies, and to reason in what respect the observed behavior was useful to induce the strategy. Uninformed signaling did not affect the acquisition of conceptual knowledge and reasoning. Informed signaling led to significantly better conceptual knowledge than uninformed signaling. It is argued that the signal-induced extraneous load exceeded the load reduction due to the signal's selection advantage in the uninformed conditions. In a third, exploratory study, nine participants were interviewed on the perception of different signals and indicated that spotlight and zoom-in signals foster processing of classroom videos.

KEYWORDS

signaling, multi-media design, self-regulated learning, cognitive load, evidence based teaching, evidence-oriented practice, evidence-informed teaching

1. Introduction

Evidence-informed practice becomes more and more a standard not only for good medicine (Sackett et al., 1996; Slavin, 2002; Diery et al., 2020, 2021; Knogler et al., 2022) but also for good teaching (Hammersley, 2005; Dagenais et al., 2012; Yeh and Santagata, 2015). This approach implies basing actions on theoretically or empirically founded evidence instead of relying on tradition or habit. Following a flexible view of evidence-informed practice (Biesta et al., 2011; Biesta, 2017), teaching should result from a diagnostic process that evaluates different possible interventions based on empirical evidence on the one hand side and situational or personal conditions on the other hand side. In this view, evidence-informed practice does not mean teaching from a cookbook in the sense of simple and unconditional application of a "scientific rule" like it is proposed by Slavin (2008). Rather, empirical evidence and the resulting pedagogical knowledge needs to become a foundation for an adaptive teaching that takes individual situations and case-specificities well into account. This flexible behavior needs to be based on a good diagnostic competence of teachers in the sense of professional vision (Sherin and van Es, 2009). Therefore, it is important to let pre-service teachers practice their professional vision in real classroom situations. In this regard, learning from authentic video material has become an important part of teacher education (Ball, 2000; Santagata et al., 2005; Spiro et al., 2007; Ball and Forzani, 2009; Blomberg et al., 2013). Video examples give insight into actual classroom situations, thereby enabling pre-service teachers to integrate theoretical knowledge with real-life teaching problems. However, even in small sequences, video material can be too complex and overwhelming, especially for non-expert viewers (Betrancourt and Tversky, 2000; Ayres and Paas, 2007). The resulting cognitive load, that is, the amount of working memory resources (Sweller, 1988) required to process the material, then tends to exceed the available capacity and to impair learning.

The present series of studies explores the potential of an instructional method that has been shown to reduce cognitive load: signaling. Signaling has hardly been researched in the context of dynamic material such as classroom video examples. First results, however, seem promising (Alpizar et al., 2020). In the present investigation the signaling method is utilized to facilitate processing of classroom video examples within a learning environment that teaches pre-service teachers how to induce learning strategies.

1.1. Reasoning

Teaching is a rather complex process that requires teachers to integrate theoretical knowledge with actual situations in order to diagnose ongoing processes and find appropriate actions to foster the learning of the students. Therefore, teachers need to choose the theoretically relevant situational information to guide their attention at (noticing) and make sense of the ongoing processes by relating the events to their theoretical knowledge (reasoning). The combined ability of noticing the relevant aspects and reasoning is called professional vision (van Es and Sherin, 2008; Stürmer et al., 2013). It is important to note that both processes are top-down in that they require theoretical (pedagogical) knowledge to be successfully performed (Grossman and McDonald, 2008), that is, both processes are evidence-informed. In the present article, the concept of professional vision is not used to understand student behavior but rather to relate teacher behavior that is their attempts to foster self-regulation in students to the theoretical concept of self-regulated learning strategies. Studies that aim at fostering professional vision typically require evidence informed noticing as a first step (van Es and Sherin, 2002, 2008; Sherin and Han, 2004; Sherin, 2007; Kersting et al., 2010; Santagata and Guarino, 2011; Stürmer et al., 2013). However, this might be rather demanding and deplete cognitive resources before reasoning can happen. There have been little attempts so far to relieve (pre-service) teachers from the cognitive load associated with noticing in order to give them the opportunity to focus on evidence-informed reasoning. Therefore, the present study aims at establishing this second component of professional vision by supporting the noticing process with content knowledge that is knowledge on learning strategy induction and signaling.

1.2. Classroom videos as a training tool

Classroom videos are a viable means to contextualize theoretical pedagogical knowledge in order to build up and practice reasoning. Additionally, they offer some advantages compared to acting in real classrooms. First, pre-service teachers can focus on diagnostic processes without having to act in parallel (Sherin, 2014). Second, whole teaching situations can be broken down into smaller, easily manageable sequences (Le Fevre, 2003) in order to reflect on them in an evidence-informed manner. Third, the material can be repeated to really work out the gist (Spiro et al., 2007). Thus, classroom videos seem promising for supporting reasoning abilities and thereby evidence-informed reasoning in teachers. However, the processing of videos might induce high cognitive load, especially in novices (Betrancourt and Tversky, 2000; Ayres and Paas, 2007). To make the high educational potential of classroom videos even more usable, overcoming these processing difficulties by a reduction of extraneous cognitive load is the aim of the present study.

1.3. Cognitive load

Cognitive load theory (Sweller, 1988) has proposed three types of cognitive load, namely intrinsic, extraneous and germane cognitive load that compete for the cognitive resources. While germane load is rather desirable and intrinsic load is immanent in matters (complexity of content), extraneous cognitive load is elicited by redundant or irrelevant information, additional stimuli, perceptually overwhelming learning environments or material and any kind of disturbances. Thus, it is this last component that needs our attention when trying to reduce the overload of (pre-service) teachers while working with classroom videos.

Those videos are known to offer options for illustrating complex issues with a high element interactivity while at the same time inducing a high amount of extraneous load (Betrancourt and Tversky, 2000; Ayres and Paas, 2007). Additionally, they are not effective in themselves. Rather, they need to be embedded in an instructional context (Blomberg et al., 2013; Seidel et al., 2013).

The cognitive theory of multimedia learning (CTML; Mayer, 2014) makes three assumptions when learning with multimedia materials, such as videos. First, multimedia material is processed via two separate channels for visual and auditory information, respectively (Baddeley, 1992). Second, each of the channels has its own limited capacity (Baddeley, 1992) and third, learning form multimedia material is effective, when learners actively engage in selection, organization and integration of the material. However, these processes are hampered when extraneous cognitive load binds too many cognitive resources (Mayer and Fiorella, 2014). Thus, the reduction of extraneous cognitive load is necessary, due to the limited capacity of the cognitive system (Chandler and Sweller, 1991, 1992; Paas, 1992; Paas and Van Merriënboer, 1994). Additionally, according to the third assumption, it might be beneficial to support selection of the important content for further processing (De Koning and Jarodzka, 2017).

1.4. Signaling

One promising way to deal with extraneous cognitive load is signaling, that is, the use of cues to help learners selecting and organizing the relevant content (Mayer, 2014). A growing literature shows advantages in learning with signaled in comparison to non-signaled static (Tversky et al., 2008; Schneider et al., 2018) and dynamic (De Koning et al., 2007; Moreno, 2007; Alpizar et al., 2020) materials. In static materials, signals, such as bold words, arrows, or different colors, point to relevant information or connect graphics and text (Mautone and Mayer, 2001; Richter et al., 2016). Furthermore, headlines and paragraphs structure the learning material. Mautone and Mayer (2001), for example, used headings, a summary, connecting words, boldface and italic for important words in their static material. So far, signaling in dynamic material has been relatively similar to that in static material. Analogously to accentuating important words or adding connecting words, Wang et al. (2020) used visual accentuation of certain areas in their animation videos as well as textual marking. Boucheix and Guignard (2005) showed a green dot on gearing wheels and arrows pointing at the relevant places on the screen. Additionally, they showed short sentences saying things like*: Look at the two wheels A and B and compare their speeds*. De Koning et al. (2007) visually enhanced the heart valves in an animation on heart function by blurring the surrounding heart and coloring the valves in red and blue.

However, most studies focus on material that has an instructional rather than exemplary character (e.g., Mautone and Mayer, 2001; overview: Alpizar et al., 2020). Thus, oftentimes signaling has been implemented into animations or other instructional visual material. There is little evidence so far on the advantages of signaling in material that exemplifies theoretical concepts or principles, such as authentic classroom videos. In addition, this type of material is different from the ones described above in that the to-be-learned content is not written or otherwise drawn onto the screen. Rather, observers are asked to infer, for example, reasons for a certain behavior from a certain situation. Thus, the relevant information is not visually printed onto the screen and therefore cannot simply be stressed by bold type letters underlining relevant words or paagraphs as a signal. Rather, signals in classroom videos might cue a certain timespan containing relevant information and or point out where in the classroom relevant behavior takes place. More complex events might additionally require verbal signals like for example "*interpretation*" to stress that a teacher in the video does not simply refer to a perceived behavior but already interpreted the reasons.

Although signaling has been shown to be advantageous in some studies, it is important to note that signals need to be appropriate and easy to process, because each new information within a scene might induce additional cognitive load (Sweller, 1988; Mayer, 2014). However, signaling might still be advantageous in authentic classroom videos because this kind of material is cognitively very demanding due to the complexity of real classroom situations. Martin et al. (2022) already gained positive effects of segmenting and self-explanation prompts, that is, methods aiming at a reduction of cognitive load, during professional-vision training.

While signaling mainly addresses the reduction of extraneous cognitive load, variables associated with other types of cognitive load may still influence the effect of signaling. In this regard, it is important to consider prior knowledge as it relates to the intrinsic load the material poses on a learner. Learners with higher prior knowledge may profit less from signaling than those with lower prior knowledge because their intrinsic cognitive load is lower and they might be less in danger of being overloaded (Van Merrienboer and Sweller, 2005; Mayer and Fiorella, 2014; Richter et al., 2016, 2018; Alpizar et al., 2020). For those learners, the threshold for positive effects of signaling is supposedly higher compared to learners, who are already heavily challenged by the high intrinsic load. When signals are presented to those high prior knowledge learners, they might not profit from them and may even feel disturbed or irritated. Additionally, it is argued that processing external cues that point at the relevance of certain information (signals) while processing relevance indicators that are derived from prior knowledge (existing knowledge structure) imposes an additional load on working memory. This effect has already been demonstrated in several studies on signaling in static material. It is a case of an expertise reversal effect (Kalyuga, 2007, 2008, 2014).

1.5. Research questions and overview of the present studies

The current work focuses on the impact of different arrangements of signaling in classroom videos on extraneous cognitive load, conceptual knowledge and reasoning. We conducted three small studies using three approaches to signaling in order to reduce the perceived extraneous cognitive load, and to improve conceptual knowledge and reasoning. Following the above argumentation on the expertise reversal effect (Kalyuga, 2007, 2008, 2014); we also investigated whether prior knowledge moderates the possible effects of the different signaling methods. Participants worked with an environment on self-regulated learning with a focus on cognitive as well as metacognitive learning strategies. The learning goal was to enable pre-service teachers to reason about how certain learning strategies are induced during class. The learning environment used authentic classroom videos presented with versus without signaling. The studies built on each other, that is, each study focused on an open question or resulting idea of the previous one. For the most part, materials were the same across all studies. Studies only differed in the respective experimental variation, the approach to signaling. Study 1 investigated the effect of a keyphrase signaling procedure, which was supposed to stress the relevant information in the instructions of the teacher. Study 2 built on the results of Study 1 and therefore contained a less demanding combination of a short tone and a red frame as signals. Based on the finding that learning is better when material and task are known in advance (Paas and Van Merriënboer, 1994; Kirschner et al., 2006; Sweller et al., 2007; Schwonke et al., 2013), Study 2b introduced an information on the utility and use of signals. The different experimental variations are reported in the respective methods section. Finally, we conducted a third exploratory survey study where participants should indicate how they experienced different signals within classroom video examples embedded in the same learning environment as in the other studies. The signals focused on information selection.

2. Study 1 – Key phrase signaling

The first study compared the conceptual knowledge and reasoning of participants who learned with simple unsignaled authentic classroom videos, with learning outcomes of participants who learned with classroom videos including a key phrase signaling. That is, we asked whether making key contents of the video permanent in the form of written text would relieve participants from extraneous cognitive load and lead to enhanced conceptual knowledge and reasoning compared to no signals. We chose key phrases as our first signal in order to stress the time as well as the content, that is, the spoken instructions of the teacher, that needed to be attended. Prior studies have found that presenting a whole transcript results in worse learning due to redundancy (Sweller, 1988; Kalyuga et al., 1999; Mayer, 2014). However, there are studies yielding good learning outcomes with key phrases instead of paraphrases (Moreno and Mayer, 2002; Mayer and Johnson, 2008). So, we decided to do the same. We expected our signals to help participants focusing on the relevant content, thereby being less distracted by irrelevant stimuli. This should result in a lower extraneous cognitive load and a better learning outcome.

2.1. Hypotheses

> *H1*: Key phrase *signaling* reduces *extraneous cognitive load*. (Cognitive load hypothesis)
>
> *H2*: Key phrase *signaling* fosters *conceptual knowledge* while working with authentic classroom videos. (Signaling knowledge hypothesis)
>
> *H3*: Key phrase *signaling* fosters *reasoning* while working with authentic classroom videos. (Signaling reasoning hypothesis)
>
> *H4: Prior conceptual knowledge* moderates key phrase *signaling* effects, that is *signaling* reduces *extraneous cognitive load* to a greater extent when *prior knowledge* is low compared to when prior knowledge is high. (Prior knowledge hypothesis)
>
> *H5: Prior reasoning* moderates key phrase *signaling* effects, that is signaling reduces *extraneous cognitive load* to a greater extent when *prior reasoning* is low compared to when *prior reasoning* is high. (Prior reasoning hypothesis)

2.2. Method

The experiment was done with videos within a learning environment. An overview of the procedure can be found in Appendix A.

2.2.1. Sample and design

Fifty-seven student teachers (M_{age} = 22.72 years, SD = 2.52 years, semester: M = 4.28, SD = 2.97) took part in this study. Participants were randomly assigned to either the key phrase signaling group (N = 30) or the no signaling group (N = 27).

2.2.2. Material

The participants worked on a computer. The materials were very similar across the present studies and will be described in detail in the following sections.

2.2.2.1. Learning environment

The study was conducted using an example-based learning environment (Renkl, 2014) with the following structure: The topic of the learning environment was *strategies of self-regulated learning* (Weinstein and Mayer, 1986). The learning phase included a theoretical and an applied part. First, participants received general theoretical knowledge on *learning strategies* and a tree diagram on the distinction between *metacognitive* and *cognitive strategies* of self-regulated learning (Glogger et al., 2013). This diagram could be accessed at any time within the environment by clicking the *help* button (see Appendix B). A description of each of the learning strategies was shown after the first presentation of the diagram. In the applied part of the learning phase, the participants watched an example video with a focus on the learning-strategy induction and were presented with the same questions that were used to assess reasoning in the pretest and posttest. In contrast to the test phases, learners received an example solution, that is, answers to the questions of the practice video (Appendix E).

2.2.2.2. Videos and experimental variation (signaling)

The presented videos were authentic sequences from school lessons with a duration of 30 to 90 s. They showed a teacher prompting cognitive or metacognitive learning strategies in students (Appendix D).

Depending on the experimental condition, the videos in the learning phase were presented without or with key phrase signaling. MOOC-courses of universities with high reputation often print the spoken text next to the respective video. The text appears as soon as it is spoken. Analogously, the signaling group in this study saw key phrases of the text, spoken by the teacher, as written text on screen next to the classroom scene, whenever the sequence was relevant to the task. That is, key phrases popped up as soon as the teacher said this phrase and the text remained on the screen until replaced by the next key phrases (e.g., *Six pictures…each representing one of these rights …assign the terms to the civil rights…make a list that contains the picture on the one side and the respective text on the other*). This procedure was intended to make the transient information more permanent. As redundant information is known to impair learning in certain applications (redundancy effect, Kalyuga et al., 1999), we only used key phrases. Such phrases have been shown to foster knowledge acquisition in short instructions (Moreno and Mayer, 2002; Mayer and Johnson, 2008).

2.2.3. Instruments

2.2.3.1. Prior knowledge

Conceptual prior knowledge was assessed by the following *self-rated item* and an *open question*: *"In my lectures, cognitive learning strategies were addressed" (1- not at all to 5-very detailed).* "*Which cognitive learning strategies do you know? Please describe concisely.*" The answer to this question was rated on a five-point scale as in Glogger-Frey et al., 2015 (see Coding scheme 1, Appendix F). The prior level of reasoning related to self-regulated learning was assessed by presenting participants with a short video of a classroom situation and asking them to indicate, whether a learning strategy was induced by the teacher and, if yes, which one and how (Appendix D). Answers were rated on a five-point scale using Coding scheme 3 (Appendix F). The answers of 12 participants (21.05%) were rated by two independent coders in order to determine inter-coder reliability (prior conceptual knowledge: $ICC(2,2) = 0.95$; prior reasoning: $ICC(2,2) = 0.91$).

2.2.3.2. Learning outcomes

We assessed learning outcomes after the learning phase by theoretical questions (conceptual knowledge) and a video task (reasoning). An example question for conceptual knowledge was: "*Please describe shortly, which cognitive learning strategies you got to know in this learning environment.*" Again, answers were rated by independent raters on a five-point scale following Coding scheme 2 (Appendix F). For reasoning participants had to watch a short classroom video. We asked participants to connect the teacher behavior in that video with their theoretical knowledge on learning strategies. Therefore they were asked to, first, fill in a gap text (Appendix D, Coding scheme 1), and second, indicate the used learning strategy and give reasons for their choice (self-description, Hilbert et al., 2008, Appendix D, Coding scheme 2), that is name the concrete behavior of the teacher that induces the respective learning strategy in the students. We rated participants' conceptual knowledge and their reasoning (Coding schemes 2 and 3, Appendix F) based on the SOLO-taxonomy (Biggs and Collis, 1982). Answers for conceptual knowledge were rated on a 5-point scale ranging from 1 (*no conceptual knowledge*) to 5 (*very clear conceptual knowledge*; Biggs and Collis, 1982). *Conceptual knowledge* was high, when participants not only named the respective learning strategy but also proved to know the nature of this strategy and to be able to relate certain behaviors and tasks to this strategy. Reasoning was rated on a 5-point-scale ranging from 1 (*no evidence-informed reasoning*) to 5 (*very good evidence-informed reasoning*). A good performance indicated that the participants had successfully used their theoretical knowledge to select, categorize and interpret relevant information within the videos and thus had demonstrated good *reasoning*. Again, the answers of 12 participants were coded by two independent coders in order to determine inter-coder reliability (conceptual knowledge: $ICC(2,2) = 0.94$; reasoning: $ICC(2,2) = 0.89$).

2.2.3.3. Cognitive load questionnaire

Cognitive load was assessed by eight items on an 11-point Likert scale (Leppink et al., 2013, 2014) ranging from 0 (*no cognitive load*) to 10 (*very high cognitive load*). The cognitive load questions referred to the three different load types, that is germane load (three items, e.g., *The learning environment has really improved my understanding of learning strategy induction.*), intrinsic load (two items, e.g., *The content of the learning environment was very complex.*), and extraneous load (three items, e.g., *The instructions and explanations were full of unclear language.*). Leppink et al. (2013) report a high internal consistency (intrinsic load: $\alpha = 0.893$, extraneous load: $\alpha = 0.785$, germane load: $\alpha = 0.947$) of the cognitive load scales. Our own consistencies are a bit lower: intrinsic load: $\alpha = 0.82$, extraneous load: $\alpha = 0.61$, germane load: $\alpha = 0.86$. For the present study only extraneous cognitive load was considered.

2.2.4. Procedure

After a demographic questionnaire, prior knowledge was assessed. Then, the learning phase started, where participants worked on the learning environment including the practice videos with example solutions. These solutions (Appendix E) were given to the participants after they had tried to answer the questions on self-regulation strategy induction on their own. The example solutions served as feedback. This helped participants learning to notice and to reason. After this, the test phase began where participants received the instructions for the posttest and the test videos, watched the videos and answered the test questions for conceptual knowledge and reasoning. Finally, participants answered the cognitive load questionnaire (Appendix A).

2.2.5. Analysis

Significance level in all analyses was $\alpha = 0.05$. All variables were z-standardized so that regression-coefficients are standardized β-coefficients that can be interpreted as effect sizes (small: <0.2; medium: <0.5; large: ≥0.5; Acock, 2014). Descriptive data of all experiments can be found in Tables 1–3. There were no statistically significant differences between the signaling and no signaling group regarding conceptual prior knowledge, $\beta = -0.05$, $F(1, 56) = 0.034$, $p = 0.86$, and reasoning, $\beta = 0.07$, $F(1, 56) = 0.072$, $p = 0.79$. To test hypothesis 1, we did a regression of extraneous cognitive load on signaling. For hypothesis 2 we conducted a regression analysis, regressing from conceptual knowledge on signaling. The same was done for hypothesis 3 with reasoning as dependent variable. In order to test hypothesis 4 we did the same regressions as for H2 and H3 with the additional factor *prior knowledge* and the interaction term of prior knowledge and signaling.

2.3. Results

H1: **Cognitive load hypothesis.** We hypothesized that *signaling* reduces *extraneous cognitive load*. There were no significant differences in *extraneous cognitive load* between the *signaling* group and the *no signaling group*, $\beta = 0.14$, $F(1, 56) = 0.263$, $p = 0.61$.

H2: **Signaling knowledge hypothesis.** The mean posttest score for *conceptual knowledge* was $M = 3.74$, $SD = 0.87$, meaning a medium to clear conceptual understanding.

H3: **Signaling reasoning hypothesis.** The posttest score for reasoning was $M = 3.72$, $SD = 0.72$. The regression weight of *signaling* was not significant for *reasoning*, $\beta = -0.45$, $F(1, 56) = 3.101$, $p = 0.09$.

H4: **Prior knowledge hypothesis.** Mean *prior conceptual knowledge* was $M = 1.97$, $SD = 1.32$, The interaction of *prior conceptual knowledge* and *signaling* did not show a significant effect on *conceptual knowledge* $\beta = -0.16$, $F(2, 55) = 0.366$, $p = 0.55$. Thus, the effect of key phrase signals on the learning outcome was not influenced by *prior conceptual knowledge* of the participants.

H5: **Prior reasoning hypothesis.** Mean *prior reasoning* was $M = 2.34$, $SD = 1.28$. There was no interaction effect of *signaling-group* by prior reasoning on *posttest reasoning*, $\beta = -0.15$, $F(2, 55) = 0.345$, $p = 0.56$. Thus, the relationship of signaling type and posttest reasoning was not moderated by prior reasoning.

2.4. Discussion

The first study showed no significant effects of signaling on conceptual knowledge, reasoning, or extraneous cognitive load. This result indicates that signaling in form of key phrases next to classroom videos neither fostered nor impaired video processing and learning. In order to understand this effect, we will have a closer look at the used signals. Important video sequences were accompanied by a written copy of the spoken content (key phrases), which was presented side to side to the video. Even if we kept the key phrases short, we assume that we have found a redundancy effect. That is, the redundant information of the key phrases in auditory and visual (written) form was suboptimal for learning, because it created cognitive load in addition to offering additional relevant information (Chandler and Sweller, 1991; Van Gog et al., 2008). Although the text in our studies was still informative for selection of information, the actual content just reflected the spoken text. Due to this redundancy, it might have created extraneous load. Because the videos were rather short, signals might not have had the chance to considerably reduce cognitive load. In sum, the amount of load induced by processing the key phrase signals might have been

 10.3389/feduc.2023.974696

TABLE 1 Means (standard deviations) of conceptual prior knowledge (range: 1–5) and prior reasoning (range: 1–5) in all studies in total and separately for the signaling conditions.

	Total		No signaling		Signaling		Informed signaling	
	Conceptual	Reasoning	Conceptual	Reasoning	Conceptual	Reasoning	Conceptual	Reasoning
Study 1	1.97 (1.32)	2.34 (1.28)	2.00 (1.18)	2.30 (1.33)	1.94 (1.46)	2.39 (1.33)		
Study 2a	1.76 (1.03)	1.09 (0.29)	1.80 (1.45)	2.91 (0.61)	1.72 (0.96)	1.17 (0.38)		
Study 2b	1.61 (0.95)	1.14 (0.35)	1.80 (1.45)	2.91 (0.61)	1.72 (0.96)	1.17 (0.38)	1.18 (0.40)	1.27 (0.47)

TABLE 2 Means (standard deviations) of cognitive load (range 0–10) in all studies in total and separately for the signaling conditions.

Study	Total	No signaling	Signaling	Informed signaling
S1	1.49 (1.46)	1.38 (1.43)	1.58 (1.50)	
S2a	2.11 (1.33)	2.21 (1.47)	2.03 (1.23)	
S2b	2.19 (1.39)	2.21 (1.47)	2.03 (1.23)	2.44 (1.60)
Total		1.88 (0.94)	1.89 (0.99)	2.44 (1.60)

equal to the amount of load that was reduced due to the selection support. The analyses, conducted on extraneous cognitive load, show no significant differences. This does not necessarily mean that there was no modulation of extraneous load by the signaling procedure. Rather, we assume, that the key phrase signaling took some extraneous load away by focusing participants on the relevant sequences but at the same time added some extraneous load by adding redundant information. However, it seems reasonable to further investigate the relation of cognitive load and learning outcomes with a more fine-grained measure that captures the material-induced load separately from the signal-induced load.

To actively reduce signal-induced load in a first step, we chose a different approach to signaling in the subsequent study. Study 2a introduced a combination of two signals that contained less redundant information and was more subtle, and therefore prone to reduce extraneous cognitive load.

3. Study 2a – Beep and frame

In order to relieve working memory by supporting the selection of relevant information, while *not* straining working memory by redundancy, this study was conducted with a combination of two very subtle signals. The signals were not related to the spoken content. We used a tone, indicating the start of a relevant scene, and a red frame around the screen for the whole duration of the relevant scene. We expected the signals to be subtle enough now to help participants selecting the relevant content without adding extraneous load.

3.1. Hypotheses

> *H1*: Frame-tone *signaling* reduces *extraneous cognitive load.* (Cognitive load hypothesis)
>
> *H2*: Frame-tone *signaling* fosters *conceptual knowledge* while working with authentic classroom videos. (Signaling knowledge hypothesis)
>
> *H3*: Frame-tone *signaling* fosters *reasoning* while working with authentic classroom videos. (Signaling reasoning hypothesis)
>
> *H4*: *Prior conceptual knowledge* moderates frame-tone *signaling* effects, that is signaling reduces *extraneous cognitive load* to a greater extent when *prior knowledge* is low compared to when *prior knowledge* is high. (Prior knowledge hypothesis)
>
> *H5*: *Prior reasoning* moderates frame-tone *signaling* effects, that is *signaling* reduces *extraneous cognitive load* to a greater extent when *prior conceptual knowledge* is low compared to when *prior conceptual knowledge* is high. (Prior reasoning hypothesis)

3.2. Method

3.2.1. Sample and design

Thirty-three student teachers (mean age: $M = 22.5$ years, $SD = 3.0$ years) participated for the chance to win a voucher for a bookstore. Seventy percent had previously completed an internship in a pedagogical setting. Eighteen participants were randomly assigned to the signaling group and 15 participants were randomly assigned to the no signaling group. The independent variables were prior knowledge and signaling (signaling, no signaling). The dependent variables were again conceptual knowledge, reasoning and cognitive load.

3.2.2. Material

The material was identical to the first study with the following differences: Instead of printing the spoken text on the screen during relevant sequences, signaling consisted of a short tone (frequency: 1 kHz, duration: 250 ms) that ended 300 ms before the start of the relevant sequences and was followed by a red frame around the scene for the whole duration of the relevant sequences (Appendix C).

3.2.3. Instruments

We used the same instruments that were used in the first study. The internal consistency of the cognitive load scales was: intrinsic load: $\alpha = 0.91$, extraneous load: $\alpha = 0.71$, germane load: $\alpha = 0.92$. Again, we only used the extraneous cognitive load scale.

3.2.4. Procedure

The procedure was identical to that of study one. Twenty percent of the open answers were double-coded, $ICC = 0.971$.

3.2.5. Analysis

Significance level in all analyses was $\alpha = 0.05$. All variables were z-standardized so that regression-coefficients are standardized β-coefficients that can be interpreted as effect sizes (small: <0.2; medium: <0.5; large: ≥0.5; Acock, 2014). Descriptive data of all

TABLE 3 Means (standard deviations) of conceptual knowledge (range: 1–5) and reasoning (range: 1–5) in all studies in total and separately for the signaling conditions.

	Total		No signaling		Signaling		Informed signaling	
	Conceptual	Reasoning	Conceptual	Reasoning	Conceptual	Reasoning	Conceptual	Reasoning
E1	3.74 (0.87)	3.72 (0.72)	3.85 (0.87)	3.89 (0.62)	3.63 (0.87)	3.56 (0.78)		
E2a	2.84 (0.85)	2.91 (0.61)	2.89 (0.87)	2.88 (0.65)	2.80 (0.86)	2.93 (0.59)		
E2b	2.98 (0.72)	2.90 (0.56)	2.89 (0.87)	2.88 (0.65)	2.80 (0.86)	2.93 (0.59)	3.42 (0.84)	2.86 (0.44)

experiments can be found in Tables 1–3. Prior conceptual knowledge, $\beta = 0.09$, $F(1, 32) = 0.045$ $p = 0.81$ and prior reasoning, $\beta = 0.57$, $F(1, 32) = 2.819$, $p = 0.10$. To test hypothesis 1, we did a regression of extraneous cognitive load on signaling. For hypothesis 2, we conducted a regression from conceptual knowledge on signaling. In order to test hypothesis 3, we regressed reasoning on signaling and for hypothesis 4 and 5 we did the same two regressions with the additional factor prior knowledge and the interaction term of prior knowledge and signaling.

3.3. Results

H1: **Cognitive load hypothesis.** *Extraneous cognitive load* was descriptively higher in the *no signaling*, $M = 2.21$, $SD = 1.47$, compared to the *signaling* condition, $M = 2.03$, $SD = 1.23$. Signaling and no signaling group did not significantly differ in extraneous cognitive load, $\beta = -0.14$, $F(1, 32) = 0.152$, $p = 0.70$.

H2: **Signaling knowledge hypothesis.** The main effect of *frame-tone signaling* on *conceptual knowledge* did not reach significance, $\beta = -0.11$, $F(1, 32) = 0.093$, $p = 0.76$.

H3: **Signaling reasoning hypothesis.** The main effect of *signaling* on *reasoning* did not reach significance, $\beta = 0.08$, $F(1, 32) = 0.048$, $p = 0.83$.

H4: **Prior knowledge hypothesis.** The interaction of frame-tone *signaling* and *prior conceptual knowledge* was not significant for *conceptual knowledge*, $\beta = -0.18$, $F(2, 31) = 0.223$, $p = 0.64$.

H5: **Prior reasoning hypothesis.** The interaction of *signaling* and *prior reasoning* could not be estimated for *reasoning* due to a lack of variance.

3.4. Discussion

There was no significant effect of signaling with tone and red frame on extraneous cognitive load, conceptual knowledge, or reasoning. Thus, frame-tone signaling was still not beneficial as a design principle to foster learning. To further support the usability of the signals, we opted at expanding this setting by an additional condition, where participants received information on how to use the signals. Previous studies suggest, that extraneous cognitive load during tasks is reduced, when information on the task can be processed in advance (Paas and Van Merriënboer, 1994; Kirschner et al., 2006; Sweller et al., 2007; Schwonke et al., 2013). Thus, signal-induced extraneous cognitive load might be smaller, when learners get to know how to work with the signals before the actual video task. Consequently, Study 2b included an additional condition, which contained the same learning environment that was used in the signaling condition of Study 2a with the same signals. However, the participants received not only the tone and frame to point out relevant sequences but also an additional information on why and how to use signals in advance (informed signaling).

4. Study 2b – Informed beep and frame

For Study 2b, we collected data of one additional experimental group to compare with the groups of Study 2a. It focused on the effect of giving participants instructional information on the signaling method in order to prepare participants for proper use and thereby reduce extraneous cognitive load (informed signaling). Therefore, the setting of Study 2b was the same as the one in Study 2a, except for an additional information on signaling. To further investigate the idea that extraneous load is induced by irritation in uninformed signaling, we also compared extraneous load in informed and uninformed participants. We expected a reduced extraneous load in the informed participants because their resources would not be strained by a signal-induced irritation.

4.1. Hypotheses

H1: *Extraneous cognitive load* is higher in the *uninformed* frame-tone *signaling* group than in the *informed* frame-tone *signaling* group. (Informed cognitive load hypothesis)

H2: *Informed* frame-tone *signaling* leads to better *conceptual knowledge* while working with authentic classroom videos than *uninformed* frame-tone signaling. (Informed signaling knowledge hypothesis)

H3: *Informed* frame-tone signaling leads to better *reasoning* while working with authentic classroom videos than *uninformed* frame-tone signaling. (Informed signaling reasoning hypothesis)

4.2. Method

In addition to the signaling group ($N = 18$) and the no signaling group ($N = 15$) of Study 2a, we included an informed signaling group ($N = 11$). Twenty percent of the open answers were double coded, ICC = 0.962.

4.2.1. Sample and design

Eleven additional student teachers took part in this study. All additional students were assigned the informed signaling condition. This led to a total sample of 44 students (77% female, age, $M = 22.73$ years, $SD = 3.29$ years). Seventy-three percent reported to have completed a pedagogical internship.

4.2.2. Material

The material was the same as in Study 2a. Additionally, participants received the following information in advance of the signaled videos: "*In the following, you will see some example videos with teachers inducing learning strategies and students who implement them. The sequences, where the induction or implementation becomes especially salient, are particularly emphasized. How? You will hear a signal-tone and directly afterwards you will see a relevant sequence that is framed in red.*"

4.2.3. Instruments

The instruments were the same as in Study 2a. The cognitive load scale yielded slightly different internal consistencies due to the 11 additional participants (intrinsic load: $\alpha = 0.89$, extraneous load: $\alpha = 0.72$, germane load: $\alpha = 0.91$).

4.2.4. Procedure

The procedure was the same as in Study 2a except for one change: participants received the aforementioned instruction on the use of signals in advance of the video-presentation.

4.2.5. Analysis

Significance level in all analyses was $\alpha = 0.05$. All variables were z-standardized so that regression-coefficients are standardized β-coefficients that can be interpreted as effect sizes. Descriptive data of all experiments can be found in Tables 1–3. There was no difference between the *informed* and *uninformed* signaling groups concerning *prior conceptual knowledge*, $\beta = -0.61$, $F(1, 43) = 3.218$, $p = 0.08$., or reasoning, $F(1, 43) = 2.335$, $p = 0.13$.

For hypothesis 1 we modeled a regression from *extraneous cognitive load* on *instruction* (informed, non-informed). The data of the uninformed group was taken from Study 2a.

To test hypothesis 2, we built a regression of conceptual knowledge on instruction (informed, uninformed).

Hypothesis 3 was tested analogously to H2, but with *reasoning* as the dependent variable.

4.3. Results

> *H1*: **Informed cognitive load hypothesis.** Study 2b showed no significant difference between *extraneous cognitive* load in the uninformed, $M = 2.03$, $SD = 1.23$, and informed, $M = 2.44$, $SD = 1.60$, signaling group, $\beta = 0.30$, $F(1, 43) = 0.61$, $p = 0.44$.
>
> *H2*: **Informed signaling knowledge hypothesis.** For the post-test results the effect of *instruction* (uninformed signaling, informed signaling) reached one-sided significance for *conceptual knowledge*, $\beta = 0.72$, $F(1, 43) = 3.650$, $p < 0.03$ (one sided). *Conceptual post-test knowledge* was significantly better in the *informed* signaling condition compared to the uninformed signaling condition.
>
> *H3*: **Informed signaling reasoning hypothesis.** *Instruction* (informed, uninformed) was not significantly related to *reasoning*, $\beta = -0.12$, $F(1, 43) = 0.092$, $p = 0.75$.

4.4. Discussion

The results of Study 2b show an influence of prior instructional information, that is, participants that were introduced to the method of signaling in advance, yielded better results in conceptual knowledge than those who were naive. Thus, Study 2b shows that informed signaling is a more promising method to improve learning, reflecting and reasoning with authentic classroom videos than the uninformed signaling. Recent findings suggest that signal-induced load reduction might have excelled the signal-induced increase in extraneous load in our experiments. The information on signaling prohibited or reduced the induction of additional extraneous load by the signals. In line with this, cognitive load has been shown to be reduced, when information on the method and the task can be processed in advance (Paas and Van Merriënboer, 1994; Kirschner et al., 2006; Sweller et al., 2007; Schwonke et al., 2013). However, we could not directly demonstrate that the improvement in our studies was due to a reduction in extraneous cognitive load. This result might be attributed to certain weaknesses of our extraneous-cognitive-load scale. Some items (e.g., "*The explanations and hints in the learning environment were very ineffective with respect to learning*") strongly suggest focusing the answers in this scale on processing aspects of the environment, like instructions, rather than on the video material itself.

To gain more information about potentially low-cost signals, we conducted an exploratory study, where we presented participants with different signals and asked questions about the usability of the signals.

5. Study 3 – Signal evaluation

5.1. Theory

Previous studies have shown different learning outcomes with different signals (De Koning et al., 2007; Alpizar et al., 2020). However, the signals were implemented into varying learning environments and used with very different types of content. In order to find out which signals are most appropriate to reduce extraneous cognitive load in our learning environment, we conducted an exploratory interview study. In this study, all signals were presented within the learning environment with the same material and tasks. Only video-signal *combinations* were varied as described below. This guaranteed a high comparability.

5.2. Method

5.2.1. Sample and design

Nine university students (three male, five female, one diverse, age: $M = 21.89$, $SD = 2.51$) took part in this exploratory study. All participants were presented with five different signals, each one in a separate video. In order to not confound signal type with video content we made up two groups of participants receiving either one of two signal-content combinations (Appendix G). Assignment to the groups was random.

5.2.2. Material

We used the same videos as in the experiments before. The videos were presented with five different signals. (1) Based on the finding, that a reduction of the image section cues relevant locations (Glaser et al., 2017) and is effective in stressing relevant information in dynamic learning material (Amadieu et al., 2011), we decided to test the zoom-in effect. Therefore, the whole scene was scaled up beginning 300 ms before start of the relevant sequence, reaching its maximum after 4 s and lasting for the whole duration of the relevant sequence with the relevant image section being the center of attention. (2) The spotlight effect has already been shown to be effective in signaling (De Koning et al., 2007), Therefore, it was worth being implemented in this study. This was done by lighting up a circle around the relevant image section starting 300 ms before start of the relevant sequence and lasting for its whole duration. (3) The countdown effect was inspired by studies in general psychology that found, that information is processed better, when the time of its appearance is known in advance (Rolke, 2008). The countdown effect consisted of a visual presentation of numbers (font size: 100 pt., duration: 1 s per number) in the right lower corner of the screen in reversed order from 3 to 1, whereby the end of presentation marked the point in time, when the relevant sequence started. (4) The beep was an acoustic signal, that is, a tone with a frequency of 1 kHz that was presented for a duration of 250 ms and ended 300 ms before the relevant scene started. (5) The increased volume is naturally used by people to stress the importance of certain verbally transferred contents (Xie et al., 2019). Therefore, we included an increase in sound volume by 20 dB for the whole duration of relevant scenes. As said, in order to unconfound the signaling type and the concrete video content, participants were assigned to one of two groups with different combinations of content and signal type (Appendix C).

5.2.3. Instruments

After each video, participants were interviewed. They were asked to indicate the used signal-type and to describe, how the respective signal would support them if they were actually learning with the material. Additionally, they had to answer questions on the usability of the signals, including disruption by the signal (Appendix I).

5.2.4. Procedure

After going through the same theoretical learning phase as participants in Studies 1–2b, each participant received an instruction for the signal judgment. Then, participants initially watched all videos to get an impression on all signals. After watching the videos, participants received task instructions for the second phase of video presentation. They were told to attend the signals, to find out, what was used as signal in each video, and to imagine, how the signals would support them if they were learning with the material. They were informed that they would be interviewed after each video (Appendix H). Then, all videos were presented again with the same signals. After each video, participants were interviewed and had to answer questions on the usability of the signals and on different effects they had (Appendix H).

5.2.5. Analysis

Participants' answers to each of the questions were counted. Table 4 shows the percentages of participants that reported the respective perceptions/evaluations.

5.3. Results

Zoom-in, countdown, and *beep* were correctly identified as signals by 100% of the participants. Only 44.4% of the participants correctly identified an *increased volume* as signal. The *countdow*n and *beep* were perceived as rather distracting and even elicited startle in two of the participants. However, most of the participants, who rated the *zoom-in* as distracting and irritating had watched the video on "*human rights*" and stressed that in this video the signaling was perceived as misplaced and that they were irritated by a cut in the scene directly after the signal. Less than half of the participants judged the *beep* or the *increased volume* as helpful to identify relevant information or to learn in general. The interviews indicated that the auditive signals were not associated with relevance and even misinterpreted as technical issues. When asked to identify the best signal, a majority of participants indicated to prefer the *spotlight*. Importantly, a majority of participants mentioned that signaling was not necessary in the present videos because the videos were short and easy to process (Tables 5–7).

5.4. Discussion

The exploratory investigation of usability of different signals revealed some interesting aspects. First, auditory signals seem to be associated with technical problems and were not helpful as relevance cues. Second, not only the nature of the used signals is important but also the adequate placement within the sequences. Third, the *zoom-in* and *spotlight*-effect are promising signals for future studies. Fourth, participants agree, that signaling might be less useful in short and simple material. Especially the last point is in accordance with the results of Studies 1 - 2b and with previous findings in cognitive load research (Sweller, 1988; Paas, 1992; Van Merrienboer and Sweller, 2005). An explanation could be that cognitive capacity is not completely occupied by the material and thus, recipients do not need help with selection of relevant sequences. This is supported by the rather low reported extraneous cognitive load in Studies 1–2b.

TABLE 4 Intercorrelations of relevant variables in Study 1.

	Prior conceptual knowledge	Prior reasoning	Post conceptual knowledge	Post reasoning	Extraneous load
Prior conceptual knowledge					
Prior reasoning	**0.36***				
Post conceptual knowledge	0.25	0.22			
Post reasoning	0.23	0.13	**0.48***		
Extraneous load	−0.06	−0.07	−0.04	−0.01	

*Significant correlation, $p < 0.05$.

TABLE 5 Intercorrelations of relevant variables in Study 2a.

	Prior conceptual knowledge	Prior reasoning	Post conceptual knowledge	Post reasoning	Extraneous load
Prior conceptual knowledge					
Prior reasoning	−0.13				
Post conceptual knowledge	−0.03	**0.35***			
Post reasoning	−0.27	0.23	**0.36***		
Extraneous load	−0.07	0.23	0.20	−0.25	

*Significant correlation, $p < 0.05$.

TABLE 6 Intercorrelations of relevant variables in Study 2b.

	Prior conceptual knowledge	Prior reasoning	Post conceptual knowledge	Post reasoning	Extraneous load
Prior conceptual knowledge					
Prior reasoning	−0.12				
Post conceptual knowledge	−0.11	0.21			
Post reasoning	−0.22	0.28	0.29		
Extraneous load	−0.06	−0.02	0.11	**−0.32***	

*Significant correlation, $p < 0.05$.

TABLE 7 Results of the exploratory interviews in Study 3.

Category of judgment	Zoom-in	Spotlight	Countdown	Beep sound	Increased volume
Correctly perceived as a cue	100.00%	88.90%	100.00%	100.00%	44.40%
Misperceived as a cue	2 times	/	/	/	2 times
Startled	11.10%	11.10%	22.20%	22.20%	11.10%
Irritated	44.40%	66.70%	66.00%	55.60%	22.20%
Distracted	55.60%	33.30%	88,90%	66.70%	33.30%
Facilitates identification of relevant information	77.80%	88.90%	77.80%	44.40%	22.20%
Facilitates learning	44.40%	88.90%	66.60%	44.40%	22.20%
Overall rated best cue	3 votes	4 votes	0 votes	1 vote	1 vote

Percentage of participants, who indicated the aspects in the left column.

6. General discussion

In two experimental studies and an exploratory study, we investigated if and under what circumstances the signaling principle is suited to support learning from authentic classroom videos in the scope of evidence-informed reasoning in student teachers. Against the hypotheses, uninformed signaling did not result in better overall, conceptual knowledge or reasoning compared to unsignaled authentic classroom videos. This finding is surprising with respect to former studies on signaling (Schneider et al., 2018; Alpizar et al., 2020). To understand this unexpected finding, we need to find out what distinguishes the present from former studies.

As already pointed out in the introduction, signaling has not typically been investigated in classroom examples for learning, but rather during expository instruction. In classroom videos, signals can help guiding attention to certain points in time, and thus, for instance, to an utterance of a teacher. However, in expository instruction videos, where signaling has been investigated in the past, signals do more than just focusing attention to a certain point in time. Typically, signals such as arrows, frames, or labels are used to explicitly stress certain contents that are shown in a graph or animation (Wang et al., 2020). In contrast to this, tone and frame signaling in our videos has been rather unfocused and left open, what exactly needs to be processed during the relevant sequences. Thus, signaling in classroom video examples most likely needs to be different from what has been done so far. Signals that are aimed at supporting information selection in classroom videos need to point at relevant situations but can only to a limited degree spatially locate the center of attention or even stress the most central information. Signals that are aimed at information organization refer to abstract concepts (e.g., inducing cognitive strategies), typically presented auditorily as words, in classroom videos rather than concrete, visible items (e.g., heart valve) like it is typical for instructional videos. Therefore, they might need to be either well-prepared (introduction of color codes for different abstract categories before video presentation) or contain verbal information (metacognition is written above all behaviors associated with metacognition). However, both procedures might induce additional extraneous cognitive load.

Additionally, in contrast to other studies, participants acquired theoretical knowledge on learning strategies before watching the example videos. In contrast to instructional videos, where the knowledge is typically exclusively presented in the video, the prior presentation of conceptual knowledge in our examples makes it somehow more difficult to find differences between the experimental conditions. However, the fact, that there were effects in Study 2b within this kind of material and procedure shows that there is a considerable amount of learning during the classroom videos although knowledge was presented beforehand. Furthermore, it also demonstrates that the signals help processing the relevant information although their reference is not as clear as in expository instruction videos.

But why did we find just this one effect of informed signaling? One typical approach to explain effects of multimedia-design principles is a reduction of extraneous cognitive load. Technically speaking, signaling is supposed to reduce extraneous cognitive load by reducing the processed information to the gist (Mayer and Fiorella, 2014; Schneider et al., 2018; Alpizar et al., 2020). However, adding information, which is not directly related to the content, even if it is a signal, adds extraneous cognitive load (Sweller, 1988; Paas, 1992). As already discussed in the context of Study 1, the amount of added extraneous load depends on the actual nature of the used signal (Sweller et al., 1998) and needs to be outperformed by the load reduction that comes along with the selection advantage. Future studies need to find signals that reach this goal in order to validate signaling as a suitable design principle for example-based learning in classroom videos. Two aspects can prevent the success of signals: either the reduction of load is too small or the signal-induced additional load is too high.

(a) **The reduction of load is too small**

Concerning (a), signals reduce a considerable amount of load if there is rather overloaded material with only a small amount of relevant information. In contrast, if a higher amount of information in the video is relevant and the scene is simple, signals are not capable of substantially reducing extraneous load, because they are not really needed. The videos in our studies were rather short (duration: $M = 50$ s) and the relevant scenes made up around 20% of the whole video. Reported extraneous cognitive load was low (<3 on a scale ranging from 0 to 10) in all studies. Additionally, information was always provided verbally by the teacher. There was marginal, if any, uncertainty of the spatial location of relevant information, as the information was auditory. It might help to look at the person who talks, but the auditory information does not have to be searched for in the video (as can be the case for visual information). Typically, a spatial uncertainty induces additional extraneous load (Kalyuga et al., 1999). As spatial uncertainty was marginal in the present studies, the processing advantage due to easier selection of relevant sequences might not have exceeded the load induced by signal processing.

That is, in the present studies the range of possible processing facilitation was rather small due to rather simple and short authentic classroom videos. This interpretation was confirmed by the participants in Study 3, who mentioned that the videos were simple and short enough to be processed without signaling. Also consistent with this interpretation is the finding that there were no differences in reported extraneous cognitive load between the uninformed signaling and no signaling group. One might argue that still there was no ceiling effect in the conceptual knowledge and reasoning. However, this generally imperfect result might not necessarily be attributed to processing difficulties but might rather be due to a general inexperience with this type of task or some weaknesses of the videos itself as indicated in the case of the video on *human rights* in Study 3. To further explore this, it could be helpful to have a closer look at the development of knowledge over time, for example by using two or more video examples that are well balanced across the sample and measuring learning outcome after each. Additionally, signaling effects might be boosted by giving additional organizational signals instead of just selection support (De Koning et al., 2009).

(b) **The signal-induced additional load is too high**

Because our approach of using signaling to foster selection of relevant aspects in authentic classroom videos is innovative, participants are not at all used to finding signals in this kind of material. Therefore, the signals might have caused an initial irritation and led to processing costs. This irritation was smaller in size when using a tone and a frame compared to the written text but might still have corroded the intended selection benefit and learning advantage. In Study 2b only those participants, who had been informed about the reason and use of signals, were able to derive advantage from them. This supports the idea that there is a tradeoff between load reduction due to a selection advantage and load induction due to the processing of the signal itself. Thus, within our rather short classroom videos, there were no positive effects of signaling on knowledge acquisition. However, some evidence, namely better learning outcomes with informed signaling compared to uninformed signaling, an overall very low reported extraneous cognitive load as well as exploratory results of Study 3 support the idea that signaling can be profitable in more complex material with less irritating signals and an instruction to the signaling method in advance.

6.1. Limitations and outlook

To verify this interpretation, future studies should implement zoom-in or spotlight signals in longer, more demanding videos as well as vary and assess cognitive load. This should be done with a greater sample, especially to find reliable results on informed signaling. Additionally, the items measuring extraneous cognitive load did not explicitly refer to the video material. Thus, participants could have focused on the video or other learning material when answering the questions. For example, answering the item: "*The explanations and advises in the learning environment were very ineffective with respect to learning*," with a high agreement could either refer to some explanations of teachers in videos or to explanations in the environment that framed the work with the videos. Therefore, the extraneous cognitive load scale should be modified in order to measure extraneous cognitive load induced by the video material. Furthermore, the aforementioned item refers to effectiveness, which is highly correlated to the subjective perception of successful processing. Likewise, there might be a general item-independent tendency of participants to ascribe a higher load to videos, where they could not successfully extract the relevant information, independently of the actual reason (e.g., lack of knowledge, low general ability). To tackle this confound of individually perceived extraneous cognitive load and learning results, one might use eye-tracking techniques like Gaze Transition Entropy (GTE; Krejtz et al., 2015; Eckstein et al., 2017), or implement an environment with two different learning contents in order to vary signaling within participants. Additionally, it is worth considering finding items that differentiate between video-induced load and signal-induced load. This is because a medium load can either be due to medium load of both sources, or high load due to signals with low load due to videos, or low load due to signals but still high load due to videos.

To compare signaling effects in the classroom examples with those that have been discovered in instructional material, it might be wise to not only cue which information should be selected, but also how this can be organized, that is, what information belongs to which This support of organization has been suggested by the CTML (Mayer, 2014) and has already been done in textual and dynamic material (Ozcelik et al., 2010; Richter and Scheiter, 2019). Thereby, it could be possible to push the learning advantages of signaled compared to not signaled material. In videos on learning strategies, this can be done by not only telling participants, which information to select, but also, to which concrete strategy the sequence refers. Adding organization cues, however, is only indicated in more complex material, where processing needs to be supported or in participants with rather low prior knowledge (Richter and Scheiter, 2019).

7. Conclusion

All in all, the present studies give rise to the hypothesis that the use of signaling in classroom videos is advantageous under certain but not all circumstances, namely, when signaling is properly introduced and the material is sufficiently complex. That is, it could be advantageous when signaling costs are low and potential signaling advantages are high. Thus, signaling remains a promising design principle, in particular for example-based learning occasions, because example videos, especially in the context of evidence-informed practice, require processing of complex interactions of a high number of acting people (e.g., students in the classroom), while also considering plenty of contextual information and theoretical and empirical knowledge. The basis for learning to reason about theory or evidence-informed concepts and rules with videos is to select the information illustrating the concept or rule (e.g., one or two teacher's statements). After finding the relevant information, learners can attempt connecting it with the concept or rule and explain how the information exemplifies the concept or rule (professional vision, Seidel and Stürmer, 2014). Thus, signaling as support to notice relevant information can be highly advantageous in educational actions targeting the incorporation of evidence and knowledge in teacher professional vision and behavior. However, it is important to keep in mind that signaling can also be disadvantageous. Therefore, besides trying to minimize load induction by the signals itself, an implementation of signaling must be carefully evaluated in light of the complexity of the material.

Our studies offer first impressions on the effectiveness of different versions of signaling in classroom video examples. We find signaling effects under very limited conditions. This can be a starting point to find out more about the interplay of different processes of extraneous load induction by multi-media design attempts on the one hand side and load relieve that is induced by processing facilitation due to the design principle on the other hand side. Thus, our work stresses the importance to always have in mind both, the signal-induced load and the potential load reduction. To make signaling a safe option for educators, it is not only important to stress that signals always need to be properly introduced but it is also vital to find signals with a generally low potential to induce additional load. Therefore, we offered a first exploratory study attempting to find appropriate signals in the context of classroom videos. Interview data suggests the zoom-in and spotlight effect as most promising signals. By and large, the present work explored the potential of several approaches to signaling in learning to reason about classroom videos. Using key phrases to signal key auditory information in a classroom video is not recommendable. Using a tone and a frame to highlight key sequences has potential when learners are informed about the signal and its function. From the learners' point of view, the zoom-in and spotlight effect are promising signals that should be investigated in future research.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AR, TS, IG-F, and JM contributed data and or material. ST has written the first draft, which was then carefully revised by the other authors. The manuscript was then finalized by ST. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.974696/full#supplementary-material

References

Acock, A. C. (2014). *A Gentle Introduction to Stata* (*4th Edn*). TX: Stata Press.

Alpizar, D., Adesope, O. O., and Wong, R. M. (2020). A meta-analysis of signaling principle in multimedia learning environments. *Educ. Technol. Res. Dev.* 68, 2095–2119. doi: 10.1007/s11423-020-09748-7

Amadieu, F., Mariné, C., and Laimay, C. (2011). The attention-guiding effect and cognitive load in the comprehension of animations. *Comput. Hum. Behav.* 27, 36–40. doi: 10.1016/j.chb.2010.05.009

Ayres, P., and Paas, F. (2007). Making instructional animations more effective: a cognitive approach. *Appl. Cogn. Psychol.* 21, 695–700. doi: 10.1002/acp.1343

Baddeley, A. (1992). Working memory. *Science* 255, 556–559. doi: 10.1126/science.1736359

Ball, D. L. (2000). Bridging practices: intertwining content and pedagogy in teaching and learning to teach. *J. Teach. Educ.* 51, 241–247. doi: 10.1177/0022487100051003013

Ball, D., and Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *J. Teach. Educ.* 60, 497–511. doi: 10.1177/0022487109348479

Betrancourt, M., and Tversky, B. (2000). Effect of computer animation on users' performance: a review. *Trav. Hum.* 63:311. doi: 10.1006/ijhc.2002.1017

Biesta, G. (2017). "The future of teacher education: evidence, competence or wisdom?" in *A Companion to Research in Teacher Education*. eds. A. M. Peters, B. Covie and I. Menter (Singapore: Springer), 435–453.

Biesta, G., Allan, J., and Edwards, R. (2011). The theory question in research capacity building in education: towards an agenda for research and practice. *Br. J. Educ. Stud.* 59, 225–239. doi: 10.1080/00071005.2011.599793

Biggs, J. B., and Collis, K. F. (1982). "Evaluating the quality of learning," in *The SOLO Taxonomy*. New York: Academic Press.

Blomberg, G., Renkl, A., Gamoran Sherin, M., Borko, H., and Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *J. Educ. Res. Online* 5, 90–104. doi: 10.25656/01:8021

Boucheix, J. M., and Guignard, H. (2005). What animated illustrations conditions can improve technical document comprehension in young students? Format, signaling and control of the presentation. *Eur. J. Psychol. Educ.* 20, 369–388. doi: 10.1007/BF03173563

Chandler, P., and Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cogn. Instr.* 8, 293–332. doi: 10.1207/s1532690xci0804_2

Chandler, P., and Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *Br. J. Educ. Psychol.* 62, 233–246. doi: 10.1111/j.2044-8279.1992.tb01017.x

Dagenais, C., Lysenko, L., Abrami, P. C., Bernhard, R. M., Ramde, J., and Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: a review of empirical research. *Evid. Policy* 8, 285–309. doi: 10.1332/174426412X654031

De Koning, B. B., and Jarodzka, H. (2017). "Attention guidance strategies for supporting learning from dynamic visualizations" in *Learning from Dynamic Visualization*. eds. R. Lowe and R. Ploetzner (Cham: Springer)

De Koning, B. B., Tabbers, H. K., Rikers, R. M., and Paas, F. (2007). Attention cueing as a means to enhance learning from an animation. *Appl. Cogn. Psychol.* 21, 731–746. doi: 10.1002/acp.1346

De Koning, B. B., Tabbers, H. K., Rikers, R. M., and Paas, F. (2009). Towards a framework for attention cueing in instructional animations: guidelines for research and design. *Educ. Psychol. Rev.* 21, 113–140. doi: 10.1007/s10648-009-9098-7

Diery, A., Knogler, M., and Seidel, T. (2021). Supporting evidence-informed practice through teacher education: a profile analysis of teacher educators' perceived challenges and possible solutions. *Int. J. Educ. Res. Open* 2:100056. doi: 10.1016/j.ijedro.2021.100056

Diery, A., Vogel, F., Knogler, M., and Seidel, T. (2020). Evidence-informed practice in higher education: teacher educators' attitudes, challenges, and uses. *Front. Educ.* 5:62. doi: 10.3389/feduc.2020.00062

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., and Bunge, S. A. (2017). Beyond eye gaze: what else can eye-tracking reveal about cognition and cognitive development? *Dev. Cogn. Neurosci.* 25, 69–91. doi: 10.1016/j.dcn.2016.11.001

Glaser, M., Lengyel, D., Toulouse, C., and Schwan, S. (2017). Designing computer-based learning contents: influence of digital zoom on attention. *Educ. Technol. Res. Dev.* 65, 1135–1151. doi: 10.1007/s11423-016-9495-9

Glogger, I., Holzäpfel, L., Kappich, J., Schwonke, R., Nückles, M., and Renkl, A. (2013). Development and evaluation of a computer-based learning environment for teachers: assessment of learning strategies in learning journals. *Educ. Res. Int.* 2013, 1–12. doi: 10.1155/2013/785065

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., and Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learn. Instr.* 39, 72–87. doi: 10.1016/j.learninstruc.2015.05.001

Grossman, P., and McDonald, M. (2008). *Back to the future: Directions for research in teaching and teacher education.*

Hammersley, M. (2005). Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice. *Evid. Policy* 1, 85–100. doi: 10.1332/1744264052703203

Hilbert, T. S., Renkl, A., Schworm, S., Kessler, S., and Reiss, K. (2008). Learning to teach with worked-out examples: a computer-based learning environment for teachers. *J. Comput. Assist. Learn.* 24, 316–332. doi: 10.1111/j.1365-2729.2007.00266.x

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: a cognitive load perspective. *Educ. Psychol. Rev.* 19, 387–399. doi: 10.1007/s10648-007-9051-6

Kalyuga, S. (2008). Relative effectiveness of animated and static diagrams: an effect of learner prior knowledge. *Comput. Hum. Behav.* 24, 852–861. doi: 10.1016/j.chb.2007.02.018

Kalyuga, S. (2014). "The expertise reversal principle in multimedia learning" in *The Cambridge Handbook of Multimedia Learning*. ed. R. E. Mayer (Cambridge: Cambridge University Press), 576–597.

Kalyuga, S., Chandeler, P., and Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Appl. Cogn. Psychol.* 13, 351–371. doi: 10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Kersting, N. B., Givvin, K. B., Sotelo, F. L., and Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: further explorations of a novel measure of teacher knowledge. *J. Teach. Educ.* 61, 1–181. doi: 10.1177/00224871093478,172

Knogler, M., Hetmanek, A., and Seidel, T. (2022). Determining an evidence base for particular fields of educational practice: a systematic review of meta-analyses on effective mathematics and science teaching. *Front. Psychol.* 13:873995. doi: 10.3389/fpsyg.2022.873995

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., et al. (2015). Gaze transition entropy. *ACM Trans. Appl. Percept.* 13, 1–20. doi: 10.1145/2834121

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1207/s15326985ep4102_1

Le Fevre, D. M. (2003). "9. Designing for teacher learning: video-based curriculum design" in *Using Video in Teacher Education*. ed. J. Brophy (Bingley: Emerald Group Publishing Limited), 235–258.

Leppink, J., Paas, F., van der Vleuten, C. P., van Gog, T., and van Merrienboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P., and Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learn. Instr.* 30, 32–42. doi: 10.1016/j.learninstruc.2013.12.001

Martin, M., Farrell, M., Seidel, T., Rieß, W., Könings, K. D., van Merriënboer, J. J. G., et al. (2022). Focused self-explanation prompts and segmenting foster pre-service teachers'professional vision – but only during training! *J. Educ. Technol. High. Educ.* 19, 1–29. doi: 10.1186/s41239-022-00331-z

Mautone, P. D., and Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *J. Educ. Psychol.* 93:377. doi: 10.1037/0022-0663.93.2.377,389

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learn. Instr.* 29, 171–173. doi: 10.1016/j.learninstruc.2013.04.003

Mayer, R. E., and Fiorella, L. (2014). "12 principles for reducing extraneous processing in multimedia learning: coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles" in *The Cambridge Handbook of Multimedia Learning. Vol. 279* (New York, NY: Cambridge University Press)

Mayer, R. E., and Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *J. Educ. Psychol.* 100:380. doi: 10.1037/0022-0663.100.2.380,386

Moreno, R. (2007). Optimising learning from animations by minimizing cognitive load: cognitive and affective consequences of signaling and segmentation methods. *Appl. Cogn. Psychol.* 21, 765–781. doi: 10.1002/acp.1348

Moreno, R., and Mayer, R. E. (2002). Verbal redundancy in multimedia learning: when reading helps listening. *J. Educ. Psychol.* 94:156. doi: 10.1037/0022-0663.94.1.156,163

Ozcelik, E., Arslan-Ari, I., and Cagiltay, K. (2010). Why does signaling enhance multimedia learning? Evidence from eye movements. *Comput. Hum. Behav.* 26, 110–117. doi: 10.1016/j.chb.2009.09.001

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistic: a cognitive-load approach. *J. Educ. Psychol.* 84:429. doi: 10.1037/0022-0663.84.4.429,434

Paas, F. G., and Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Pschol. Rev.* 6, 351–371. doi: 10.1007/BF02213420

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cogn. Sci.* 38, 1–37. doi: 10.1111/cogs.12086

Richter, J., and Scheiter, K. (2019). Studying the expertise reversal of the multimedia signaling effect at a process level: evidence from eye tracking. *Instr. Sci.* 47, 627–658. doi: 10.1007/s11251-019-09492-3

Richter, J., Scheiter, K., and Eitel, A. (2016). Signaling text-picture relations in multimedia learning: a comprehensive meta-analysis. *Educ. Res. Rev.* 17, 19–36. doi: 10.1016/j.edurev.2015.12.003

Richter, J., Scheiter, K., and Eitel, A. (2018). Signaling text–picture relations in multimedia learning: the influence of prior knowledge. *J. Educ. Psychol.* 110:544. doi: 10.1037/edu0000220,560

Rolke, B. (2008). Temporal preparation facilitates perceptual identification of letters. *Percept. psychophys* 70, 1305–1313. doi: 10.3758/PP.70.7.1305

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence-informed medicine: what it is and what it isn't. *BMJ* 312, 71–72. doi: 10.1136/bmj.312.7023.71

Santagata, R., Gallimore, R., and Stigler, J. W. (2005). "The use of video teaching for teacher education and professional development" in *Preparing Teachers to Teach With Technology. Current Perspectives on Applied Information Technologies*. eds. C. Vrasidas and G. V. Glass (Greenwich: Preparing teachers to teach with technology, Information Age Publishing), 151–167.

Santagata, R., and Guarino, J. (2011). Using video to teach future teachers to learn from teaching. *ZDM* 43, 133–145. doi: 10.1007/s11858-010-0292-3

Schneider, S., Beege, M., Nebel, S., and Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educ. Res. Rev.* 23, 1–24. doi: 10.1016/j.edurev.2017.11.001

Schwonke, R., Ertelt, A., Otieno, C., Renkl, A., Aleven, V., and Salden, R. J. (2013). Metacognitive support promotes an effective use of instructional resources in intelligent tutoring. *Learn. Instr.* 23, 136–150. doi: 10.1016/j.learninstruc.2012.08.003

Seidel, T., Blomberg, G., and Renkl, A. (2013). Instructional strategies for using video in teacher education. *Teach. Teach. Educ.* 34, 56–65. doi: 10.1016/j.tate.2013.03.004

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educ. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Sherin, M. G. (2007). "The development of teachers' professional vision in video clubs" in *Video Research in the Learning Sciences*. eds. R. Goldman, R. Pea, B. Barron and S. Derry (Mahwah, NJ: Erlbaum), 383–396.

Sherin, M. (2014). "The development of teachers' professional vision in video clubs" in *Video Research in the Learning Sciences* (New York: Routledge), 397–410.

Sherin, M. G., and Han, S. Y. (2004). Teacher learning in the context of a video club. *Teach. Teach. Educ.* 20, 163–183. doi: 10.1016/j.tate.2003.08.001

Sherin, M., and Van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *J. Teach. Educ.* 60, 20–37. doi: 10.1177/0022487108328155

Slavin, R. E. (2002). Evidence-based education policies: transforming educational practice and research. *Educ. Res.* 31, 15–21. doi: 10.3102/0013189X031007015

Slavin, R. E. (2008). Evidence-based reform in education: what will it take? *Eur. Educ. Res. J.* 7, 124–128. doi: 10.2304/eerj.2008.7.1.124

Spiro, R. J., Collins, B. P., and Ramchandran, A. (2007). "Reflections on a post-Gutenberg epistemology for video use in ill-structured domains: fostering complex learning and cognitive flexibility," in *Video Research in the Learning Science*. eds. R. Goldman, R. Pea, B. Barron, and S. J. Derry (New York: Routledge), 93–100.

Stürmer, K., Könings, K. D., and Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: effect of courses in teaching and learning. *Br. J. Educ. Psychol.* 83, 467–483. doi: 10.1111/j.2044-8279.2012.02075.x

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Sweller, J., Kirschner, P. A., and Clark, R. E. (2007). Why minimally guided teaching techniques do not work: a reply to commentaries. *Educ. Psychol.* 42, 115–121. doi: 10.1080/00461520701263426

Sweller, J., Van Merrienboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205

Tversky, B., Heiser, J., Mackenzie, R., Lozano, S., and Morrison, J. (2008). Enriching animations. *Learning with animation: Research implications for design*, 263–285.

van Es, E. A., and Sherin, M. G. (2002). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *J. Technol. Teach. Educ.* 10, 571–596.

van Es, E. A., and Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teach. Teach. Educ.* 24, 244–276. doi: 10.1016/j.tate.2006.11.005

Van Gog, T., Paas, F., and Van Merriënboer, J. J. (2008). Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. *Learn. Instr.* 18, 211–222. doi: 10.1016/j.learninstruc.2007.03.003

Van Merrienboer, J. J., and Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educ. Psychol. Rev.* 17, 147–177. doi: 10.1007/s10648-005-3951-0

Wang, X., Lin, L., Han, M., and Spector, J. M. (2020). Impacts of cues on learning: using eye tracking technologies to examine the functions and design of added cues in short instructional videos. *Comput. Hum. Behav.* 107:106279. doi: 10.1016/j.chb.2020.106279

Weinstein, C. E., and Mayer, R. E. (1986). "The teaching of learning strategies" in *Handbook of Research in Teaching*. ed. C. M. Wittrock (New York, NY, USA: Macmillan), 315–327.

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., and Schuller, B. (2019). "Speech emotion classification using attention-based LSTM," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 1675–1685.

Yeh, C., and Santagata, R. (2015). Preservice teachers' learning to generate evidence-based hypotheses about the impact of mathematics teaching on learning. *J. Teach. Educ.* 66, 21–34. doi: 10.1177/0022487114549470

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 05 May 2023
DOI 10.3389/feduc.2023.1001523

Check for updates

OPEN ACCESS

EDITED BY
Katharina Loibl,
University of Education Freiburg, Germany

REVIEWED BY
Nikol Rummel,
Ruhr University Bochum, Germany
Anne Deiglmayr,
Leipzig University, Germany

*CORRESPONDENCE
Theresa Krause-Wichmann
✉ theresa.krause-wichmann@uni-saarland.de
Martin Greisel
✉ martin.greisel@phil.uni-augsburg.de
Christina Wekerle
✉ christina.wekerle@phil.uni-augsburg.de

SPECIALTY SECTION
This article was submitted to Teacher Education, a section of the journal Frontiers in Education

RECEIVED 23 July 2022
ACCEPTED 10 March 2023
PUBLISHED 05 May 2023

CITATION
Krause-Wichmann T, Greisel M, Wekerle C, Kollar I and Stark R (2023) Promoting future teachers' evidence-informed reasoning scripts: Effects of different forms of instruction after problem-solving.
Front. Educ. 8:1001523.
doi: 10.3389/feduc.2023.1001523

Promoting future teachers' evidence-informed reasoning scripts: Effects of different forms of instruction after problem-solving

Theresa Krause-Wichmann[1]*, Martin Greisel[2]*, Christina Wekerle[2]*, Ingo Kollar[2] and Robin Stark[1]

[1]Department of Education, Saarland University, Saarbrücken, Germany, [2]Faculty of Philosophy and Social Sciences, Augsburg University, Augsburg, Germany

Pre-service teachers face difficulties when dealing with problem situations in the classroom if their evidence-informed reasoning script (EIRS) is not adequately developed. An EIRS might be promoted by demonstrating how to implement evidence-informed reasoning after a problem-solving activity on an authentic case. However, it is unclear what form of instruction is appropriate to promote pre-service teachers in the development of an EIRS. The present 2×3-factorial experimental intervention study investigated how different forms of instruction on functional procedures (example-free vs. example-based) and on dysfunctional procedures (without vs. example-free vs. example-based) affect the development of an EIRS. *N*=384 pre-service teachers worked on a written case vignette of a problem situation in a problem-solving phase, in which the crucial steps of the EIRS were prompted externally. In the subsequent instruction phase, the participants compared their own solution with an example-free or example-based instruction on functional procedures, which was either supplemented by an example-free or example-based instruction on typical dysfunctional procedures or not at all. The participants' learning success (declarative EIRS; near and far transfer problem-solving performance) and error awareness were assessed. The results revealed that the example-based instruction on functional procedures led to a higher learning success than the example-free instruction. Both forms of instruction on dysfunctional procedures improved learning success compared to learning without one. During learning, error awareness was higher for learners who worked with an example-free instruction on dysfunctional procedures. In order to promote the development of an EIRS in pre-service teachers, it is promising to provide instruction after problem-solving that presents a functional example of evidence-informed reasoning for the given problem and that also points out typical dysfunctional approaches to solving the problem. The results highlight the importance of selecting appropriate scaffolds in case-based learning approaches that aim to develop cognitive schemata. The mechanisms that explain when and why instructions on dysfunctional procedures work need to be further explored.

KEYWORDS

teacher education, evidence-informed reasoning, problem-solving prior to instruction, script theory, example-based learning, error-based learning

1. Introduction

1.1. Background and relevance

For teachers it is part of their daily teaching practice to be confronted with various problem situations in the classroom, for example, when students have difficulties in grasping the learning material or are unmotivated. Against the background of the constant demand for *evidence-informed teaching practice*, teachers are increasingly asked to base their decisions not (only) on subjective theories or experiential knowledge, but especially on educational theories and empirical findings (Joyce and Cartwright, 2020; Ferguson, 2021; Slavin et al., 2021). For example, when confronted with a student who is frustrated because of a poor grade, it might be appropriate to frame the feedback in terms of Weiner's (1986) attribution theory of motivation and emotion by attributing the performance to variable characteristics (e.g., the student's learning effort). Teachers must be able to understand theories and/or empirical findings, and to apply them in an appropriate and meaningful way. In particular, when reflecting on problem situations, the cognitive processes during problem-solving should be well selected and systematically carried out to avoid hasty and possibly dysfunctional decisions (Brown, 1987; Chen and Bradshaw, 2007; Jonassen, 2011; Csanadi et al., 2021). Therefore, this paper is concerned with how pre-service teachers can be supported in reflecting on problem situations through a systematic, coherent, and evidence-informed reasoning process.

1.2. Evidence-informed reasoning

Problem situations that teachers face in their daily practice can be distinguished into (a) problems that require immediate judgments and routines of action under time pressure, and (b) problems that allow for retrospective analysis and reflection; this latter type of problems provides an opportunity for evidence-informed reasoning (Renkl, 2022; Leuders et al., in press). In the educational context, evidence-informed reasoning can be defined as a process of thinking about a problem and forming an argument in a systematic and coherent way, underpinning the argument with educational theories and/or empirical findings (Csanadi et al., 2021; Wilkes and Stark, 2022). Evidence-informed reasoning tends not only to require knowledge of theories and empirical findings, but also knowledge of *what* actions or steps to take and *how* to take them to solve the problem at hand (van Gog et al., 2019). Research on teachers' reasoning skills (e.g., Seidel and Stürmer, 2014; Kiemer and Kollar, 2018; Kramer et al., 2021) suggests a sequence of four key activities that are useful in retrospectively reflecting on problem situations: Having identified a problem, the teacher must (1) reconstruct the problem by developing an understanding of which particular aspect of the given complex situation is actually the core of the problem (*problem description*). (2) The teacher must explain the problem by developing a causal model that represents relevant cause-and-effect relations (*problem explanation*). (3) The teacher must derive student-related consequences or target states by deducing from the previous step which alternative state should be aimed (*goal setting*). (4) The teacher must derive self-related consequences, i.e., concrete options for action that are suitable for achieving the goals set in the previous step (*setting options for action*).

From a script-theoretical perspective (Schank and Abelson, 1977; Schank, 1999; Fischer et al., 2013), knowledge about *what* actions to perform and *how* to perform them is mentally organized in so-called *scripts*, which are a particular form of cognitive schemata. Teachers' knowledge about how to solve problem situations in the classroom can be conceptualized as a dynamic knowledge structure that guides teachers in solving a problem – an *evidence-informed reasoning script* (*EIRS*).

Several studies indicated that pre-service teachers and in-service teachers rarely display systematic, coherent, and evidence-informed reasoning: they show deficits in applying the crucial cognitive processes and/or evidence to the problem situation (e.g., Hetmanek et al., 2015; Lysenko et al., 2015; Yeh and Santagata, 2015; Kiemer and Kollar, 2018, 2021; Csanadi et al., 2021). These deficits could be explained not only by a lack of educational knowledge or affective barriers, such as negative beliefs regarding the usefulness of educational evidence (e.g., Dagenais et al., 2012; Lysenko et al., 2014; Kiemer and Kollar, 2021; Thomm et al., 2021), but also by an insufficiently developed EIRS (e.g., Kiemer and Kollar, 2018; Csanadi et al., 2021). Therefore, it is the responsibility of teacher education to find ways to support future teachers in developing an EIRS.

1.3. Fostering the development of an EIRS through authentic cases

When it comes to solving new, unfamiliar problem situations, teachers must be able to identify a problem, apply acquired knowledge and solve the problem systematically. To enable future teachers to tackle different problems based on a stable, well-developed EIRS, an important goal of teacher education is to teach for knowledge transfer. Previous research has extensively addressed the cognitive and situational aspects of learning and transfer, considering *what* is to be learned (e.g., abstract concepts, procedures), *to which situation or task* it is to be transferred (e.g., near transfer within a domain, far transfer beyond a domain), and *which instructional approach* is effective (e.g., problem-solving, metacognitive prompts; van Gog et al., 2019; Jacobson et al., 2020).

In terms of fostering transferable, evidence-informed reasoning skills in pre-service teachers, reflection on authentic and problematic case scenarios from pedagogical practice has proven valuable (e.g., Piwowar et al., 2018; Thiel et al., 2020; Helleve et al., 2021). From a cognitive perspective, case-based reasoning is beneficial for learning as it encourages learners to solve new problem situations by remembering previous situations and adapting their solutions (Kolodner, 1993). Despite the widespread use of case-based approaches, simply exposing pre-service teachers to complex problems from educational practice without further instructional guidance may not be sufficient to foster the development of an EIRS: Complex problems, such as classroom situations, place high demands on cognitive and metacognitive abilities, so that learners may perceive the difficulty of the problem situation as quite high; the capacity of working memory is likely to be overloaded, which may lead to cognitive overload (Ge and Land, 2003, 2004; van Merriënboer et al., 2003; Ge et al., 2005; Kirschner et al., 2006; van Merriënboer and Sweller, 2010; Jonassen, 2011). Therefore, learning with authentic cases needs to be carefully instructed in teacher education to be beneficial for learning.

1.3.1. Supporting student learning

In terms of supporting student learning in problem-solving or reasoning tasks, instructional means that monitor cognitive processes and direct attention to critical aspects, such as external prompts, show promise (Ge and Land, 2003; Ge et al., 2005; Chen and Bradshaw, 2007; Wilkes et al., 2022). However, learners are not able to judge whether they have performed the requested activities in an appropriate way if they are only guided to solve the problem by external prompts (Spensberger et al., 2021). This means learners only know *which* actions they are supposed to perform in a templated manner, but they do not become aware of whether the way they performed the activities was *functional* (i.e., more likely to be correct) or *dysfunctional* (i.e., more likely to be incorrect). Therefore, it seems promising to help pre-service teachers not only to follow the sequence of the four reasoning steps mentioned above (i.e., *problem description, problem explanation, goal setting, setting options for action*) in problem-solving, but also to teach them *how* to perform these steps.

1.3.2. Instruction after problem-solving

One approach that could help to promote the acquisition and transfer of an EIRS is the *instruction after problem-solving approach* (*PS-I*; also referred to as *problem-solving prior to instruction*), which includes both an initial problem-solving phase and a subsequent instruction phase. In the initial problem-solving phase, learners attempt to solve a problem that requires the application of yet to-be-learned principles, concepts, or strategies, and often fail to solve the problem successfully; in the subsequent instruction phase, learners are explicitly taught the content to be learned (e.g., principles, concepts or strategies that should be applied to the given problem; e.g., Loibl et al., 2017; Sinha et al., 2020; Sinha and Kapur, 2021a).

Research on instruction after problem-solving in STEM fields. The benefits of PS-I have been empirically demonstrated and replicated in a variety of contexts, especially in comparison to approaches with direct instruction or instruction prior to problem-solving; PS-I has become particularly popular in STEM domains, with the goal of promoting conceptual learning and transfer (for an overview cf. Loibl et al., 2017; Sinha and Kapur, 2019, 2021b). For example, a typical problem-solving task used in PS-I for learning mathematical concepts is that students are given data of different athletes and asked to identify the most consistent athlete based on a mathematical calculation (e.g., Kapur, 2012, 2014). In the instruction phase, students are taught the canonical solution based on the mathematical concept.

Key mechanisms of PS-I. Three recent reviews have addressed the possible reasons of when and why PS-I is effective – or is not effective (Loibl et al., 2017; Sinha and Kapur, 2019, 2021b). These reviews indicate that the effectiveness of PS-I does not seem to be rooted in its individual components (i.e., problem-solving and instruction), but in the way they are combined and the sequencing of the phases. From these reviews, at least three key mechanisms can be derived as to why PS-I is conducive to learning, namely (1) that prior knowledge is activated and differentiated during problem-solving, (2) that learners' attention is directed to the principles, concepts or strategies to be learned in the instruction phase, and (3) that learners become aware of their dysfunctional procedures (i.e., errors) by questioning their own solutions (*error awareness*; Loibl et al., 2017; Sinha and Kapur, 2019, 2021b). It should be noted that the implementation of both the problem-solving phase and the instruction phase differs across PS-I studies (Loibl et al., 2017; Sinha and Kapur, 2019, 2021b; Nachtigall et al., 2020). The popular *productive failure* approach, for which Sinha and Kapur (2021b) have formulated several fidelity criteria (e.g., providing problems that afford multiple representations, instruction building on students' solution), can be considered as a subtype of PS-I, but not all PS-I designs are examples of productive failure (for all fidelity criteria cf. Sinha and Kapur, 2021b). Overall, while the PS-I approach is very specifically characterized by its two phases, there is no single design of the two phases *per se* within the PS-I research; this is especially true for PS-I designs in non-STEM domains.

Instruction after problem-solving in fields beyond STEM. Given the large number of studies on PS-I in STEM fields, it is striking that the evidence on the impact of PS-I in less structured domains such as teacher education appears to be insufficient and inconsistent: Only a few studies investigated the effects of PS-I in less structured domains and these studies did not consistently indicate positive effects on learning (for an overview cf. Nachtigall et al., 2020). As different learning goals require different means of instruction, the inconsistent evidence could be explained by divergent learning goals (e.g., conceptual vs. procedural knowledge) and/or divergent design features (Loibl et al., 2017; Sinha and Kapur, 2019, 2021b; Nachtigall et al., 2020). For example, Schwartz and Bransford (1998) showed beneficial effects of PS-I with college students in the context of psychology, compared with instruction after reading or summarizing a text or with problem-solving without instruction. In the problem-solving phase of their PS-I condition, students had to analyze contrasting cases of data from simplified classical psychology experiments before engaging with a text or lecture on the relevant psychological phenomena. In a study by Glogger-Frey et al. (2015), who used PS-I in the domain of educational psychology with pre-service teachers to promote their abilities to assess learning strategies in learning journals, the PS-I condition was outperformed by a condition in which students studied a worked-out solution for the same problem-solving task before instruction. In the PS-I condition, participants were first given samples of learning journals written by (high-school) students and asked to develop criteria in order to assess the application of learning strategies. In the subsequent instruction phase, they were taught the evaluation criteria to be learned. However, the studies by Schwartz and Bransford (1998) and Glogger-Frey et al. (2015) differ from rather traditional PS-I studies (such as in STEM domains; Loibl et al., 2017; Sinha and Kapur, 2019, 2021b) regarding the control condition, which is that PS-I was not compared to instruction prior to problem-solving or direct instruction. Overall, it is difficult to draw conclusions about whether, when and why PS-I is also effective in domains that are rather less structured than STEM domains (such as teacher education).

Implications for the design of the problem-solving phase in less structured domains: Implementing structuring scaffolds. One design feature that is particularly different across PS-I studies (both in rather well- and less-structured domains) is the form of scaffolding in the problem-solving phase. In more traditional PS-studies, students received little or no support in solving a particular problem (e.g., Kapur, 2012, 2014). An advantage of unguided problem-solving is seen primarily in the fact that learners are given the opportunity to explore the problem, considering their own intuitive ideas (e.g., Sinha et al., 2020). To specifically promote comprehensive exploration processes, Sinha et al. (2020) and Sinha and Kapur (2021a) offered so-called *failure-driven scaffolds* that explicitly encourage learners to explore the problem with suboptimal representations and solution paths. Other studies offered rather *success-driven scaffolds* that guide students, structure their problem-solving process, and thereby help them to perform better – at the expense of less opportunity to explore the problem. Examples of more success-driven scaffolds include contrasting cases (e.g., Schwartz and Bransford, 1998; Loibl and Rummel, 2014a; Glogger-Frey et al., 2015; Chase and Klahr, 2017), self-explanation prompts (e.g., Roll et al., 2012; Fyfe et al., 2014), interaction support (e.g., Roll et al., 2012; Westermann and Rummel, 2012) and accuracy feedback (e.g., Fyfe et al., 2014; for an overview see Loibl et al., 2017; Sinha and Kapur, 2019, 2021b; Nachtigall et al., 2020). The above-mentioned study by Sinha et al. (2020) indicated that failure-driven scaffolding is more effective in learning with the PS-I approach than success-driven scaffolding with high specificity (i.e., definite advice for the optimal solution), but similar to success-driven scaffolding with low specificity (i.e., external prompts or hints that structure the problem-solving process into subtasks, or tell students what to do, but not how to do it). Based on the extensive research on scaffolding, albeit mainly in STEM fields, it could be postulated that especially for complex problems in rather less structured domains, it might be promising to structure the problem-solving process using success-driven scaffolds in the form of low-specific problem-solving prompts (Jonassen, 1997, 2000, 2011; Chen and Bradshaw, 2007). In the learning environment of the present study, which aims to encourage pre-service teachers to apply a functional EIRS on problem situations in the classroom, learners were supported by the means of more success-driven prompts that structured their problem-solving process along the EIRS with low specificity, i.e., without suggesting precise procedures or instructing how to perform them.

Implications for the design of the instruction phase in less structured domains: The form of instruction deserves more attention. However, compared to the state of research on the implementation of guidance in the problem-solving phase of PS-I, only little research has addressed the design features of the instruction phase (Loibl et al., 2017; Sinha and Kapur, 2019, 2021b). In line with Sinha and Kapur (2021b), encouraging learners to compare their own solutions with a sample solution and its critical features in the instruction phase can be seen as a central part of the PS-I approach. Working through these critical features and becoming aware of specific knowledge gaps can encourage learners to rethink their mental models, trigger active processing of the content to be learned, and promote knowledge acquisition (Chi, 2000; Loibl and Rummel, 2014b; Sinha and Kapur, 2021b). In contrast to more well-structured STEM problems, there is rarely *one single* canonical, i.e., "correct" solution for problems teachers face in the classroom. There are usually several functional options for action and even more dysfunctional options. Against this background, we argue that the form of instruction deserves special attention when it comes to helping pre-service teachers to develop an EIRS. It seems important to consider not only the potential benefits of guidance in the problem-solving phase, but also to explore the question of what features of the instruction phase make learning with the PS-I approach beneficial, especially in rather less structured domains such as teacher education. We argue that pre-service teachers should not be taught (only) a single canonical solution for a problem in the instruction phase, but especially how to apply a systematic, coherent, and evidence-informed problem-solving approach. This raises the question of how the instruction phase in PS-I can be designed to foster pre-service teachers' EIRS.

1.4. Designing instruction after problem-solving

1.4.1. Form of instruction on functional procedures

When there are several potentially "correct" options to solve a problem, as it is the case with most classroom problems, it is crucial to understand the rationale behind the problem and to apply powerful strategies or heuristics to tackle it (van Gog et al., 2004). If learners are presented with (only) an exemplary, worked-out solution to functionally deal with a given complex problem, there is a risk that learning will be hindered; learners might focus their attention on non-essential parts of the exemplary, worked-out solution (e.g., the specific wording) rather than on the underlying concepts, principles, or strategies (Renkl, 2002, 2017). It would be particularly precarious if learners misunderstand the meaning behind the exemplary, worked-out solution and do not become aware of their own dysfunctional procedures. In the worst case, not recognizing dysfunctional approaches could lead to learners internalizing and applying these approaches to problem situations in practice (Metcalfe, 2017).

Lange et al. (2021) therefore hypothesized that instructional explanations that do not include worked-out solutions (which they call *example-free instruction*) might be more effective for learning with complex problems than exemplary, functional worked-out solutions without instructional explanations. In their study, they examined the effects of such worked-out solutions without instructional explanations vs. example-free instruction on university students' critical thinking skills. While the worked-out solution only illustrated an exemplary, functional solution without instructional explanations, the example-free instruction provided extensive explanations for solving critical thinking problems. The example-free instruction has been proven to be superior to the worked-out solution without instructional explanations in promoting skill acquisition. Transferring Lange et al.' (2021) findings to teacher education, one could argue that it could be more appropriate to provide pre-service teachers with an example-free instruction after a problem-solving activity, explaining how to functionally manage the situation according to a normative EIRS in an abstract, general form.

On the other hand, research on example-based learning has revealed that examples are particularly promising to promote learning and transfer (van Gog and Rummel, 2010, 2018). Illustrating the content to be learned, such as abstract concepts, principles, or strategies by an example, can encourage learners to encode and interconnect both these abstract concepts, principles, or strategies and specific application possibilities; in this way, examples foster learners' ability to transfer acquired knowledge to new problem situations, and prevent them from acquiring so-called *inert knowledge* (i.e., knowledge that can be expressed but not applied to solve problems; Renkl et al., 1996; van Gog and Rummel, 2010, 2018; Renkl, 2017; Mayer, 2020). In example-based learning, examples are usually implemented in a way that concretely illustrates the application of the abstract concepts, principles, or strategies that are to be explained to the learners; to this end, a written step-by-step sample solution is provided of how a particular problem can be solved in a functional way (commonly known as *worked example*; e.g., Atkinson et al., 2000; van Gog and Rummel, 2010, 2018). In terms of promoting the acquisition and transfer of the EIRS, pre-service teachers might benefit from an *example-based instruction after problem-solving* that combines both an abstract, general description of the normative EIRS with a worked example that illustrates a functional problem-solving approach according to the EIRS (not to be mistaken with the approach of using worked examples as preparatory activity *prior to instruction*; cf. Glogger-Frey et al., 2015). Providing students with both a description of the EIRS and an example of how to apply it would allow students to better understand the rationale of the EIRS; moreover, they may become more easily aware of their own dysfunctional procedures (Loibl et al., 2017).

Thus, the question arises whether instruction after problem-solving should describe the functional operations of the above four reasoning steps in a still abstract, general form (i.e., *example-free instruction on functional procedures*), or in a worked-out form, in which it is – in addition to a general description of the normative EIRS – concretely illustrated how to solve the given problem by applying the operations of the EIRS (i.e., *example-based instruction on functional procedures*).

1.4.2. Form of instruction on dysfunctional procedures

Providing instruction on functional procedures helps to build knowledge about *what to do best* when faced with a particular problem or task (Oser et al., 2012; Renkl, 2017). Learners could benefit not only from instruction that focuses on *best practice*, but also from instruction that focuses on *dysfunctional practice* (e.g., Loibl and Rummel, 2014b; Loibl and Leuders, 2018, 2019). In terms of reflecting on problem situations in the classroom, a typical example of dysfunctional practice is that the third step of the EIRS *goal setting* is skipped. In other words, when a teacher has analyzed the problem (i.e., EIRS step 1: *problem description*) and its possible reasons or consequences (i.e., EIRS step 2: *problem explanation*) in a functional way, the teacher already formulates concrete options for action (i.e., EIRS step 4: *setting options for action*). It would have been important to first consider target states for the student to be achieved from the perspective of theory or empirical findings (i.e., EIRS step 3: *goal setting*) in order to avoid jumping to conclusions. Another typical dysfunctional practice is that the EIRS step *goal setting* is implemented in an inappropriate way: For example, even if a student-related target state is formulated before concrete options for action are determined (i.e., EIRS step 4), the target state is not coherent with the previous explanation of the problem and, therefore, may not contribute to solving the problem.

Tracing how a dysfunctional procedure differs from a functional one and understanding why a procedure is dysfunctional could help learners to correctly update schemata of functional procedures and to create schemata of dysfunctional procedures (i.e., *negative knowledge*; Oser et al., 2012). When learners thoroughly elaborate the features of dysfunctional procedures, they are more likely to address their own knowledge gaps (e.g., Große and Renkl, 2007; Durkin and Rittle-Johnson, 2012; Barbieri and Booth, 2020).

Error-based learning is criticized because it risks learners internalizing dysfunctional procedures (Metcalfe, 2017). It is important that learners are enabled to reflect on dysfunctional procedures in comparison to corresponding functional procedures (Durkin and Rittle-Johnson, 2012; Oser et al., 2012). In several PS-I studies demonstrating the benefits of presenting dysfunctional procedures learners were guided to compare both functional and dysfunctional solution attempts in whole-class discussions, i.e., only the teacher/instructor built upon learners' typical dysfunctional procedures and contrasted their features with those of functional procedures (e.g., Kapur, 2012, 2014; Kapur and Bielaczyc, 2012; Loibl and Rummel, 2014b). Loibl and Leuders (2018, 2019), therefore, investigated learners' individual cognitive processes during instruction. Both studies underline the advantages of providing both a solution of a functional procedure and typical dysfunctional solution attempts, supplemented by comparison prompts, as opposed to providing a solution of a functional procedure and typical dysfunctional solution attempts without comparisons prompts or only solutions of a functional procedure. Moreover, the study by Loibl and Leuders (2019) indicated that the elaboration of the dysfunctional procedures seemed to mediate this effect.

Looking at the PS-I approach from a conceptual change perspective (e.g., Vosniadou, 2013, 2019), one could argue that it might be promising for students' error awareness not to prompt the comparison of the functional procedure and typical dysfunctional procedures with each other, but of *their own* procedure with both a functional procedure and typical dysfunctional procedures. By comparing their own solution approach with both a solution of a functional procedure and of typical dysfunctional procedures, students might become more aware of the appropriateness of their own approach. When learners are explicitly encouraged to recognize dysfunctional procedures in their own solution, they can reorganize their mental schemata (Posner et al., 1982), and, as a result, be prevented from internalizing dysfunctional solution approaches (Metcalfe, 2017). A study by Heemsoth and Heinze (2016), for example, indicated the benefits of prompted reflection on the rationale behind one's own dysfunctional procedures (i.e., error-centered reflection) over the reflection on the correct solution corresponding to one's own dysfunctional procedures (i.e., solution-centered reflection) after problem-solving on knowledge acquisition. The authors explained the effect by suggesting that error-centered reflection would lead to more elaborated learning. Through error-centered reflection, students might have become more aware of the extent to which the procedures they used fostered or hindered the problem-solving process (Heemsoth and Heinze, 2016).

It remains unclear what form of instruction on dysfunctional procedures as a complement to instruction on functional procedures is appropriate to promote error awareness and the development of an EIRS in pre-service teachers. On the one hand, specific exemplification of *how not to do something* using *dysfunctional* (or *erroneous*) *examples* might encourage learners to identify, comprehend, explain, and/or remedy own dysfunctional procedures by referring to underlying concepts, principles, or strategies (e.g., Große and Renkl, 2007; Durkin and Rittle-Johnson, 2012; Barbieri and Booth, 2020). On the other hand, students might not benefit from exemplifications, if their own solution approach does not resemble the dysfunctional procedures presented; specific exemplifications of other students' dysfunctional procedures that have not been used by the learners themselves might even distract them from becoming aware of the correctness of their own approach (Loibl and Leuders, 2019).

Overall, it is unclear whether pre-service teachers learning with instruction on functional procedures after problem-solving would benefit from supplementary instruction on typical dysfunctional procedures of other pre-service teachers – be it in the form of an abstract, general description of typical dysfunctional procedures (i.e., *example-free instruction on dysfunctional procedures*) or in a worked-out form that – in addition to such an abstract, general description – also presents a specific exemplification of the described procedures for the given problem (i.e., *example-based instruction on dysfunctional procedures*).

1.5. Research questions and hypotheses

In the present experimental intervention study, pre-service teachers learned how to analyze educational problems in a systematic, coherent, and evidence-informed way, based on the PS-I approach. The study focused on the effects of different forms of instruction: On the one hand, we investigated which form of instruction on functional procedures (i.e., *example-free instruction on functional procedures* vs. *example-based instruction on functional procedures*) would be more suitable for pre-service teachers' ability to deal with problem situations according to the EIRS. Secondly, we aimed to find out whether and in what form instruction on functional procedures should be complemented by instruction on typical dysfunctional procedures (i.e., *without any instruction on dysfunctional procedures* vs. *example-free instruction on dysfunctional procedures* vs. *example-based instruction on dysfunctional procedures*). We formulated the following research questions and hypotheses:

Research Question 1. To what extent do different forms of instruction on functional procedures (i.e., *example-free* vs. *example-based*) and on dysfunctional procedures (i.e., *without* vs. *example-free* vs. *example-based*) affect pre-service teachers' learning success, i.e., the development of an EIRS (i.e., *the declarative EIRS*) and its application in similar and unfamiliar problem situations (i.e., *near and far transfer problem-solving performance*)?

> *Hypothesis 1*: Against the background of the benefits of examples (e.g., van Gog and Rummel, 2010, 2018), we expected that the example-based instruction on functional procedures would lead to a higher learning success than learning with the example-free instruction on functional procedures. We further expected that learning with the example-based instruction on dysfunctional procedures would be superior to learning with the example-free instruction and without any instruction on dysfunctional procedures. Yet, the example-free instruction on dysfunctional procedures should still work better than no instruction on dysfunctional procedures. We further hypothesized that a combination of the example-based instruction on functional procedures and dysfunctional procedures would lead to the highest learning success.

Research Question 2: To what extent do different forms of instruction on functional procedures (i.e., *example-free* vs. *example-based*) and on dysfunctional procedures (i.e., *without* vs. *example-free* vs. *example-based*) affect pre-service teachers' *error awareness*, with special emphasis on the written comparison of the students' solution and the instruction during learning?

> *Hypothesis 2*: We assumed that learning with the example-based instruction on functional procedures would have a greater potential to help learners to become aware of the correctness of their own approach than learning with the example-free instruction on functional procedures. We postulated that the example-based instruction on dysfunctional procedures would promote learners' error awareness more than the example-free instruction or no instruction on dysfunctional procedures. Learning with the example-free instruction on dysfunctional procedures should still be superior to learning without any instruction on dysfunctional procedures to promote students' error awareness. The combination of example-based instruction on functional and dysfunctional procedures was expected to be superior to all the other conditions in terms of promoting error awareness.

Research Question 3. To what extent is the postulated effect of the form of instruction on functional procedures on near and far transfer problem-solving performance serially mediated by pre-service teachers' error awareness and the declarative EIRS, moderated by the different forms of instruction on dysfunctional procedures (i.e., *without*, *example-free*, and *example-based*)?

> *Hypothesis 3*: Following Loibl and Leuders (2019), we expected that the hypothesized effect of the form of instruction on functional procedures on near and far transfer problem-solving performance would be serially mediated by the participants' error awareness and the declarative EIRS for the three different forms of instruction on dysfunctional procedures. The instruction on dysfunctional procedures is viewed as complementary to the instruction on functional procedures, and thus is conceptualized as a moderator. We assumed that (a) the form of instruction on functional procedures would influence the participants' error awareness. Subsequently (b), the higher the participants' error awareness, the more pronounced their declarative EIRS should be. Finally (c), the better the participants' declarative EIRS would be developed, the better the students should be able to solve near and far transfer problems. Further, we postulated that (d) these associations would be moderated by the form of instruction on dysfunctional procedures, respectively (for the postulated moderated serial mediation, see Figure 1).

2. Method

2.1. Participants and design

$N = 384$ pre-service teachers ($M_{Age} = 27.72$, $SD = 3.54$; 76% female) participated as part of their regular university courses. On average, students were in their third semester ($M_{Sem} = 3.43$, $SD = 1.64$). It was mandatory to take part in the training elements of the study, but it was voluntary to participate in the data collection. As no one opted out, the full sample was included in the analyses.

In a randomized 2×3-factorial between-subjects design, the factors *form of instruction on functional procedures* and *form of instruction on dysfunctional procedures* were varied, resulting in six

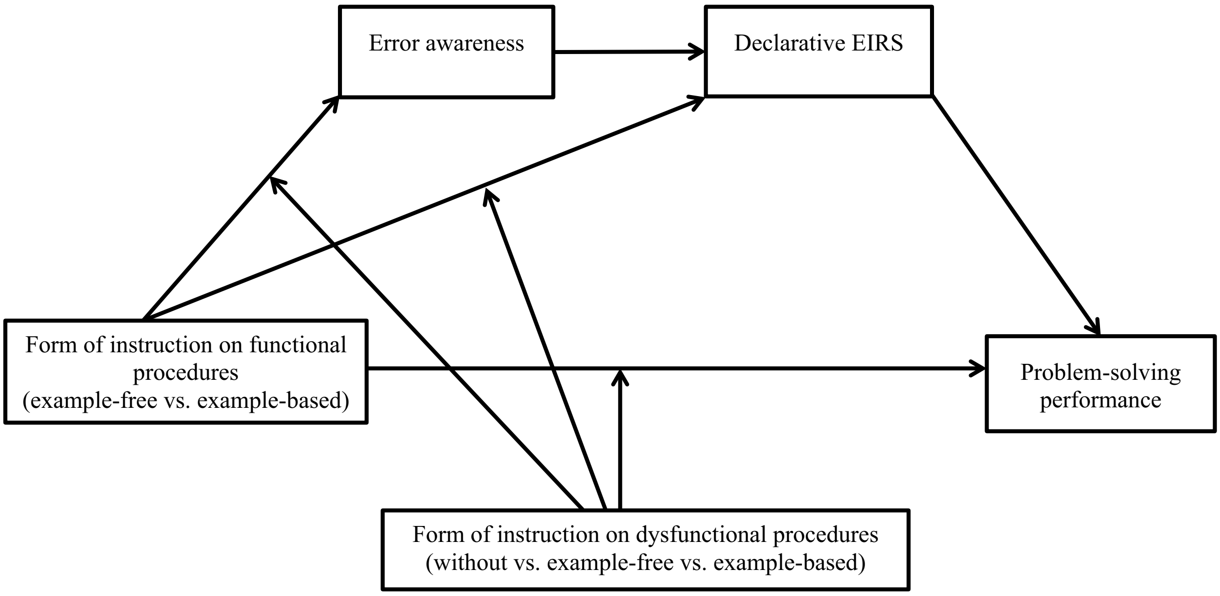


FIGURE 1
Conceptual regression model (Hypothesis 3).

TABLE 1 2×3-factorial, experimental design.

	Factor 2: Form of instruction on dysfunctional procedures		
Factor 1: Form of instruction on functional procedures	Without [0]	Example-free [D^-]	Example-based [D^+]
Example-free [F^-]	F^- 0 ($n=64$)	F^- D^- ($n=64$)	F^- D^+ ($n=64$)
Example-based [F^+]	F^+ 0 ($n=64$)	F^+ D^- ($n=64$)	F^+ D^+ ($n=64$)

experimental conditions with $n=64$ participants each (Table 1). With respect to the *form of instruction on functional procedures*, we varied whether students received either an *example-free* [F^-] or an *example-based* [F^+] instruction. Regarding the *form of instruction on dysfunctional procedures,* the participants were either not provided with any instruction on dysfunctional procedures [*0*], with an example-free [D^-], or with an example-based instruction [D^+]. A power analysis for 2×3-ANOVA with $\alpha=0.05$, $1-\beta=0.80$ indicated that the sample size of $N=384$ would be sufficient to identify small- to medium-sized effects.

2.2. Procedure and material

The study comprised two online sessions that were conducted independently and individually on a private computer with no time limit. An overview of the entire procedure is presented in Figure 2. In the first session, 2 weeks before the start of the training, participants answered a pre-questionnaire on socio-demographic data and a pre-test. The second session consisted of two parts: (1) the training with the actual intervention (i.e., problem-solving phase and instruction phase) and (2) a post-test.

The training followed the PS-I approach: in the initial problem-solving phase, participants worked through a written problematic classroom case scenario of 200 words using a written summary of corresponding educational theories and empirical findings. The case scenario involves a student who is angry at her mathematics teacher because, in her opinion, a test she failed had been too difficult. As a result, she declares that she does not want to study for the subject anymore. The summary of educational theories and findings focused on the attribution theory of motivation and emotion (Weiner, 1986), control-value theory of achievement emotions (Pekrun, 2006), achievement goal theory (Wigfield et al., 2016), and corresponding empirical findings. To reduce the complexity of the case analysis, all participants received success-driven prompts that structured the problem-solving process along the EIRS with low specificity (cf. *Instruction after Problem-Solving*): Per prompt, learners were instructed to perform one of the four EIRS steps (i.e., *problem description, problem explanation, goal setting, setting options for action*). In doing so, the participants were not instructed *how* to perform the steps, i.e., no precise procedures or operations were suggested (e.g., for *goal setting*: "Please set general goals to improve the situation"). In the subsequent instruction phase, participants were asked to compare their own solution with the provided instruction, to explain what they did the same or similar, and what they did differently or incorrectly. The instructions explained how to perform the operations of the EIRS in different ways that varied depending on condition (cf. *Participants and Design* and *Operationalization of the Independent Variables*). After the training phase, the post-test was administered.

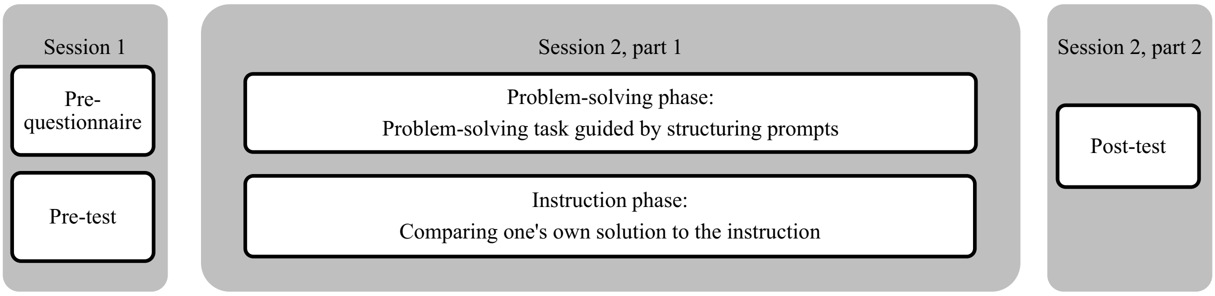


FIGURE 2
Procedure.

2.3. Operationalization of the independent variables

After the problem-solving phase, the participants with the *example-free instruction on functional procedures* received an instruction that described the crucial operations of a normative EIRS (i.e., how to ideally solve problem situations in a functional evidence-informed manner) in an abstract, general form, without any detailed elaboration of the EIRS for the given problem. For example, regarding the third step of the EIRS *goal setting*, the instruction focused on the two crucial operations of deriving student-related target states, and of connecting these relations with evidence (cf. Table 2).

The participants with the *example-based instruction on functional procedures* learned with an instruction that comprised the above-mentioned abstract description of the crucial operations of a functional EIRS, and a detailed "good practice" elaboration of these operations for the previous problem-solving task. For the step *goal setting*, for instance, it was concretely worked out which student-related goals could be derived coherently from the previously determined cause-and-effect relations, taking into account the evidence presented in the summary (cf. Table 2).

In the conditions that combined the instruction on functional procedures with an example-free instruction on dysfunctional procedures, the participants were also presented with several typical dysfunctional problem-solving procedures of pre-service teachers in an abstract form. Yet, this description did not include a precise elaboration of these typical dysfunctional procedures that might be applied to the problem in the previous training task. Regarding the step *goal setting*, for instance, students were told that pre-service teachers often select target states that do not fit to the cause-and-effect relations established before (cf. Table 2).

In the conditions that combined the instruction on functional procedures with an example-based instruction on dysfunctional procedures, participants also received the above-mentioned abstract, general description of typical dysfunctional procedures, along with a detailed elaboration of these dysfunctional procedures for the problem situation of the previous training task. In the step *goal setting*, for example, a typical student solution was illustrated that did not address the actual cause-effect-relations of the problem (cf. Table 2).

In the two conditions without any instruction on dysfunctional procedures, participants received only the assigned instruction on functional procedures (either in an example-free or an example-based version, depending on condition).

2.4. Measures

2.4.1. Dependent measures

In the post-test, three indicators of knowledge acquisition were measured: (a) declarative EIRS, (b) near-transfer problem-solving performance, and (c) far-transfer problem-solving performance.

Declarative EIRS. To measure the participants' declarative EIRS, they were asked to indicate how they would proceed when analyzing problematic classroom situations by describing in an open-ended format how they would implement the individual steps in detail. Two coders, who were blind to condition, were trained to code the participants' answers. The answers were coded with respect to whether they included the operations that define the EIRS (i.e., *problem description, problem explanation, goal setting, setting options for action*). To calculate inter-rater reliability, 10 % of the sample were double-coded by the two independent coders, with satisfactory inter-rater reliability (*Cohen's* $\kappa > 0.80$). Students were able to reach a maximum of five points for the definition of the four EIRS steps (two of the four EIRS steps, i.e., *problem explanation* and *goal setting*, also include an appropriate reference to evidence, for which 0,5 points each were awarded; cf. Table 3).

Near transfer problem-solving performance. Near transfer problem-solving performance was measured by an analysis of the participants' written analyses of a case vignette that described a problem similar to the training scenario in 200 words. Essentially, the problem was that a student is dissatisfied with the grade in the test and while rethinking the main reasons for the grade (i.e., lack of talent), he finally concludes that future learning would be a waste of time. The students were asked to analyze the situation along the EIRS with the aid of educational evidence, while the summary of corresponding educational theories and empirical findings was not provided anymore. Since there is more than one possible "correct" option to handle a complex problem situation, the answers were coded with respect to whether the participants had appropriately applied the operations that define the EIRS (i.e., *problem description, problem explanation, goal setting, setting options for action*). In other words, two coders assessed *how* the students reflected on the problem and whether it was done in a systematic, coherent, and evidence-informed manner, based on the operations of the EIRS taught in the training, or not. The right column of Table 3 presents a coding

TABLE 2 Exemplary illustration of the different forms of instruction for the step *goal setting* (translated and abbreviated).

Condition	Instruction on functional procedures		Instruction on dysfunctional procedures		
	Example-free	Example-based	Without	Example-free	Example-based
	Concluding from the previous analysis and based on the evidence (theories and findings) you received, you need to derive student-related consequences, i.e., target states. In other words, you need to consider what alternative state should be targeted for the situation by starting at the adjusting screws of the previously identified negative cause-effect relations, so that positive cause-effect relations are established. Thereby, theoretical terms/evidence must be assigned appropriately.	[Text of respective example-free instruction, cf. left column.] In line with Weiner's (1986) attribution theory and Pekrun's (2006) control-value theory, Mr. Schuster must ensure that Sarah develops an internally variable attribution pattern and a higher control appraisal. Sarah further needs to realize that she can improve through focused learning (learning goal orientation; Wigfield et al., 2016). In order to do so, however, she needs to feel a sense of control so that she does not become frustrated. Moreover, Sarah should view poor grades less as a failure and more as an opportunity to close the identified gaps.		Typical errors of other students: (1) Immediately after the analysis, concrete options for teacher action are already formulated without deriving student-related target states that should be achieved from the perspective of the theory or empirical findings. (2) The derived target state does not fit the previously explained cause-effect relation.	[Text of respective example-free instruction, cf. left column.] (1) The teacher might encourage Sarah that she can get a better grade next time and give her individual exercises to do at home. (2) In order for Sarah to stop being frustrated, her grades need to improve.
$F^- 0$	x		x		
$F^- D^-$	x			x	
$F^- D^+$	x				x
$F^+ 0$		x	x		
$F^+ D^-$		x		x	
$F^+ D^+$		x			x

example for the near transfer problem-solving performance. The coders, who were again blind to condition, reached satisfactory inter-rater reliability in 10 % of the sample (*Cohen's* $\kappa > 0.80$). Participants were able to score up to five points for correctly applying each operation of the EIRS to the problem situation.

Far transfer problem-solving performance. A second scenario measuring far transfer presented a completely novel situation, which had to be analyzed along the EIRS with the help of a summary of corresponding theories and empirical findings that was different from the training (self-determination theory; Ryan and Deci, 2000). In the 120-words-scenario, a science teacher fails to motivate his students to actively participate in his lesson. Two coders (blind to condition) estimated with satisfactory inter-rater reliability (in 10 % of the sample; *Cohen's* $\kappa > 0.80$) whether the four reasoning steps were correctly implemented or not. As with the near transfer problem-solving performance, participants could receive a maximum of five points (cf. Table 3).

To measure the participants' *error awareness* as a process variable in the instruction phase, we asked the participants to compare their own solution to the provided instruction, and to state what they did differently or what similarly when they were working on the problem. Two independent coders rated whether the participants had assessed their approach correctly, i.e., whether the students identified their functional (i.e., rather correct) and dysfunctional (i.e., rather incorrect) solution steps in terms of the EIRS (max. 5 points; cf. Table 3). Again, 10 % of the sample were double-coded. Inter-rater reliability was satisfying (*Cohen's* $\kappa > 0.80$).

2.4.2. Control measures

Prior problem-solving performance. To capture the participants' prior problem-solving performance, the pre-test presented a 200-words-scenario, which resembled the scenario measuring near transfer problem-solving performance. Participants were asked to analyze the scenario using educational theories and/or empirical findings, if possible, but they were not given a summary of corresponding evidence at this point. As with the post-test, learners could score up to max. 5 points (cf. Table 3). Ten percent of the sample were double-coded with satisfactory inter-rater reliability (*Cohen's* $\kappa > 0.80$). Reliability was sufficient (*Cronbach's* $\alpha = 0.70$).

Beliefs regarding the usefulness of educational evidence. In addition, participants had to rate five items measuring their beliefs regarding the usefulness of educational evidence (adapted from Wagner et al., 2016; e.g., "Educational knowledge is helpful to make good teaching decisions"; 1 = not at all true, 5 = very much true; Cronbach's $\alpha = 0.78$).

TABLE 3 Coding scheme and coding example (translated).

EIRS operations (declarative EIRS)	Max. Points	Coding example for the application of the EIRS operations (near transfer problem-solving performance)	Coding (Points)
(1) Problem description: The core of the problem must be described.	1	The core of the problem is the lack of motivation to deal with the learning content. The student perceives help as annoying and rejects it.	1
(2) Problem explanation: a) A correct model of cause-and effect relations of the problem situation must be formulated…	1	According to Weiner and Pekrun, the student attributes his failures internally and stably, i.e., he attributes his failures to a lack of talent. Consequently, he does not see any sense in trying harder and dealing intensively with the contents, because he sees himself as not gifted. In the course of this, resignation sets in and the help of others is rejected. Here possibly a work avoidance goal orientation (goal orientation theory) threatens, in order to escape the shame that the student feels. The student does not see the everyday relevance of the topic, which affects additionally negatively his motivation.	1
b) …and theoretical terms/evidence must be assigned appropriately.	0,5		0,5
(3) Goal setting: a) Student-related target states must be derived immediately logical from the result in cause-and-effect relations of the problem…	1	In line with the provided evidence on the benefits of reattribution, reattribution should take place on the part of the student in order to improve his self-concept. He should learn that through stronger effort, he will achieve better grades (attribution: internal and variable). The goal orientation should also change, towards a learning goal orientation.	1
b) …and theoretical terms/evidence must be assigned appropriately.	0,5		0,5
(4) Setting options for action: Concrete options for action must be directly derived from the target states.	1	The teacher could conduct partner dictations, which are corrected by the partner. In this way, the students give a positive value to the task and have a high level of control, since they can control each other's tasks. Consequently, they might feel pleasure in solving the task.	0,5

2.5. Analytic strategy

As level of significance, $\alpha = 0.05$ was applied for all global tests of significance. To answer Research Question 1, 2×3-factorial multivariate analyses of variance (MANOVAs) with the experimental factors *form of instruction on functional procedures* and *form of instruction on dysfunctional procedures* were calculated for the three interrelated learning success measures. To check effects on *error awareness* (Research Question 2), 2×3-factorial analyses of variance (ANOVAs) were computed. Planned contrast tests were performed (Research Question 1 and 2) to examine the presumed differences between the single conditions for significant main effects of the factor *form of instruction on dysfunctional procedures* and for significant interaction effects. If the ANOVA had revealed a significant main effect of the *form of instruction on dysfunctional procedures*, six contrasts tests were calculated that analyzed the postulated between-group differences in dependence of the three individual factor levels (i.e., without [*0*], example-free [D^-] example-based [D^+]).[1] In case of a significant interaction effect, the postulated superiority of the combination of the example-based instruction on functional procedures and the example-based instruction on dysfunctional procedures was checked by comparing this condition with all other conditions.[2] To account for potential alpha error inflation, the Bonferroni-Holm correction was applied.

Hayes' (2021) PROCESS macro for SPSS version 4.0 was used to examine the postulated mediation effect of the form of instruction on functional procedures on the near and far transfer problem-solving performance, moderated by the different forms of instruction on dysfunctional procedures (Research Question 3; Figure 1). Two moderated serial mediation models were computed with two mediators in a row (mediator 1: error awareness, mediator 2: declarative EIRS; Figure 1). The form of instruction on dysfunctional procedures was conceptualized as moderator. As dependent measure, the near transfer problem-solving performance was used in Model 1, and the far transfer problem-solving performance in Model 2. For mediation analyses, all mediating and dependent variables were z-standardized. To test the significance of indirect effects, we used 10,000 bootstrapped samples. We interpreted the indirect effect size as significant if its 95%-confidence interval did not include zero.

3. Results

3.1. Preliminary analyses

There were no *a-priori* differences between the experimental conditions regarding the control measures prior problem-solving

1 In case of a significant main effect of the *form of instruction on dysfunctional procedures*, the following conditions were compared with each other: (1) $F^+ D^+$ vs. $F^+ D^-$, (2) $F^+ D^+$ vs. $F^+ 0$, (3) $F^+ D^-$ vs. $F^+ 0$, (4) $F^- D^+$ vs. $F^- D^-$, (5), $F^- D^+$ vs. $F^- 0$, (6) $F^- D^-$ vs. $F^- 0$).

2 In case of a significant interaction effect, the condition $F^+ D^+$ was compared to (1) $F^+ D^-$, (2) $F^+ 0$, (3) $F^- D^-$, (4) $F^- D^+$, (5) $F^- 0$.

performance, $F(5, 378) = 0.99$, $p = 0.424$, and beliefs regarding the usefulness of educational evidence, $F(5, 378) = 0.737$, $p = 0.595$ (for means and standard deviations cf. Table 4).

3.2. Learning success (Research Question 1)

The means and standard deviations of the three learning success measures (i.e., declarative EIRS, near transfer problem-solving performance, far transfer problem-solving performance) for the six experimental conditions are displayed in Table 4. Overall, it is remarkable that all values were rather far away from the theoretical maximum to be reached in the post-test. The correlations between the three learning success measures (i.e., declarative EIRS, near transfer problem-solving performance, far transfer problem-solving performance) are presented in Table 5. The correlations were small to moderate and positive (Table 5). The results of the MANOVAs, the subsequent ANOVAs and the planned contrasts are reported below. The results of the planned contrast tests for a significant main effect of the form of instruction on dysfunctional procedures are presented in Table 6, those for a significant interaction effect in Table 7.

The 2×3-factorial MANOVA, using the *form of instruction on functional procedures* as well as the *form of instruction on dysfunctional procedures* as between-subject factors, and the three learning success measures as dependent measures revealed no significant interaction effect, $\Lambda = 0.98$, $F(6, 752) = 1.60$, $p = 0.144$. There was a significant main effect with moderate effect size for the form of instruction on functional procedures, $\Lambda = 0.93$, $F(3, 376) = 9.22$, $p < 0.001$, $\eta_p^2 = 0.07$. There was also a significant main effect with small effect size for the form of instruction on dysfunctional procedures, $\Lambda = 0.94$, $F(6, 752) = 3.69$, $p < 0.001$, $\eta_p^2 = 0.03$.

Concerning the *declarative EIRS*, in line with the MANOVA, the subsequent 2×3-factorial ANOVA showed no interaction effect, $F(2, 378) = 1.50$, $p = 0.225$. The results showed a small significant main effect of the form of instruction on functional procedures, $F(1, 378) = 5.25$, $p = 0.023$, $\eta_p^2 = 0.014$, indicating that participants working with the example-based instruction reproduced more operations of the EIRS than those participants working with the example-free instruction (Table 4). Further, there was a small significant main effect of the form of instruction on dysfunctional procedures, $F(2, 378) = 8.09$, $p < 0.001$, $\eta_p^2 = 0.041$. The planned contrasts showed that participants who were provided with any kind of instruction on dysfunctional procedures (i.e., example-free or example-based) outperformed those who were not, but these differences were only significant in the conditions with the example-based instruction on functional procedures (Tables 4, 6: Contrasts 2 and 3). For participants who learned with the example-free instruction on functional procedures the superiority of the two forms of instruction on dysfunctional procedures could only be observed at a descriptive level (Tables 4, 6: Contrasts 5 and 6). There were no significant differences between the conditions with the example-based instruction on dysfunctional procedures and the respective conditions with the example-free instruction on dysfunctional procedures with regard to influencing the acquisition of the EIRS (Table 6: Contrasts 1 and 4).

A similar picture emerged regarding the *near transfer problem-solving performance*. Consistent with the MANOVA, the interaction between both factors failed to reach significance, $F(2, 378) = 0.94$, $p = 0.392$. There was a small significant main effect of the form of instruction on functional procedures, $F(1, 378) = 7.04$, $p = 0.008$, $\eta_p^2 = 0.018$, indicating that the example-based instruction was superior to the example-free instruction in terms of fostering learning success (Table 4). Results further revealed a small significant main effect of the form of instruction on dysfunctional procedures, $F(2, 378) = 6.51$, $p = 0.002$, $\eta_p^2 = 0.033$. Contrast tests showed that students who worked with the example-free instruction on dysfunctional procedures outperformed those students who worked without any instruction on dysfunctional procedures, but only in the condition with the example-based instruction on functional procedures (Tables 4, 6: Contrast 3). Although the other calculated contrasts were not significant, it should be noted that for the participants in the condition with the example-free instruction on functional procedures, the latter finding could be observed at a descriptive level (Tables 4, 6: Contrast 6). The superiority of learning with the example-based instruction on dysfunctional procedures over learning without any instruction on dysfunctional procedures could be revealed only on a descriptive level, for both the conditions with the example-free and example-based instruction on functional procedures (Tables 4, 6: Contrasts 2 and 5).

Although the interaction effect of the MANOVA was not significant, the ANOVA for the *far transfer problem-solving performance* revealed a small significant interaction effect; it indicated that the example-based instruction on functional procedures was only superior, if it was combined with one of the instructions on dysfunctional procedures, $F(2, 378) = 3.27$, $p = 0.039$, $\eta_p^2 = 0.017$ (Figure 3; Table 4). The planned contrast tests revealed that participants who learned with the example-based instruction on functional procedures in combination with the example-based instruction on dysfunctional procedures showed a better performance than participants in all the other conditions (Tables 4, 7: Contrasts 2–5) – except for the condition that combined the example-based instruction on functional procedures with an example-free instruction on dysfunctional procedures; here, the result was on a descriptive level (Table 7: Contrast 1). Since the interaction was ordinal, main effects can be interpreted (Figure 3). The main effect of the form of instruction on functional procedures was significant with moderate effect size, $F(1, 378) = 24.54$, $p < 0.001$, $\eta_p^2 = 0.061$. Participants who received the example-based instruction showed a higher learning success than those who received the example-free instruction (Table 4). However, there was no significant main effect of the form of instruction on dysfunctional procedures, $F(2, 378) = 1.06$, $p = 0.347$.

3.3. Error awareness (Research Question 2)

Regarding *error awareness*, neither the interaction effect, $F(2, 378) = 1.68$, $p = 0.188$, nor the main effect of the instruction on functional procedures, $F(1, 378) = 2.07$, $p = 0.151$ reached the level of statistical significance in the 2×3 ANOVA. Yet, there was a small significant main effect of the form of instruction on dysfunctional procedures, $F(2, 378) = 7.36$, $p < 0.001$, $\eta_p^2 = 0.038$. The contrast tests revealed that the participants who learned with the example-free instruction on dysfunctional procedures as a supplement for the example-based instruction on functional procedures had a higher error awareness than those who received the example-based instruction on dysfunctional procedures as a supplement (Tables 4, 6: Contrast 1). All other contrasts (Table 6) did not reach statistical significance. However, it should be noted, that on a descriptive level,

TABLE 4 Means and standard deviations for all variables depending on the experimental condition.

Variable	Condition					
	Example-based instruction on functional procedures			Example-free instruction on functional procedures		
	Example-based instruction on dysfunctional procedures	Example-free instruction on dysfunctional procedures	Without instruction on dysfunctional procedures	Example-based instruction on dysfunctional procedures	Example-free instruction on dysfunctional procedures	Without instruction on dysfunctional procedures
	($F^+ D^+$)	($F^+ D^-$)	(F^+ 0)	($F^- D^+$)	($F^- D^-$)	(F^- 0)
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Control variables						
Prior problem-solving performance	1.30 (1.14)	1.63 (1.32)	1.60 (1.16)	1.53 (1.31)	1.72 (1.26)	1.41 (1.21)
Beliefs regarding the usefulness of educational evidence	3.74 (0.58)	3.76 (0.61)	3.85 (0.54)	3.85 (0.65)	3.74 (0.66)	3.89 (0.57)
Learning success						
Declarative EIRS	1.69 (1.14)	1.66 (1.19)	1.03 (0.93)	1.30 (1.12)	1.31 (0.91)	1.05 (0.91)
Near transfer problem-solving performance	2.33 (1.03)	2.55 (0.97)	2.01 (0.85)	2.10 (0.96)	2.16 (0.69)	1.91 (0.81)
Far transfer problem-solving performance	2.86 (1.02)	2.73 (1.10)	2.40 (1.00)	2.06 (0.96)	2.21 (0.92)	2.23 (0.92)
Learning process						
Error awareness	1.72 (0.60)	2.06 (0.50)	1.88 (0.48)	1.94 (0.55)	2.08 (0.54)	1.88 (0.51)

learning with the example-free instruction on dysfunctional procedures turned out to be more helpful in terms of estimating the own approach compared to learning without any instruction on dysfunctional procedures, with both forms of instruction on functional procedures (Tables 4, 6: Contrasts 3 and 6). The correlation analyses revealed significant positive correlations between error awareness and the three learning success variables with small to moderate effect sizes (Table 5).

3.4. Serial mediation (Research Question 3)

Two moderated serial mediation analyses were computed to examine to what extent the effect of the form of instruction on functional procedures (i.e., example-free vs. example-based) on near transfer problem-solving performance (Model 1) and far transfer problem-solving performance (Model 2) was serially mediated by *error awareness* as first stage mediator, and the *declarative EIRS* as second stage mediator for all three forms of instruction on dysfunctional procedures (Hypothesis 3; Figure 1). The indirect pathways are displayed in Table 8.

The participants' error awareness and declarative EIRS turned out to mediate the effect of the form of instruction on functional procedures on the students' near and far problem-solving performance (Model 1 and 2). In both models, this indirect effect was only significant for the conditions with the example-free instruction on dysfunctional procedures and without any instruction on dysfunctional procedures: when the instruction on functional procedures was supplemented by an example-free instruction on dysfunctional procedures or was not supplemented, the example-based instruction on functional procedures increased the students' error awareness. Consequently, the more the students were aware of the correctness of their own approach, the more developed was their declarative EIRS. The more developed the students' declarative EIRS was, the better they were able to solve near and far transfer problems.

It should be noted that the aforementioned effect of the form of instruction on functional procedures on near transfer problem-solving performance (Model 1) was completely mediated by error awareness and the declarative EIRS, as the direct effect did not reach the level of statistical significance (Table 8). The aforementioned effect of the form of instruction on functional procedures on far transfer problem-solving performance (Model 2) was also completely mediated in the condition without any instruction on dysfunctional procedures, and partially in the condition with the example-free instruction on dysfunctional procedures, as the direct effect was significant (Table 8).

TABLE 5 Bivariate Pearson correlations of control and dependent variables.

		1	2	3	4	5	6
1	Prior problem-solving performance	–					
2	Beliefs regarding the usefulness of educational evidence	0.04	–				
3	Declarative EIRS	0.07	0.13*	–			
4	Near transfer problem-solving performance	0.12*	0.14**	0.40**	–		
5	Far transfer problem-solving performance	−0.08	0.10*	0.20**	0.29**	–	
6	Error awareness	0.12*	0.15**	0.22**	0.34**	0.18**	–

**$p < 0.01$; *$p < 0.05$ (two-tailed).

TABLE 6 Results of planned contrast tests for dependent variables with significant main effect of the form of instruction on dysfunctional procedures.

Contrast	1		2		3		4		5		6	
Dependent variable	$F^+ D^+$ vs. $F^+ D^-$		$F^+ D^+$ vs. $F^+ 0$		$F^+ D^-$ vs. $F^+ 0$		$F^- D^+$ vs. $F^- D^-$		$F^- D^+$ vs. $F^- 0$		$F^- D^-$ vs. $F^- 0$	
	t	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Declarative EIRS	0.11[a]	0.455	3.58[a]	**<0.001**	3.37[a]	**<0.001**	−0.09[a]	0.465	1.39[a]	0.083	1.66[a]	0.050
Near transfer problem-solving performance	−1.25[a]	0.107	1.97[a]	0.026[b]	3.40[a]	**<0.001**	−0.40[a]	0.344	1.18 [a]	0.121	1.84[a]	0.034[b]
Error awareness	−3.65	**<0.001**	−1.70	0.045[b]	1.95	0.026[b]	−1.49	0.068	0.62	0.267	2.12	0.017[b]

p one-tailed. Adjusted level of significance by Bonferroni-Holm correction. Significant results are printed in bold letters.
[a]Corrected degrees of freedom due to violating homoscedasticity.
[b]Not significant due to Bonferroni-Holm correction.

Regarding the example-based instruction on dysfunctional procedures, none of the indirect paths reached statistical significance – neither in Model 1 nor in Model 2. Only the direct path from the form of instruction on functional procedures to far transfer problem-solving performance (Model 2) turned out to be significant (Table 8). The example-based instruction on functional procedures yielded better problem-solving performance without increasing error awareness and the declarative EIRS as long as an example-based instruction on dysfunctional procedures was also present.

4. Discussion

The aim of the present study was to examine the effects of a PS-I-based learning environment on pre-service teachers' ability to deal with problem situations in the classroom. The study focused on the effects of different forms of instruction after problem-solving in the domain of teacher education, with the aim of fostering a systematic, coherent, and evidence-informed reasoning approach. The study addressed evidence-informed reasoning not only from a perspective focusing on the use of scientific knowledge in practical situations, but especially from a perspective focusing on the cognitive processes of evidence-informed reasoning (Kiemer and Kollar, 2018; Csanadi et al., 2021). The internal validity of the study can be regarded as secured, as the six experimental conditions did not differ significantly with respect to possible confounding variables.

Regarding the students' learning success, the study focused on both basic declarative knowledge dimensions (i.e., declarative EIRS) and more complex procedural knowledge dimensions (i.e., near and far transfer problem-solving performance). Hypothesis 1 was partly confirmed. As expected, the acquisition of the declarative EIRS as well as its application in the near and far transfer was supported by the example-based instruction on functional procedures. The findings suggest that comparing one's own solution approach with a specific exemplification of how to deal with a problem following the EIRS in a functional way contributes to the memorization of the script and helps to deal with problem situations in a more systematic, coherent, and evidence-informed way. The specific exemplification of the EIRS seems to promote pre-service teachers' ability to apply the EIRS not only to familiar, but even to new, unfamiliar situations. The results contribute to the extensive literature pointing to the benefits of example-based learning, and especially of worked examples (e.g., Renkl, 2017; van Gog and Rummel, 2018; Mayer, 2020).

However, the postulated superiority of providing an example-based instruction on dysfunctional procedures over providing an example-free instruction in terms of fostering pre-service teachers' learning success could not be confirmed. Regarding the development of a declarative EIRS and its application to a familiar problem situation (i.e., near transfer problem-solving performance), it was important for the students to learn with any kind of instruction on dysfunctional procedures. The analysis of typical dysfunctional procedures obviously helps students to internalize the EIRS and to apply it to a situation that is comparable to the training situation – regardless the specific form of instruction. The specific form of instruction on dysfunctional procedures does not seem to be relevant for the internalization of the EIRS, but particularly the fact that dysfunctional strategies of solving a problem situation are presented at all. Apparently, underlining typical dysfunctional solution approaches of other pre-service teachers leads attention to the principles to be learned (Große and Renkl, 2007; Oser et al., 2012). Surprisingly, when it came to applying the EIRS to

TABLE 7 Results of contrast tests for dependent variables with significant interaction effects.

Contrast	1		2		3		4		5	
Dependent variable	$F^+ D^+$ vs. $F^+ D^-$		$F^+ D^+$ vs. $F^+ 0$		$F^+ D^+$ vs. $F^- D^+$		$F^+ D^+$ vs. $F^- D^-$		$F^+ D^+$ vs. $F^- 0$	
	t	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Far transfer problem-solving performance	0.74	0.230	2.67	**0.004**	4.59	**<0.001**	3.74	**<0.001**	3.65	**<0.001**

p one-tailed. Adjusted level of significance by Bonferroni-Holm correction. Significant results are printed in bold letters.

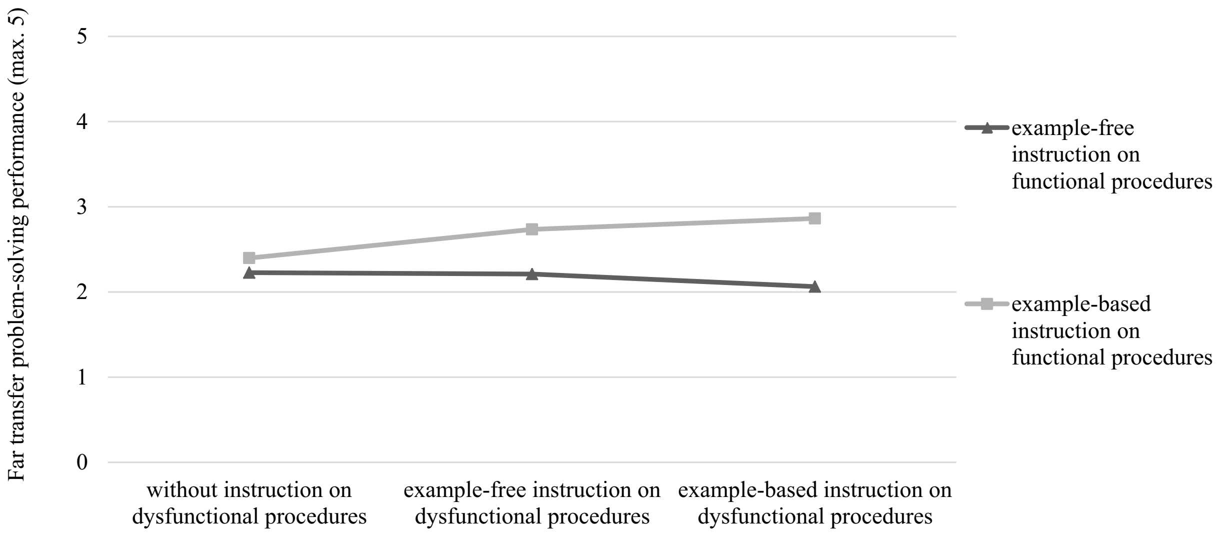


FIGURE 3
Interaction plot regarding the far transfer problem-solving performance.

a novel, unfamiliar problem situation (i.e., far transfer problem-solving performance), instructing students on typical dysfunctional procedures did not directly foster their ability to solve the problem, but the specific combination of any kind of instruction on dysfunctional procedures with the example-based instruction on functional procedures did. The exemplification of a functional approach seems to play out its potential when it is complemented by typical dysfunctional approaches.

Error-based learning approaches are often criticized for the risk of learners internalizing dysfunctional procedures (Metcalfe, 2017). This critique cannot be supported by the findings of the present study. The contrary is true: the presentation of dysfunctional approaches seems to be crucial to promote pre-service teachers' problem-solving performance. Looking at the present findings on the benefits of instructions on dysfunctional procedures as a whole, acquiring knowledge about what not to do in principle when dealing with problem situations (i.e., *negative knowledge*; Oser et al., 2012) seems to help learners to apply the EIRS to other situations. It even seems necessary to provide both a specific illustration of a "good practice" approach and at least a description or illustration of a dysfunctional approach in the instruction phase of PS-I to enable the transfer of the EIRS to unfamiliar situations. In sum, the benefits of error-based learning (e.g., Große and Renkl, 2007; Loibl and Leuders, 2018, 2019) can be replicated by the present study.

The expectations set under Hypothesis 2 regarding error awareness was also partially confirmed. Surprisingly, although the ANOVA results indicated that the form of instruction on functional procedures (i.e., example-free vs. example-based) had no impact on the quality of the student's assessments of their own approach, it must be noted that the mediation analyses hinted at this factor to be decisive for error awareness (Hypothesis 3). Considering the mediation results, it can be cautiously stated that pre-service teachers especially seem to benefit from instruction containing a specific illustration of a functional approach, in order to assess the correctness of their own approach.

Both the correlation analyses and the mediation analyses indicated that error awareness seems to be decisive for learning (Hypothesis 3), which is in line with the considerations on the key mechanisms of PS-I (Loibl et al., 2017; Nachtigall et al., 2020): Comparing one's own approach with an exemplification of a functional approach helps learners to become aware of their own errors, which then helps to further develop a mental schema of how to solve problems in principle. When dealing with problem situations, it seems important for pre-service teachers to have internalized the principle and operations behind the EIRS at a declarative level (Fischer et al., 2013), which in turn is promoted by comparing one's own approach with approaches of others. Therefore, the findings highlight the importance of estimating one's own approach in instruction after problem-solving for learning and correspond with findings on the benefits of comparisons for learning and transfer in general (e.g., Durkin and Rittle-Johnson, 2012; Loibl and Leuders, 2018, 2019).

However, the mechanism described above seems to apply particularly to instruction on functional procedures that is not supplemented by any

TABLE 8 Conditional direct and indirect effects for the moderated serial mediation models regarding near and far transfer problem-solving performance.

Tested paths	Model 1: Near transfer problem-solving performance						Model 2: Far transfer problem-solving performance					
	Without instruction on dysfunctional procedures		Example-free instruction on dysfunctional procedures		Example-based instruction on dysfunctional procedures		Without instruction on dysfunctional procedures		Example-free instruction on dysfunctional procedures		Example-based instruction on dysfunctional procedures	
	B	*CI* [LL; UL]	*B*	*CI* [LL; UL]	*B*	*CI* [LL; UL]	*B*	*CI* [LL; UL]	*B*	*CI* [LL; UL]	*B*	*CI* [LL; UL]
pv-ov	−0.038	[−0.264; 0.339]	0.257	[−0.046; 0.559]	0.118	[−0.244; 0.479]	0.112	[−0.216; 0.439]	**0.410***	**[0.057; 0.764]**	**0.720***	**[0.387; 1.053]**
pv-m1-ov	**0.068***	**[0.010; 0.144]**	**0.064***	**[0.009; 0.135]**	0.014	[−0.027; 0.064]	0.059	[−0.006; 0.151]	0.055	[−0.006; 0.146]	0.012	[−0.025; 0.062]
pv-m2-ov	−0.035	[−0.148; 0.071]	−0.087	[−0.033; 0.229]	0.122	[−0.003; 0.262]	−0.014	[−0.066; 0.033]	0.035	[−0.013; −0.104]	0.049	[−0.003; 0.125]
pv-m1-m2-ov	**0.030***	**[0.007; 0.063]**	**0.028***	**[0.005; 0.063]**	0.006	[−0.012; 0.027]	**0.012***	**[0.002; 0.028]**	**0.011***	**[0.001; 0.028]**	0.002	[−0.005; 0.013]

pv = predictor variable (factor: form of instruction on functional procedures); m1 = error awareness; m2 = declarative EIRS; ov = outcome variable (problem-solving performance). Mediating and dependent variables were z-standardized. Significance of regression coefficients (B) was interpreted referring to 95% confidence intervals (CI) not including zero. Significant results are printed in bold letters. *Significant result.

instructions on dysfunctional procedures or complemented only by example-free instructions on dysfunctional procedures. If there are no specific examples of typical dysfunctional procedures presented, it is important to have recognized one's own dysfunctional procedures, to build knowledge on how to solve and how not to solve problem situations in principle (i.e., *negative knowledge*; Oser et al., 2012). In contrast, when instruction on functional procedures is combined with a specific exemplification of dysfunctional procedures for a given problem, this knowledge seems to be acquired without the need to having recognized one's own errors before. This finding underlines the benefits of erroneous examples for learning success (e.g., Große and Renkl, 2007; Durkin and Rittle-Johnson, 2012; Barbieri and Booth, 2020). However, when it comes to estimating one's own approach, it might be more beneficial if the specific exemplifications of dysfunctional procedures match exactly with the dysfunctional procedures applied by the learners themselves (Loibl and Leuders, 2019); non-matching exemplifications might rather distract learners from recognizing own errors, which does not necessarily interfere with learning from the example itself.

4.1. Limitations and future directions

The present study is not without limitations. Concerning the methodology of the study, the sample size was too small to apply more refined statistical procedures such as structural equation modeling. The regression-based approach applied in the present study did not account for potential variable reciprocity or autoregressive effects. Since the study was considered as a short-term intervention, no claims can be made about long-term effects.

It is striking that even though all groups seem to have benefited from the PS-I approach, learning success was clearly below expectations; the mean values of the learning success variables proved to be far from the achievable maximum. The entire procedure, with the exception of the pre-test, was completed in one session to avoid sample failures, forgetting effects and between-group communication. However, working on three scenarios in one session was perhaps too strenuous for some of the participants, which might have influenced their learning success negatively. The internalization of an EIRS might require more time and practice (Anderson, 1996). However, the collected log data does not provide any information about time-on-task and whether work was done on the tasks during the time the computer window was open or not. It should be noted that the benefits of the PS-I approach have been mainly shown for conceptual learning and transfer, but less for learning complex problem-solving skills and procedural knowledge facets (Loibl et al., 2017; Sinha and Kapur, 2019, 2021b). Since, in contrast to many other PS-I studies, the approach was not compared with another approach, such as direct instruction or instruction prior to problem-solving (Loibl et al., 2017; Sinha and Kapur, 2019, 2021b; Nachtigall et al., 2020), no conclusion can be drawn as to whether another approach would have been more appropriate to promote the acquisition and transfer of an EIRS.

Furthermore, the PS-I approach was not implemented in a "traditional" way (e.g., traditional productive failure; e.g., Kapur, 2012, 2014). In contrast to the fidelity criteria formulated by Sinha and Kapur (2021b) for PS-I and productive failure in particular, the students in the present study worked individually and online and had no social surround that might have influenced their learning process – partly for COVID-19-pandemic reasons and partly for testing reasons. Thus, students were neither able to work collaboratively nor were they advised and supported by a teacher during the instruction phase. An opportunity to work in groups and a social surround facilitation in both PS-I phases might have fostered the identification, elaboration, and organization of the critical features of the EIRS (Sinha and Kapur, 2021b). One further difference from traditional PS-I studies is that in the problem-solving phase of the present study the students did not generate totally intuitive solution ideas to the given problem (e.g., Kapur and Bielaczyc, 2012). They were provided with educational evidence and external guidance to use the steps of evidence-informed reasoning, in order to reduce complexity. Therefore, the crucial mechanism of activating prior knowledge may have been mitigated in the training. However, the pre-test might have been more comparable to the traditional kind of problem-solving phase of the PS-I approach, which leads to another limitation of the study: the students might have also reflected on their procedure in the pre-test (on the influence of the presence and nature of pre-testing in PS-I see Sinha and Kapur, 2021b). But it should be noted that the pre-test was administered in a separate session 2 weeks before the training and that the problem scenario used in the pre-test was different from the problem scenario used in the intervention.

Another methodological limitation is that negative knowledge was not explicitly assessed. As one key goal of the present error-based approach is to make students aware of *how not to do something*, it would have been enlightening if statements could have been made about whether the students also remembered typical dysfunctional procedures or not (Loibl and Leuders, 2018).

Further, in previous PS-I studies, error awareness was usually measured at a subjective level using questionnaire items (e.g., Loibl and Rummel, 2014b; Glogger-Frey et al., 2015). In the present study, error awareness was operationalized by an objective measure to account for self-report biases (i.e., the task of comparing one's own solution to the given instruction). The instrument was conceptualized as a genuine part of the instruction phase of the implemented learning approach. Therefore, we are only able to draw conclusions about a rather specific form of error awareness that arises from the form of instruction depending on the experimental condition; no conclusions can be drawn about a rather *global awareness of knowledge gaps* that might have been triggered by cognitive processes in the problem-solving phase, without clarity about which specific component of the EIRS was lacking (Loibl and Rummel, 2014b). Since the written comparisons were at a rather superficial, principled level, it was not possible to analyze which specific errors the students became aware of. We cannot draw any inferences on the students' reflection quality, and whether they have really reflected on the rationale behind their dysfunctional procedures (Heemsoth and Heinze, 2016). Future studies should incorporate measures that allow precise conclusions on both global and specific awareness of knowledge gaps as well as the quality of reflections, in order to understand the key mechanisms of learning from instruction on functional and dysfunctional procedures after problem-solving.

The external validity might be another limitation of the present study. As the participants were already further advanced in their teacher studies, it is unclear whether novices would benefit in a similar way from both the treated topic and the learning approach or not (e.g., Kalyuga et al., 2001). In general, it needs to be clarified how much prior knowledge is needed to learn effectively from the PS-I approach incorporating typical dysfunctional procedures, and if rather adaptive forms of feedback are more efficient.

In addition to that, when there is a problem situation that a teacher needs to react immediately (such as a classroom disruption), the teacher has no time to work through the problem based on the EIRS. In time-pressured situations, teachers cannot search for appropriate evidence and reflect on the problem systematically (Renkl, 2022; Leuders et al., in press). Systematic, coherent, and evidence-informed reasoning based on the EIRS is more suitable for reflecting retrospectively on problem situations after teaching, and for making considerations for further lessons in a scientific way. However, as the EIRS could become automated with time and experience, it might serve as heuristic to make short-term decisions in the classroom (such as giving immediate feedback in terms of Weiner's (1986) attribution theory). If teachers acquire additionally a "toolbox" (Renkl, 2022) of theories and empirical findings with time, they can solve problems in an evidence-informed manner, even if the time is limited.

A further limitation results from the fact that the summary of educational theories and findings that was provided in the training (and in the post-test) was precisely aligned with the problems described in the written case vignettes. In pedagogical practice, teachers are confronted with a multitude of information that they can apply in problematic situations. Therefore, ecological validity may be limited to that end, too. Future studies should consider addressing ecologically more valid problem-solving tasks (e.g., by means of authentic classroom videos) and providing multiple resources of information, containing for example also irrelevant, less suitable, and/or contradicting information.

In this context, it must be noted that the setting of the present study – at least implicitly – suggests that reasoning processes in which teachers apply educational knowledge are of higher quality than reasoning processes that are based on experiential knowledge. This position can be criticized as being too distant from practice (Wilkes and Stark, 2022). Preliminary confirmation of this critique can be found in empirical findings by Gegenfurtner et al. (2020) indicating that in-service teachers and school principals not only use evidence-based bodies of knowledge in their reasoning, but also episodic that is experience-based knowledge. Overall, research should broaden the theoretical perspective on evidence-informed reasoning in a way that is in line with assumptions on the development of (adaptive) expertise (Tsui, 2009; Bohle Carbonell et al., 2014) and focuses on the reflexive integration of both evidence-based and experience-based information.

4.2. Conclusion

The findings of the present study have strong implications not only for future research, but especially for teacher education practice, as they provide valuable indications for the design of learning environments that target the evidence-informed reasoning skills of pre-service teachers. Against the background of the calls for evidence-informed teaching practice, the study reveals that case-based learning approaches can support pre-service teachers in developing competences that might help in dealing with future problem situations in a systematic, coherent, and evidence-informed way. Although the present implementation of the PS-I approach differs from "traditional" PS-I studies conducted in STEM domains, it can be cautiously stated that the present case- and error-based approach with the sequence *instruction after problem-solving* works in the rather unstructured domain of teacher education – especially when examples of functional and dysfunctional procedures are included. In this context, the script perspective as outlined by Fischer et al. (2013), as well as Kiemer and Kollar (2018), has proven to be beneficial. The internalization of the EIRS and its transfer seem to particularly require concretization by illustrating how to proceed and how not to proceed when being confronted with various problem situations in the classroom. However, further research on the mechanisms of why and when instruction on typical dysfunctional procedures works is needed. Motivational and attitudinal conditions should also be considered.

Overall, the findings underline the relevance of supporting future teachers in reflecting on problem situations by implementing learning opportunities in teacher education that not only focus on the acquisition of scientific knowledge (e.g., pedagogical-psychological theories and empirical findings), but also incorporate scaffolds for developing systematic evidence-informed problem-solving schemata. Future research should investigate how such problem-solving schemata can be further developed and consolidated, in order to be applied in everyday practice. Thus, research should also focus on the transition from learning with fictitious cases to learning with real-world situations in practice.

Data availability statement

The datasets presented in this article are not readily available because of legal reasons. Participants gave their consent to store, process and analyze the data as well as to the publication of the analysis results. Participants were assured that the data will not be distributed. Requests to access the datasets should be directed to TK-W, theresa.krause-wichmann@uni-saarland.de.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

TK-W and RS conceived of the presented idea and planned the experiment. TK-W and MG carried out the experiment. TK-W developed the theory and performed the computations. RS, MG, and IK verified the analytical methods. MG, CW, IK, and RS provided critical feedback and helped shape the research, analysis, and manuscript. TK-W took the lead in writing the manuscript with support and input from all co-authors. RS and IK supervised the project. All authors contributed to the article and approved the submitted version.

Funding

The research reported in this article was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, STA 596/8-1, KO 3426/3-1).

Acknowledgments

We thank Hannah Laass for her help in the preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Anderson, J. R. (1996). ACT: a simple theory of complex cognition. *Am. Psychol.* 51, 355–365. doi: 10.1037/0003-066X.51.4.355

Atkinson, R. K., Derry, S. J., Renkl, A., and Wortham, D. (2000). Learning from examples: instructional principles from the worked examples research. *Rev. Educ. Res.* 70, 181–214. doi: 10.3102/00346543070002181

Barbieri, C. A., and Booth, J. L. (2020). Mistakes on display: incorrect examples refine equation solving and algebraic feature knowledge. *Appl. Cogn. Psychol.* 34, 862–878. doi: 10.1002/acp.3663

Bohle Carbonell, K., Stalmeijer, R. E., Könings, K. D., Segers, M., and van Merriënboer, J. J. (2014). How experts deal with novel situations: a review of adaptive expertise. *Educ. Res. Rev.* 12, 14–29. doi: 10.1016/j.edurev.2014.03.001

Brown, A. (1987). "Metacognition, executive control, self-regulation, and other more mysterious mechanisms" in *Metacognition, motivation, and understanding*. eds. F. E. Weinert and R. H. Kluwe (Hillsdale, NJ: Lawrence Erlbaum), 65–116.

Chase, C. C., and Klahr, D. (2017). Invention versus direct instruction: for some content, it's a tie. *J. Sci. Educ. Technol.* 26, 582–596. doi: 10.1007/s10956-017-9700-6

Chen, C. H., and Bradshaw, A. C. (2007). The effect of web-based question prompts on scaffolding knowledge integration and ill-structured problem solving. *J. Res. Technol. Educ.* 39, 359–375. doi: 10.1080/15391523.2007.10782487

Chi, M. T. H. (2000). "Self-explaining expository texts: the dual processes of generating inferences and repairing mental models" in *Advances in instructional psychology: Educational design and cognitive science*. ed. R. Glaser, vol. *5* (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc), 161–238.

Csanadi, A., Kollar, I., and Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Dagenais, C., Lysenko, L., Abrami, P. C., Bernard, R. M., Ramde, J., and Janosz, M. (2012). Use of research-based information by school practitioners and determinants of use: a review of empirical research. *Evid. Policy* 8, 285–309. doi: 10.1332/174426412X654031

Durkin, K., and Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learn. Instr.* 22, 206–214. doi: 10.1016/j.learninstruc.2011.11.001

Ferguson, L. E. (2021). Evidence-informed teaching and practice-informed research. *Z. Pädagog. Psychol.* 35, 199–208. doi: 10.1024/1010-0652/a000310

Fischer, F., Kollar, I., Stegmann, K., and Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educ. Psychol.* 48, 56–66. doi: 10.1080/00461520.2012.748005

Fyfe, E. R., DeCaro, M. S., and Rittle-Johnson, B. (2014). An alternative time for telling: when conceptual instruction prior to problem solving improves mathematical knowledge. *Br. J. Educ. Psychol.* 84, 502–519. doi: 10.1111/bjep.12035

Ge, X., Chen, C.-H., and Davis, K. A. (2005). Scaffolding novice instructional designers' problem-solving processes using question prompts in a web-based learning environment. *J. Educ. Comput. Res.* 33, 219–248. doi: 10.2190/5F6J-HHVF-2U2B-8T3G

Ge, X., and Land, S. M. (2003). Scaffolding students' problem-solving processes in an ill-structured task using question prompts and peer interactions. *Educ. Technol. Res. Dev.* 51, 21–38. doi: 10.1007/BF02504515

Ge, X., and Land, S. M. (2004). A conceptual framework for scaffolding ill-structured problem-solving processes using question prompts and peer interactions. *Educ. Technol. Res. Dev.* 52, 5–22. doi: 10.1007/BF02504836

Gegenfurtner, A., Lewalter, D., Lehtinen, E., Schmidt, M., and Gruber, H. (2020). Teacher expertise and professional vision: examining knowledge-based reasoning of pre-service teachers, in-service teachers, and school principals. *Front. Educ.* 5:59. doi: 10.3389/feduc.2020.00059

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., and Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learn. Instr.* 39, 72–87. doi: 10.1016/j.learninstruc.2015.05.001

Große, C. S., and Renkl, A. (2007). Finding and fixing errors in worked examples: can this foster learning outcomes? *Learn. Instr.* 17, 612–634. doi: 10.1016/j.learninstruc.2007.09.008

Hayes, A. F. (2021). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*, *3rd Edn*. New York, NY: The Guilford Press.

Heemsoth, T., and Heinze, A. (2016). Secondary school students learning from reflections on the rationale behind self-made errors: a field experiment. *J. Exp. Educ.* 84, 98–118. doi: 10.1080/00220973.2014.963215

Helleve, I., Eide, L., and Ulvik, M. (2021). Case-based teacher education preparing for diagnostic judgement. *Eur. J. Teach. Educ.* 46, 50–66. doi: 10.1080/02619768.2021.1900112

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M., Gräsel, C., et al. (2015). Wozu nutzen Lehrkräfte welche Ressourcen? Eine Interviewstudie zur Schnittstelle zwischen bildungswissenschaftlicher Forschung und professionellem Handeln im Bildungsbereich [for what do teachers use which kind of resources? An interview study on the interface between research and professional practice in education]. *Unterrichtswissenschaft* 43, 193–208.

Jacobson, M., Goldwater, M., Markauskaite, L., Lai, P. K., Kapur, M., Roberts, G., et al. (2020). Schema abstraction with productive failure and analogical comparison: learning designs for far across domain transfer. *Learn. Instr.* 65:101222. doi: 10.1016/j.learninstruc.2019.101222

Jonassen, D. H. (1997, 1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educ. Technol. Res. Dev.* 45, 65–94. doi: 10.1007/BF02299613

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educ. Technol. Res. Dev.* 48, 63–85. doi: 10.1007/BF02300500

Jonassen, D. H. (2011). *Learning to solve problems: A handbook for designing problem-solving learning environments*. New York: Routledge.

Joyce, K. E., and Cartwright, N. (2020). Bridging the gap between research and practice: predicting what will work locally. *Am. Educ. Res. J.* 57, 1045–1082. doi: 10.3102/0002831219866687

Kalyuga, S., Chandler, P., Tuovinen, J., and Sweller, J. (2001). When problem solving is superior to studying worked examples. *J. Educ. Psychol.* 93, 579–588. doi: 10.1037/0022-0663.93.3.579

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instr. Sci.* 40, 651–672. doi: 10.1007/s11251-012-9209-6

Kapur, M. (2014). Productive failure in learning math. *Cogn. Sci.* 38, 1008–1022. doi: 10.1111/cogs.12107

Kapur, M., and Bielaczyc, K. (2012). Designing for productive failure. *J. Learn. Sci.* 21, 45–83. doi: 10.1080/10508406.2011.591717

Kiemer, K., and Kollar, I. (2018). "Evidence-based reasoning of pre-service teachers: a script perspective." in *Conference: International conference of the learning sciences: London, UK*.

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Z. Pädag. Psychol.* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1207/s15326985ep4102_1

Kolodner, J. L. (1993). "Introduction" in *Case-based learning*. ed. J. L. Kolodner (Boston: Springer), 1–5.

Kramer, M., Förtsch, C., Seidel, T., and Neuhaus, B. J. (2021). Comparing two constructs for describing and analyzing teachers' diagnostic processes. *Stud. Educ. Eval.* 68:100973. doi: 10.1016/j.stueduc.2020.100973

Lange, C., Almusharraf, N., Koreshnikova, Y., and Costley, J. (2021). The effects of example-free instruction and worked examples on problem-solving. *Heliyon* 7:e07785. doi: 10.1016/j.heliyon.2021.e07785

Leuders, T., Wiedmann, M., and Loibl, K. (in press). "Evidenzorientierung in der Lehrkräftebildung. [Evidence-Based Teacher Education]" in *Evidenzbasierung in Lehrer*innenbildung. Schule und Bildungspolitik und -administration – Neue Befunde zu alten Problemen. [evidence-based teacher education. School and educational policy and administration – New findings on old problems.]*. eds. K. Besa, D. Demski, J. Gesang and H. Hinzke (Wiesbaden: Springer VS).

Loibl, K., and Leuders, T. (2018). Errors during exploration and consolidation – the effectiveness of productive failure as sequentially guided discovery learning. *J. Math.-Didakt.* 39, 69–96. doi: 10.1007/s13138-018-0130-7

Loibl, K., and Leuders, T. (2019). How to make failure productive: fostering learning from errors through elaboration prompts. *Learn. Instr.* 62, 1–10. doi: 10.1016/j.learninstruc.2019.03.002

Loibl, K., Roll, I., and Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educ. Psychol. Rev.* 29, 693–715. doi: 10.1007/s10648-016-9379-x

Loibl, K., and Rummel, N. (2014a). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instr. Sci.* 42, 305–326. doi: 10.1007/s11251-013-9282-5

Loibl, K., and Rummel, N. (2014b). Knowing what you don't know makes failure productive. *Learn. Instr.* 34, 74–85. doi: 10.1016/j.learninstruc.2014.08.004

Lysenko, L. V., Abrami, P. C., Bernard, R. M., and Dagenais, C. (2015). Research use in education: an online survey of Canadian teachers. *Brock Educ. J.* 25, 1–20. doi: 10.26522/brocked.v25i1.431

Lysenko, L. V., Abrami, P. C., Dagenais, C., and Janosz, M. (2014). Educational research in educational practice: predictors of use. *Can. J. Educ.* 12, 1–17. doi: 10.2190/mhur-4fc9-b187-t8h4

Mayer, R. E. (2020). Advances in designing instruction based on examples. *Appl. Cogn. Psychol.* 34, 912–915. doi: 10.1002/acp.3701

Metcalfe, J. (2017). Learning from errors. *Annu. Rev. Psychol.* 68, 465–489. doi: 10.1146/annurevpsych-010416-044022

Nachtigall, V., Serova, K., and Rummel, N. (2020). When failure fails to be productive: probing the effectiveness of productive failure for learning beyond STEM domains. *Instr. Sci.* 48, 651–697. doi: 10.1007/s11251-020-09525-2

Oser, F., Näpflin, C., Hofer, C., and Aerni, P. (2012). "Towards a theory of negative knowledge (NK): almost-mistakes as drivers of episodic memory amplification" in *Professional and practice-based learning*. eds. J. Bauer and C. Harteis, *Human Fallibility: The Ambiguity of Errors for Work and Learning*, vol. *6* (Dordrecht: Springer), 53–70.

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* 18, 315–341. doi: 10.1007/s10648-006-9029-9

Piwowar, V., Barth, V. L., Ophardt, D., and Thiel, F. (2018). Evidence-based scripted videos on handling student misbehavior: the development and evaluation of video cases for teacher education. *Prof. Dev. Educ.* 44, 369–384. doi: 10.1080/19415257.2017.1316299

Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: towards a theory of conceptual change. *Sci. Educ.* 66, 211–227. doi: 10.1002/sce.3730660207

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learn. Instr.* 12, 529–556. doi: 10.1016/s0959-4752(01)00030-5

Renkl, A. (2017). "Instruction based on examples" in *Educational Psychology handbook series. Handbook of research on learning and instruction*. eds. R. E. Mayer and P. A. Alexander (New York, NY: Routledge), 325–348.

Renkl, A. (2022). Meta-analyses as a privileged information source for informing teachers' practice? A plea for theories as primus inter pares. *Zeitschrift für Pädagogische Psychol.* 36, 217–231. doi: 10.1024/1010-0652/a000345

Renkl, A., Mandl, H., and Gruber, H. (1996). Inert knowledge: analyses and remedies. *Educ. Psychol.* 31, 115–121. doi: 10.1207/s15326985ep3102_3

Roll, I., Holmes, N. G., Day, J., and Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. *Instr. Sci.* 40, 691–710. doi: 10.1007/s11251-012-9208-7

Ryan, R. M., and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68

Schank, R. C. (1999). *Dynamic memory revisited*. Cambridge, MA: Cambridge University Press.

Schank, R. C., and Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.

Schwartz, D. L., and Bransford, J. D. (1998). A time for telling. *Cogn. Instr.* 16, 475–5223. doi: 10.1207/s1532690xci1604_4

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educ. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Sinha, T., and Kapur, M. (2019). "When productive failure fails." in *41st annual conference of the cognitive science*. Montreal, Canada.

Sinha, T., and Kapur, M. (2021a). Robust effects of the efficacy of explicit failure-driven scaffolding in problem-solving prior to instruction: a replication and extension. *Learn. Instr.* 75:101488. doi: 10.1016/j.learninstruc.2021.101488

Sinha, T., and Kapur, M. (2021b). When problem solving followed by instruction works: evidence for productive failure. *Media Psychol.* 91, 761–798. doi: 10.3102/00346543211019105

Sinha, T., Kapur, M., West, R., Catasta, M., Hauswirth, M., and Trninic, D. (2020). Differential benefits of explicit failure-driven and success-driven scaffolding in problem-solving prior to instruction. *J. Ed. Psychol.* 113, 530–555. doi: 10.1037/edu0000483.?

Slavin, R. E., Cheung, A. C. K., and Zhuang, T. (2021). How could evidence-based reform advance education? *ECNU Rev. Educ.* 4, 7–24. doi: 10.1177/2096531120976060

Spensberger, F., Kollar, I., and Pankofer, S. (2021). Effects of worked examples and external scripts on fallacy recognition skills: a randomized controlled trial. *J. Soc. Work. Educ.* 58, 622–639. doi: 10.1080/10437797.2021.1963892

Thiel, F., Böhnke, A., Barth, V. L., and Ophardt, D. (2020). How to prepare preservice teachers to deal with disruptions in the classroom? Differential effects of learning with functional and dysfunctional video scenarios. *Prof. Dev. Educ.* 49, 108–122. doi: 10.1080/19415257.2020.1763433

Thomm, E., Sälzer, C., Prenzel, M., and Bauer, J. (2021). Predictors of teachers' appreciation of evidence-based practice and educational research findings. *Z. Pädagog. Psychol.* 35, 173–184. doi: 10.1024/1010-0652/a000301

Tsui, A. B. (2009). Distinctive qualities of expert teachers. *Teach. Teach.* 15, 421–439. doi: 10.1080/13540600903057179

van Gog, T., Paas, F., and van Merriënboer, J. J. G. (2004). Process-oriented worked examples. Improving transfer performance through enhanced understanding. *Instr. Sci.* 32, 83–98. doi: 10.1023/B:TRUC.0000021810.70784.b0

van Gog, T., and Rummel, N. (2010). Example-based learning: integrating cognitive and social-cognitive research perspectives. *Educ. Psychol. Rev.* 22, 155–174. doi: 10.1007/s10648-010-9134-7

van Gog, T., and Rummel, N. (2018). "Example-based learning" in *International handbook of the learning sciences*. eds. F. Fischer, C. E. Hmelo-Silver, S. R. Goldman and P. Reimann (New York, NY: Routledge), 201–209.

van Gog, T., Rummel, N., and Renkl, A. (2019). "Learning how to solve problems by studying examples" in *The Cambridge handbook of cognition and education*. eds. J. Dunlosky and K. A. Rawson (Cambridge, Ma: Cambridge University Press), 183–208.

van Merriënboer, J. J. G., Kirschner, P. A., and Kester, L. (2003). Taking the load off a learner's mind: instructional design for complex learning. *Educ. Psychol.* 38, 5–13. doi: 10.1207/S15326985EP3801_2

van Merriënboer, J. J. G., and Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Med. Educ.* 44, 85–93. doi: 10.1111/j.1365-2923.2009.03498.x

Vosniadou, S. (2013). "Conceptual change in learning and instruction: the framework theory approach" in *The international handbook of conceptual change*. ed. S. Vosniadou. *2nd* ed (New York, NY: Routledge), 11–30.

Vosniadou, S. (2019). The development of students' understanding of science. *Front. Educ.* 4:32. doi: 10.3389/feduc.2019.00032

Wagner, K., Klein, M., Klopp, E., and Stark, R. (2016). Förderung anwendbaren bildungswissenschaftlichen Wissens anhand kollaborativem Lernen aus Fehlern [facilitation of applicable knowledge in teacher education by collaborative learning from instructional errors]. *Unterrichtswissenschaft* 44, 373–390.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer.

Westermann, K., and Rummel, N. (2012). Delaying instruction: evidence from a study in a university relearning setting. *Instr. Sci.* 40, 673–689. doi: 10.1007/s11251-012-9207-8

Wigfield, A., Tonks, S., and Klauda, S. L. (2016). "Expectancy-value theory" in *Handbook of motivation in school*. eds. K. R. Wentzel and D. B. Mielecpesnm. *2nd Edn*. (New York, NY: Routledge), 55–74.

Wilkes, T., and Stark, R. (2022). Probleme evidenzorientierter Unterrichtspraxis. Anregungen und Lösungsvorschläge. [problems of an evidence-oriented educational practice. Suggestions and solution approaches.]. *Unterrichtswissenschaft*. doi: 10.1007/s42010-022-00150-1

Wilkes, T., Stark, L., Trempler, K., and Stark, R. (2022). Contrastive video examples in teacher education: a matter of sequence and prompts. *Front. Educ.* 7:869664. doi: 10.3389/feduc.2022.869664

Yeh, C., and Santagata, R. (2015). Preservice teachers' learning to generate evidence-based hypotheses about the impact of mathematics teaching on learning. *J. Teach. Educ.* 66, 21–34. doi: 10.1177/0022487114549470

Frontiers in Education

Explores education and its importance for individuals and society

A multidisciplinary journal that explores research-based approaches to education for human development. It focuses on the global challenges and opportunities education faces, ultimately aiming to improve educational outcomes.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact