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.

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.

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