EPISTEMOLOGICAL AND ETHICAL ASPECTS OF RESEARCH IN THE SOCIAL SCIENCES

EDITED BY: Ulrich Dettweiler, Barbara Hanfstingl and Hannes Schröter
PUBLISHED IN: Frontiers in Psychology and Frontiers in Education

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88963-688-4
DOI 10.3389/978-2-88963-688-4

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: researchtopics@frontiersin.org

EPISTEMOLOGICAL AND ETHICAL ASPECTS OF RESEARCH IN THE SOCIAL SCIENCES

Topic Editors:
Ulrich Dettweiler, University of Stavanger, Norway
Barbara Hanfstingl, Alpen-Adria-Universität Klagenfurt, Austria
Hannes Schröter, German Institute for Adult Education (LG), Germany

Citation: Dettweiler, U., Hanfstingl, B., Schröter, H., eds. (2020). Epistemological and Ethical Aspects of Research in the Social Sciences. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-688-4

Table of Contents

04 ***Editorial: Epistemological and Ethical Aspects of Research in the Social Sciences***
Ulrich Dettweiler, Barbara Hanfstingl and Hannes Schröter

07 ***The Heuristic Value of* p *in Inductive Statistical Inference***
Joachim I. Krueger and Patrick R. Heck

23 ***Commentary: The Need for Bayesian Hypothesis Testing in Psychological Science***
Jose D. Perezgonzalez

26 ***Commentary: Psychological Science's Aversion to the Null***
Jose D. Perezgonzalez, Dolores Frías-Navarro and Juan Pascual-Llobell

28 ***Should We Say Goodbye to Latent Constructs to Overcome Replication Crisis or Should We Take Into Account Epistemological Considerations?***
Barbara Hanfstingl

36 ***On the Development of a Computer-Based Tool for Formative Student Assessment: Epistemological, Methodological, and Practical Issues***
Martin J. Tomasik, Stéphanie Berger and Urs Moser

53 ***Why is Implementation Science Important for Intervention Design and Evaluation Within Educational Settings?***
Taryn Moir

62 ***Linearity vs. Circularity? On Some Common Misconceptions on the Differences in the Research Process in Qualitative and Quantitative Research***
Nina Baur

77 ***The Rationality of Science and the Inevitability of Defining Prior Beliefs in Empirical Research***
Ulrich Dettweiler

81 ***How to Crack Pre-registration: Toward Transparent and Open Science***
Yuki Yamada

84 ***Confounds in "Failed" Replications***
Paola Bressan

96 ***Quantitative Data From Rating Scales: An Epistemological and Methodological Enquiry***
Jana Uher

123 ***Book Review: Another Science is Possible***
Jose D. Perezgonzalez, Dolores Frías-Navarro and Juan Pascual-Llobell

frontiers in Psychology

EDITORIAL
published: 11 March 2020
doi: 10.3389/fpsyg.2020.00428

Editorial: Epistemological and Ethical Aspects of Research in the Social Sciences

Ulrich Dettweiler[1]*, *Barbara Hanfstingl*[2] and *Hannes Schröter*[3]

[1] Department of Cultural Studies and Languages, Faculty of Arts and Education, University of Stavanger, Stavanger, Norway, [2] Faculty of Interdisciplinary Studies, Institute of Instructional and School Development, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria, [3] Department "Teaching, Learning, Counselling", German Institute for Adult Education – Leibniz Centre for Lifelong Learning, Bonn, Germany

Keywords: epistemology, methodology, data science, implementation science, Bayesian approach, replicability

Editorial on the Research Topic

Epistemological and Ethical Aspects of Research in the Social Sciences

This Research Topic focuses on the questions "behind" empirical research in the social sciences, especially in psychology, sociology and education, and presents various ideas about the nature of empirical knowledge and the values knowledge is or should be based on.

The questions raised in the contributions are central for empirical research, especially with respect to disciplinary and epistemological diversity among researchers. This diversity is also mirrored by the variety of article types collected in this issue, "Hypotheses & Theory," "Methods," "Conceptual Analyses," "Review," "Opinion," "Commentary," and "Book Review."

Krueger and Heck explore in their "Hypotheses & Theory" article "The Heuristic Value of *p* in Inductive Statistical Inference." Taking up a very lively debate on the significance of null-hypothesis testing, they explore how well the *p*-value predicts what researchers presumably seek: the probability of the hypothesis being true given the evidence, and the probability of reproducing significant results. They furthermore investigate the effect of sample size on inferential accuracy, bias, and error. In a series of simulation experiments, they find that the *p*-value performs quite well as a heuristic cue in inductive inference, although there are identifiable limits to its usefulness. Krueger and Heck conclude that despite its general usefulness, the *p*-value cannot bear the full burden of inductive inference; it is but one of several heuristic cues available to the data analyst. Depending on the inferential challenge at hand, investigators may supplement their reports with effect size estimates, Bayes factors, or other suitable statistics, to communicate what they think the data say.

The argumentation of this article is flanked with a "Comment" on the article "The Need for Bayesian Hypothesis Testing in Psychological Science" (Wagenmakers et al., 2017) by Perezgonzalez. He argues that Wagenmakers et al. fail to demonstrate the illogical nature of *p*-values, while, secondarily, they succeed to defend the philosophical consistency of the Bayesian alternative. He comments on their interpretation of the logic underlying *p*-values without necessarily invalidating their Bayesian arguments. A second contribution by Perezgonzalez et al. deals with a comment on epistemological, ethical, and didactical ideas to the debate on null hypothesis significance testing, chief among them ideas about falsificationism, statistical power, dubious statistical practices, and publication bias presented by Heene and Ferguson (2017). The authors of this commentary conclude that frequentist approaches only deal with the probability of data under H0 [$p(D|H_0)$]. If anything about the (posterior) probability of the hypotheses is at question, then a Bayesian approach is needed in order to confirm which hypothesis is most likely given both the likelihood of the data and the prior probabilities of the hypotheses themselves.

OPEN ACCESS

Edited and reviewed by:
Douglas F. Kauffman, Medical University of the Americas – Nevis, United States

Correspondence:
Ulrich Dettweiler
ulrich.dettweiler@uis.no

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *06 February 2020*
Accepted: *24 February 2020*
Published: *11 March 2020*

Citation:
Dettweiler U, Hanfstingl B and Schröter H (2020) Editorial: Epistemological and Ethical Aspects of Research in the Social Sciences. Front. Psychol. 11:428. doi: 10.3389/fpsyg.2020.00428

Hanfstingl argues in her "Hypotheses & Theory" article "Should We Say Goodbye to Latent Constructs to Overcome Replication Crisis or Should We Take Into Account Epistemological Considerations?", that a lack of theoretical thinking and an inaccurate operationalization of latent constructs leads to problems that Martin Hagger calls "déjà variables," which ultimately also contribute to a lack of replication power in the social sciences. She proposes to use assimilation and accommodation processes instead of induction and deduction to explicate the development and validation of latent constructs and theories.

In the "Methods" article "On the development of a computer-based tool for formative student assessment: epistemological, methodological and practical questions," Tomasik et al. present a computer-based tool for formative student assessment. They deal with epistemological and methodological challenges as well as challenges in the practical implementation of these instruments. Overall, the authors show how formative assessment can not only increase efficiency, but also increase the validity of such feedback processes.

Closely related to this topic is the "Review" article by Moir. She defines components necessary to promote authentic adoption of evidence-based interventions and assessments in education, thereby increasing their effectiveness and investigates, how the quality of implementation has directly affected the sustainability of two such successful interventions. By analyzing implementation science, some of the challenges currently faced within this field are highlighted and areas for further research discussed. Furthermore, this article links to the implications for educational psychologists and concludes that implementation science is crucial already to the design and evaluation of interventions, and that the educational psychologist is in an ideal position to support sustainable positive change.

In "Linearity vs. Circularity? On Some Common Misconceptions on the Differences in the Research Process in Qualitative and Quantitative Research," Baur discusses the exaggeratedly simplified distinction between quantitative and qualitative paradigms in research methods and explains why we must assume a fluent transition between the two approaches. She points to similarities between the two supposedly antagonistic approaches in the use of induction, deduction and abduction, the roundness of the applied research phases and the analyses performed.

Closely related to that article, Dettweiler argues in his "Opinion" article that in both, so-called qualitative and quantitative research, it is inevitable for the research to define his or her prior beliefs, and that it is deeply irrational to believe that research methods are purely formal, distinct and free from value-judgements. There is also an informal part inherent to rationality in science which depends on the changing beliefs of scientists (Dettweiler).

Another "Opinion" article deals with some ethical challenges with pre-registration. Yamada argues that pre-registration, which should secure the transparency in the research process, including the experimental and analytical methods, the researchers' motivation and hypotheses, can easily be "cracked." She introduces the idea that to prevent such cracking, registered research reports should not be completely accepted as secure and valid just because "they were registered"; instead, several replications of the reported research with pre-registration should be performed. In addition, outsourcing experiments to multiple laboratories and agencies that do not share profitable interests with those of the registered researchers can be an effective means of preventing questionable research practices.

Where, Yamada refers to replication as a remedy to questionable research practice, Bressan presents a "Conceptual Analysis" and puts her finger into such questionable practice in the "Open Science Collaboration's Reproducibility Project," where a replication proved to be confounded. She shows in a case study on a "failed replication" that the dataset contained a bias which was absent in the original dataset; controlling for it replicated the original study's main finding. She concludes that, before being used to make a scientific point, all data should undergo a minimal quality control. Because unexpected confounds and biases can be laid bare only after the fact, we must get over our understandable reluctance to engage in anything *post-hoc*. The reproach attached to *p*-hacking cannot exempt us from the obligation to (openly) take a good look at our data.

In her contribution "Quantitative Data From Rating Scales: Quantitative Data From Rating Scales: An Epistemological and Methodological Enquiry," classified as a "Methods" type article, Uher presents yet another perspective on the "replication crisis" and fundamentally criticizes some traditions of psychological measurement and evaluation. Referring to the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS Paradigm), she investigates psychological and social science concepts of measurement and quantification. Uher proposes to apply metrological measurement concepts with a more precise focus on data generation.

Lastly, a "Book-review" by Perezgonzalez et al. on "Another science is possible: a manifesto for slow science" (Stengers and Muecke, 2018) is completing this collection.

We sincerely hope that this collection can in fact contribute to such "another science," a science that does not build on shallow dichotomies, such as "qualitative" or "quantitative," a science that is transparent, rigorous, epistemologically informed, and ethical.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The University of Stavanger supported UD's work with a sabbatical and a grant in the program for Yngre Fremragende Forskere financed by the Norwegian Research Council (Internal Project No. IN11714).

ACKNOWLEDGMENTS

We would like to thank all authors who have contributed with their ideas to the Research Topic in its present form. We would also like to thank those who, as editors and reviewers, have contributed to a significant increase in quality.

REFERENCES

Heene, M., and Ferguson, C. J. (2017). "Psychological science's aversion to the null, and why many of the things you think are true, aren't," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions,* eds S. O. Lilienfeld and I. D. Waldman (Chichester: John Wiley & Sons), 34–52.

Stengers, I., and Muecke, S. (2018). *Another Science Is Possible : A Manifesto for Slow Science (English edition. ed.).* Cambridge: Polity.

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). "The need for Bayesian hypothesis testing in psychological science," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions* (Hoboken, NJ: Wiley-Blackwell), 123–138.

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.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 09 June 2017
doi: 10.3389/fpsyg.2017.00908

The Heuristic Value of *p* in Inductive Statistical Inference

Joachim I. Krueger and Patrick R. Heck**

Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, United States

Many statistical methods yield the probability of the observed data – or data more extreme – under the assumption that a particular hypothesis is true. This probability is commonly known as 'the' *p*-value. (Null Hypothesis) Significance Testing ([NH]ST) is the most prominent of these methods. The *p*-value has been subjected to much speculation, analysis, and criticism. We explore how well the *p*-value predicts what researchers presumably seek: the probability of the hypothesis being true given the evidence, and the probability of reproducing significant results. We also explore the effect of sample size on inferential accuracy, bias, and error. In a series of simulation experiments, we find that the *p*-value performs quite well as a heuristic cue in inductive inference, although there are identifiable limits to its usefulness. We conclude that despite its general usefulness, the *p*-value cannot bear the full burden of inductive inference; it is but one of several heuristic cues available to the data analyst. Depending on the inferential challenge at hand, investigators may supplement their reports with effect size estimates, Bayes factors, or other suitable statistics, to communicate what they think the data say.

Keywords: statistical significance testing, null hypotheses, NHST, Bayes' theorem, replicability, reverse inference

OPEN ACCESS

Edited by:
Ulrich Dettweiler,
University of Stavanger, Norway

Reviewed by:
Jose D. Perezgonzalez,
Massey University, New Zealand
Torbjørn Waaland,
University of Stavanger, Norway

****Correspondence:***
Joachim I. Krueger
joachim@brown.edu
Patrick R. Heck
pheck1000@gmail.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *17 March 2017*
Accepted: *17 May 2017*
Published: *09 June 2017*

Citation:
Krueger JI and Heck PR (2017) The Heuristic Value of p in Inductive Statistical Inference. Front. Psychol. 8:908. doi: 10.3389/fpsyg.2017.00908

INTRODUCTION

The casual view of the p-value as posterior probability of the truth of the null hypothesis is false and not even close to valid under any reasonable model.
∼ Gelman (2013, p. 69)

Gelman's (2013) observation that many views of *p*-values are too casual to be accurate is itself surprisingly casual. If the *p*-value cannot be equated with the probability of the tested hypothesis, what does it convey? In this article, we explore the association between the *p*-value produced by significance testing and the posterior (after study) probability of the (null) hypothesis. To anticipate our conclusion, we find logical (i.e., built into Bayes' theorem) and quantitative (after simulation) reasons to think the *p*-value 'significantly' predicts the probability of the hypothesis being true. These associations, being neither trivial nor perfect, suggest that the *p*-value is best understood as a useful diagnostic cue for the task of statistical inference. It should neither be ignored nor burdened with the expectation that it reveals everything the researcher wishes to know.

Although our objective is squarely focused on the inductive power of the *p*-value, we find it impossible to dissociate our investigation from the debate over Null Hypothesis Significance Testing. NHST is the preponderant form of significance testing and thus the main producer of *p*-values in psychology and many other fields of empirical research. Yet, the jerry-built framework of NHST invites a host of other types of criticism that lie beyond the scope of this article. For exposition's sake, we refer to significance testing or specifically to NHST throughout this article as

we explore the properties of *p*-values, but this presentational device does not mean that we endorse all aspects of NHST as it is currently practiced.

Significance testing in its various forms has a long tradition in psychological science, and so do statisticians' concerns and search for alternatives. Significance testing, whether or not it involves null hypotheses, is flawed on logical and probabilistic grounds. It has systematic biases and blind spots. Yet, logical and methodological limitations afflict all methods of inductive inference (García-Pérez, 2016). Hume (1739/1978) famously observed the impossibility of a rational justification of inductive inference. The question he asked, and which we should ask today, is a pragmatic one: how well does a method perform the task placed before it? And by what criteria can we judge a method's worth? In psychological science, much of the critical debate has been focused on NHST, presumably because many researchers use it ritualistically with a narrow focus on the *p*-value, and without understanding its meaning (Meehl, 1998; Gigerenzer, 2004; see also Mayo, 1996; Perezgonzalez, 2015b). Greenland et al. (2016) list no fewer than 25 misconceptions regarding *p*, chief among them the idea that *p* reflects the probability of the research hypothesis being true, that is, Gelman's gripe. Here, we can only briefly sketch the main themes of criticism before considering a specific set of questions in greater depth: what is the association between the *p*-value and the revised probability of the tested hypothesis? What are some of the factors that affect this association? Should these factors matter to the working researcher?

We address these questions with computer simulations. As we progress, it will become clear that we freely draw from distinctive statistical traditions, including Fisher's framework, the Neyman–Pearson paradigm, and Bayesian ideas. We follow this eclectic and pragmatic route in order to obtain answers to our chief questions that may translate into applied practice. We will conclude with reflections on the place of the *p*-value in psychological research and the role it may play in informing, however tentatively, theoretical considerations. Seeing some value in the use of the *p*-value, we do not end with a wholesale condemnation of significance testing (while granting that there may be other sufficient reasons). If, in the course of events, significance testing is abandoned or replaced with, for example, estimation methods (Cumming, 2014) or techniques of Bayesian model comparison (Kruschke, 2013; Kruschke and Lidell, 2017), our analysis might be remembered as a requiem for significance testing and NHST. Then, looking back from the future, we may come to see what we have lost, for better or for worse.

A BRIEF HISTORY OF CRITICISM

A radical conclusion from the critical reception of significance testing is surgical: remove such testing and the *p*-value from research altogether (e.g., Schmidt and Hunter, 1997). Indeed, the journal *Basic and Applied Social Psychology* no longer accepts research articles reporting significance tests (Trafimow and Marks, 2015), while *Psychological Science* nudges authors toward other "preferred methods" (Eich, 2014).[1] We think it self-evident that a decision to ban any particular method should clear a rational threshold. Perhaps a ban is justified if significance testing (and the resulting *p*-value) causes more harm than good. Some believe this to be so (Ioannidis, 2005; but see Fiedler, 2017), but harm and good are elastic concepts; they are difficult to define and measure in a probabilistic world. A more cautious position is to say that the *p*-value should be abandoned if its contribution to scientific progress is too small and if other measures perform better. Here, a difficulty lies in what is meant by 'too small,' or 'better.' Recall Hume's skepticism regarding the appraisal of induction. Scientists trying to evaluate a particular method have no access to truth outside of the inductive enterprise itself – if they did, they would not need induction. A method of inductive inference can be evaluated only indirectly with the help of other inductions. Recognizing this constraint, we attempt to estimate the usefulness of the *p*-value by pragmatically relying on other (mainly Bayesian) modes of induction.

Criticism of *p*-values and significance testing takes several forms. One prominent concern is that researchers misunderstand the process of inference and fail to comprehend the meaning of the *p*-value (Bakan, 1966; Cohen, 1994; Goodman, 2008; Bakker et al., 2016; Greenland et al., 2016). Gelman's epigraphic warning is a notable expression of this view. Another, more serious, criticism is that researchers deliberately or unwittingly engage in practices resulting in depressed *p*-values (Simmons et al., 2011; Masicampo and Lalande, 2012; Head et al., 2015; Perezgonzalez, 2015b; Kunert, 2016; Kruschke and Lidell, 2017). For our purposes, it is essential to note that both these criticisms are matters of education and professional ethics, which need to be confronted on their own terms. We will therefore concentrate on criticism directed at the intrinsic properties of *p*. Chief among these is the recognition that *p*-values show a high degree of sampling variation (Murdoch et al., 2008; Cumming, 2014). Variability suggests unreliability, and unreliability limits validity. The strongest reaction is to conclude that the evidentiary value of *p* is highly uncertain, or even nil. By implication, all substantive claims resting on significance testing should be ignored. Again, this may be an over-reaction. We know of no critics willing to ignore the entire archival record built on significance tests. Can we truly say that we have learned nothing (Mayo, 1996)? If we have learned something, the question is: how much?

Assuming that significance testing has taught us *something*, there remains a strong concern that much of what we think we have learned is – or will turn out to be – false (Murayama et al., 2014). Significance testing is not neutral with respect to the hypothesis being tested. At the limit, as samples become very large, even very small deviations from the hypothesized point (e.g., 0) will pass the significance threshold (Kruschke, 2013; Kruschke and Lidell, 2017). Significance testing is thus biased against the hypothesis being tested (Greenwald, 1975; Berger and Sellke, 1987). Even when the statistical hypothesis (most often the null) is true, the *p*-value will be < 0.05 in 5%

[1] The "preferred methods" include frequentist and Bayesian methods that advocates of each school would regard as incommensurable.

of the cases, and by definition so (Lindley, 1957; Wagenmakers et al., 2016). At the same time, there is also the concern that most empirical samples are not large enough to detect important effects (Cohen, 1962; Sedlmeier and Gigerenzer, 1989). That is, significance testing is not only liable to produce false positives, but also false negatives. Increases in statistical power – which is typically achieved with increases in sample size – will lower *p*-values (see Hoenig and Helsey, 2001, for a formal proof). Both of these (seemingly opposite) concerns, the risk of false positives and the risk of false negatives, imply that many exact replications will fail (Open Science Collaboration, 2015).[2] The meta-problem of uncertain (and low) replicability has caught the attention of the scientific community as well as the general public as it goes to the heart of the question of how much of a contribution scientific research can make to the well-being of those who pay for it.

More criticism does not always do more damage. The idea that *p*-values have no validity conflicts with the view that samples are too small. Yet, both lines of criticism raise the specter of false positives results. Anticipating this concern, Fisher (1935/1971) recommended a *p*-value of 0.05 as a prudent threshold the data should pass before meriting the inference of significance. He regarded this threshold as a *heuristic* rather than a firm or logical one and the *p*-value as a "crude surprise index." "No scientific worker," Fisher (1956, p. 42) wrote, "has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." A variant of the idea that significance testing is biased toward 'positive' results is the argument that the method does not allow for a corroboration of the tested hypothesis. It cannot, by design, detect true negatives. There is only refutation but no confirmation. Some Bayesian scholars consider it critical that the evidence must be allowed to support the inference that the tested hypothesis is indeed true (Kruschke and Lidell, 2017; Rouder et al., 2017). According to this view, it is a prime task of scientific research to detect and document 'invariances,' that is, to show that important phenomena *do not change* even when salient contextual factors suggest that they would (Wagenmakers, 2007; Rouder et al., 2009).[3] Conversely however, and as noted above, significance testing may also miss true effects due to lack of power or precision in measurement (Dayton, 1998; Vadillo et al., 2016) and it may thereby retard scientific exploration (Fiedler et al., 2012; Baumeister, 2016).

One general response to these diverse and partially contradictory criticisms is to place one's hope in very large samples. The call 'Let the data be big!' might draw more applause were it not for the ecological constraints of laboratory research and reduced efficiency of scientific work. Baumeister (2016) recalls that 10 observations per cell used to be the standard in social psychology, but that recently expectations have risen fivefold. Baumeister observes that a commitment to gather very many observations will decelerate the trial-and-error exploration of creative ideas. Sakaluk (2016) observes that many researchers must work with small to medium samples because they lack the resources to collect large samples for every scientific question they ask. Classic methods were developed to provide small-sample statistics whose fidelity should be evaluated. Aside from such constraints, the pursuit of large samples is understandable. Large samples make estimates more reliable and reduce error. In a very large sample, the obtained effect size (for example, *d*) approximates the population effect size (δ) and the *p*-value is highly diagnostic. If the null hypothesis is false, *p* converges on 0; if the null is true, the probability of a false positive is 0.05. Any reduction in sample size reduces this validity, but does not eliminate it.[4] As part of our investigation, we will explore the effect of increasing sample size on the two types of errors, false positives and false negatives.

THE BAYESIAN CONTEXT

If one is to reject a statistical hypothesis, there needs to be sufficient reason for the belief that the hypothesis is false. There needs to be an estimate of the probability of the hypothesis being true given the data, or p(H|D). However, the standard *p*-value is the inverse of this conditional probability, namely the probability of the data (or data more extreme) given the hypothesis, p(D|H) (Wasserstein and Lazar, 2016). When researchers reject the hypothesis, they have presumably inferred a low p(H|D) from a low p(D|H). They cannot simply equate these two conditional probabilities because this would assume a symmetry that is rare in the empirical world (Dawes, 1988; Gelman, 2013). Conversely, they cannot assume that p(D|H) tells them nothing. Kruschke and Lidell (2017) warn that "the frequentist *p*-value has little to say about the probability of parameter values." But how much is little? A lack of symmetry does not mean a lack of association. If there is a positive association between p(D|H) and p(H|D), the former has heuristic validity for the estimation of the latter.

Bayes' Theorem formalizes the matter of inverse probability (Jeffreys, 1961; Lindley, 1983). Before turning to the likelihood version of Bayes' theorem, which is preferred in formal analysis, we consider the probability version, which is more familiar. Here, the probability of the hypothesis given the data is equal to the probability of the data given the hypothesis times the ratio of two unconditional probabilities:

$$p\,(H|D) = p\,(D|H) \times \frac{p\,(H)}{p\,(D)}$$

The unconditional probability of the hypothesis, p(H), is its prior probability, that is, the estimated probability of this hypothesis being true in the absence of evidence. The unconditional probability of the data, p(D), is the probability of the empirical

[2]Replications will fail because samples are too small to detect a true effect, or because they are large enough to expose the original result as a false positive.

[3]A phenomenon must first be discovered before it can be shown to be invariant over contexts, that is, before it can be generalized.

[4]If the population is finite with size N, a sample of size N is exhaustive and necessarily valid. A sample of N-1 is only slightly inferior, and a sample of $N = 1$ remains more informative than no sample at all (Dawes, 1989).

evidence found in light of *any* hypothesis, which comprises the statistical hypothesis (H) and its alternative(s) (∼H). Bayes' Theorem can thus be written as:

$$p\,(H|D) = \frac{p\,(H)\ \times p\,(D|H)}{p\,(H) \times p\,(D|H) + p\,(\sim H) \times p\,(D|\sim H)}$$

The theorem teaches two lessons. First, to simply equate p(H|D) with p(D|H) is to commit a fallacy of reverse inference (Krueger, 2017). Second, to dismiss p(D|H) is to ignore the fact that it is one of the determinants of p(H|D) (Nickerson, 2000; Krueger, 2001; Trafimow, 2003; Hooper, 2009).

Some scholars have noted the association between the *p*-value and the posterior probability of the hypothesis (Greenland and Poole, 2013). Using simple assumptions (see below), one of us estimated the association between p(D|H) and p(H|D) to be $r = 0.38$ (Krueger, 2001). This result offered a clue for why many researchers continue to use practice of significance testing, but it was too weak to have normative force. Trafimow and Rice (2009) replicated this result and concluded that significance testing has little value. How large should this correlation be? It would be reassuring to see a correlation as large as a typical reliability coefficient, that is, a coefficient greater than 0.70. Reliability coefficients rise with the reduction of measurement error. Yet, the correlation between p(D|H) and p(H|D) is not a matter of reliability but a matter of predictive validity. Even if both probabilities were measured with precision, they would not be perfectly correlated. Beliefs of what constitutes an acceptable level of predictive validity vary. For measures that are considered subtle and sensitive, even validity correlations of around 0.3 have been presented as feats of prediction (e.g., Greenwald et al., 2009). We propose that a validity correlation of 0.5 is large enough to warrant scientific and practical interest. This is a realistic aim, and we ask if the *p*-value can meet it.

SAMPLING PROBABILITIES

How well does the *p*-value, p(D|H), predict the criterion measure, p(H|D), that researchers seek when conducting a significance test? Bayes' Theorem implies a positive association. As the *p*-value falls, so does the criterion of truth, p(H|D). If p(H) and p(D|∼H) were constant, the correlation between p(D|H) and p(H|D) would be perfect. Krueger (2001) and Trafimow and Rice (2009) assumed flat and independent distributions for p(H), p(D|H), and p(D|∼H). We replicated their finding ($r = 0.372$) with 100,000 sets of three input probabilities drawn randomly from uniform distributions. The distribution of p(H) was bounded by 0 and 1 and the distributions of p(D|H) and p(D|∼H) were bounded by 0 and 0.5. We then proceeded to use both likelihood ratios and probabilities to compute p(H|D) and we found very similar results. Here, we report only the results obtained with likelihood ratios in line with the Bayesian notion that "only the data actually observed – and not what might have occurred – are needed, so why use the might-have-been at all? (Lindley, 1983, p. 6).[5] Compared with probability ratios, likelihood ratios are less biased against the null hypothesis.[6] When using likelihoods to compute p(H|D), the criterion correlation between p(D|H) and p(H|D) dropped to $r = 0.263$.[7]

Assuming that researchers reject a hypothesis when $p < 0.05$, we asked whether the posterior probability was less than 0.5, that is, whether the hypothesis was more likely to be false than true. This threshold is a heuristic choice; it is prudent in that it avoids judgments of value, importance, or need. Other (especially lower) thresholds may be proposed in light of relevant utility considerations (Lindley, 1983). We then categorized each of the 100,000 simulated experiments in a decision-theoretic outcome table (cf. Swets et al., 2000). The rejection of an improbable hypothesis is a Hit in that this hypothesis is less likely than its alternative in light of the data. In contrast, the rejection of a hypothesis that is still more probable than its alternative is a False Alarm. The retention of a probable statistical hypothesis is a Correct Rejection in standard decision-theoretic terms, but we will refer to it as a Correct Retention (i.e., retaining a probable hypothesis) for ease of exposition. Finally, the failure to reject an improbable hypothesis is a Miss. **Figure 1** displays the four decision-theoretic outcomes[8].

Figure 2A plots the posterior probability of the hypothesis, p(H|D), against the *p*-value, p(D|H). A linear model predicts p(H|D) as 0.585p(D|H) + 0.359; $R^2 = 0.072$. For $p = 0.05$, 0.01, and 0.001, respectively, p(H|D) = 0.389, 0.365, and 0.360. The plot shows a mild concavity, and a second-order polynomial model provides a slightly better fit with $-2.352\text{p(D|H)}^2 + 1.735\text{p(D|H)} + 0.267$; $R^2 = 0.092$. The predicted values for p(H|D) are 0.348, 0.284, and 0.269 for the three benchmarks of *p*. That is, the predicted posterior probability of the hypothesis is in each case below 0.5. Yet, these predicted posterior probabilities are not as low as the corresponding *p*-values, and they decrease more slowly. Statistical regression guarantees this result.[9]

Figure 2A and the top of **Table 1** show the classification of the results. With $p = 0.05$, there are few False Alarms (1.94%). The division of the percent of False Alarms by the total percent of significant results (Hits + False Alarms) yields a 'false alarm ratio' (Barnes et al., 2009). We find that for 19.34% of the significant results the null hypothesis remains more probable than its alternative. A 'miss ratio' is obtained by dividing the percent of Misses by the total percent of non-significant results (Misses + Correct Retentions, 42.03/[42.03+47.95]). For 46.71% of the non-significant results, the null hypothesis is less probable than its alternative. The middle and the bottom parts of **Table 1**

[5]In the standard normal distribution, the correlation between the probability density [$\phi_z(z)$, the height of the curve at point z] and the complement of the cumulative probability (the area under the curve to the right of z) is 0.965 when computed for 400 z-values ranging from 0 to 3.99. When both indices are log transformed, the correlation rises to 0.989.

[6]There is no consensus among Bayesians as to whether probability or likelihood ratios are to be preferred.

[7]We obtained p(H|D) as $\frac{1}{\frac{1}{x}+1}$, where $x = \frac{\text{pdf}(D|H)}{\text{pdf}(D|\sim H)} \times \frac{p(H)}{p(\sim H)}$ and pdf refers to probability density function.

[8]Note that here we refer to any hypothesis as the topic of rejection or retention.

[9]The value of ST can be expressed in terms of Bayesian updating. The posterior odds against the null were 0.367, 0.575, and 0.563 respectively for $p = 0.05$, 0.01, and 0.001.

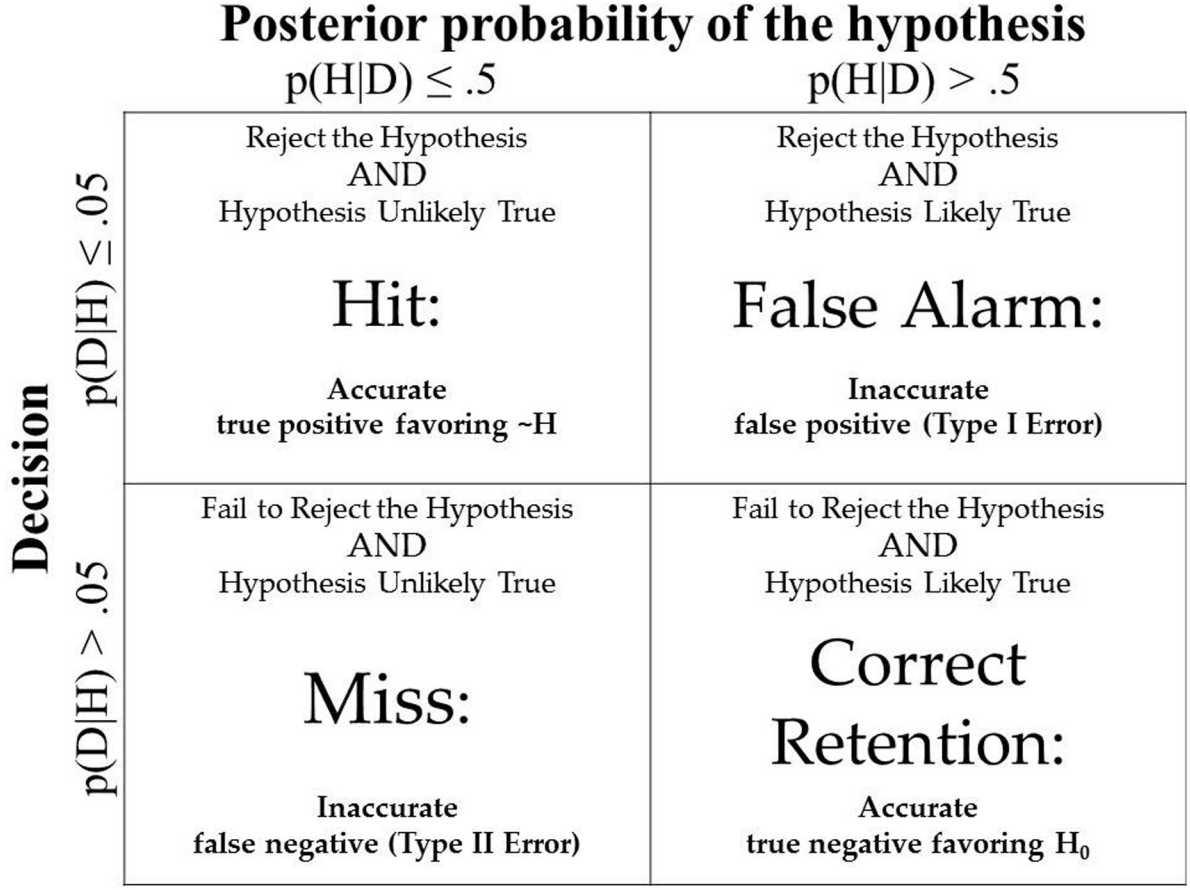

FIGURE 1 | The decision-theoretic context of significance testing.

show that as the *p*-value decreases to 0.01 and 0.001, the false alarm ratio decreases, whereas the miss ratio does not change. In other words, setting a more conservative criterion for the rejection of the hypothesis provides better insurance against false positive inferences, although it does not protect against missing important effects.[10]

Bayes' Theorem treats prior and conditional probabilities as conditionally independent. For any value of p(H), p(D|H) is – in theory – free to vary. Yet, the assumption of independence may not hold in empirical research. Theoretical considerations, past research, and experience-based hunches allow researchers to gauge the riskiness of their hypotheses (Meehl, 1998; Kruschke, 2013; Kruschke and Lidell, 2017). Doing so, researchers will select hypotheses non-randomly, and as a result, the prior probability of the hypothesis, p(H), and the obtained *p*-values become positively correlated. A risky alternative hypothesis (∼H, e.g., Uri can mentally bend spoons when primed with the name 'Geller') means that the probability of the statistical null hypothesis, p(H), is high and it makes a non-significant outcome (p(D|H) > 0.05) likely. With a large effect (∼H: $\delta = 0.8$) being initially either probable (p(H) = 0.1) or improbable (p(H) = 0.9), data will more likely be sampled from the ∼H or the H distribution, respectively. The *p*-value will be smaller in the first case than in the second case, which yields a positive correlation between p(H) and p(D|H). As the effect (d) becomes smaller, the same argument holds, but less strongly so.[11]

We will elaborate this argument in a simulation below. For now we treat it as an ecological constraint and we consider a simulation in which the correlation between p(H) and p(D|H) varied from 0 to 0.9 in steps of 0.1. **Table 2** shows a sharp rise in the criterion correlation between p(D|H) and p(H|D), but only small changes in the prevalence of the two types of error and the overall accuracy of classification (the phi coefficient). Consider the case of r(p(H),p(D|H)) = 0.5. The criterion correlation is 0.628 and p(H|D) is predicted as 1.4p(D|H) + 0.159, $R^2 = 0.395$ (see also **Figure 2B**). For $p = 0.05$, 0.01, and 0.001, respectively, the predicted values of p(H|D) are 0.229, 0.173, and 0.160. The polynomial model is -1.683p(D|H)^2 + 2.243p(D|H) + 0.088; $R^2 = 0.404$, with predicted values of p(H|D) being 0.207, 0.111,

[10]False alarm and miss ratios are frequentist indices. The tabulated data can submitted to Bayesian calculations with identical results (Gigerenzer and Hoffrage, 1995).

[11]Simonsohn et al. (2013) reach the same conclusion with *p*-curve analysis. If p(H) = 1, p(D|H) is uniformly distributed. If p(H) = 0, the distribution becomes increasingly left-skewed (more small *p*-values) as effects become larger.

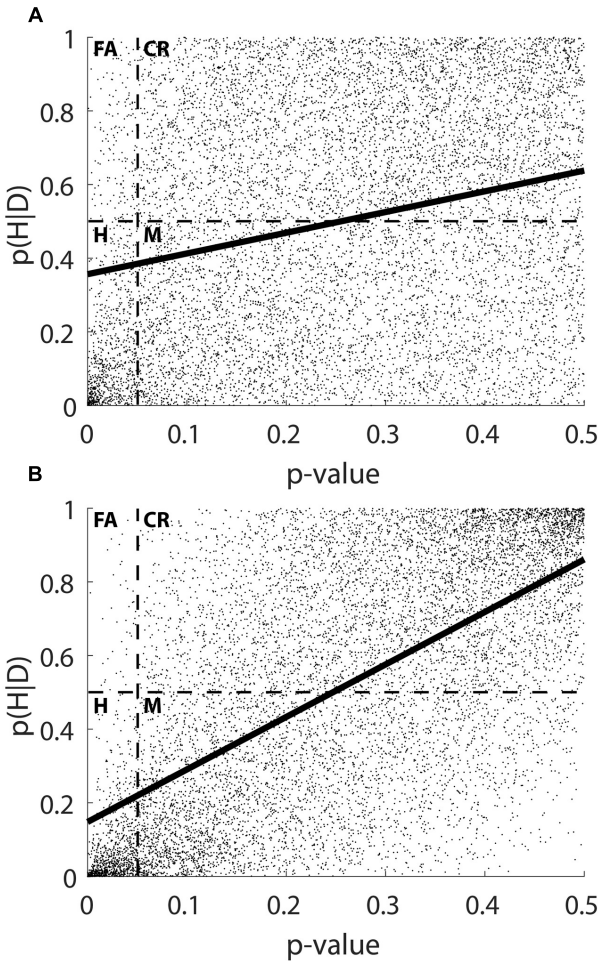

FIGURE 2 | Distribution of p(D|H) and p(H|D) when drawing input terms from uniform distributions. Dashed lines indicate boundary points for classification, with the resulting rectangles capturing each category type (bottom left: Hit; top left: False Alarm; bottom right: Miss; top right: Correct Retention). **(A)** Shows the case for r(p(H),p(D|H)) = 0; **(B)** shows the case for r(p(H),p(D|H)) = 0.5.

and 0.090. In short, the *p*-value predicts the posterior probability of the hypothesis more effectively if it is already correlated with the prior probability. As a comparison, we ran a simulation using a negative correlation, $r = -0.5$, between p(H) and p(D|H), and found a criterion correlation of −0.189. These results suggest that the *p*-value works well when it should, and that it does not when it should not.

We then asked how the correlation between *p* and the probability of the data under the alternative hypothesis, p(D|∼H) affects posterior probabilities. Strong theory provides clear alternatives to the statistical null hypothesis so that the data are either probable under the null or probable under the alternative. In other words, the correlation between p(D|H) and p(D|∼H) should be negative *a priori*. **Table 3** shows that over a range from 0 to −0.9 for this correlation, the criterion correlation became stronger, the false alarm ratio dropped, and the miss ratio varied little. We also used a positive correlation [r between p(D|H) and p(D|∼H) = 0.5] as input and found a very low criterion correlation to $r = 0.132$. In short, a research design that pits two hypotheses against each other so that the data cannot be improbable (or probable) under both allows the *p*-value to reach its greatest inductive potential.

To recapitulate, we saw in the first set of simulations that [1] the *p*-value predicts the posterior probability of the tested hypothesis, [2] this correlation is strongest under the most realistic assumptions, [3] false positive inferences are least likely under the most realistic settings, and that [4] the probability of false negative inferences (Misses) is high. The *p*-value thus appears to have heuristic value for inductive inference. Yet, these simulations are only first approximations. They were limited in that input correlations varied only one at a time. Further, these simulations did not involve a sampling of data from which correlations were computed; they instead sampled probability values and stipulated specific correlations among them. We designed the next round of simulations to address these limitations.

SAMPLING OBSERVATIONS

To obtain values for p(D|H) and p(D|∼H) from sampled data, we generated sets of two normal distributions with 100,000 cases each. In each set, one distribution ($M = 50$, $SD = 10$) was paired with an alternative distribution (M ranging from 50.1 to 60 in steps of 0.1 and $SD = 10$). Standardized effect sizes, δ, thus varied from 0.01 up to 1.0. We then drew mixed samples of 100 observations from each pair of populations, letting the number of observations drawn from the lower distribution range from 10 to 90 in steps of 10. We drew 50 sets of samples for each combined setting of effect size and mixed sampling to generate distributions of means. For each of these 900 distributions, we obtained the z score, its one-tailed values of p(D|H) and p(D|∼H), and the corresponding probability densities. Finally, we varied the prior probability of the hypothesis that μ = 50, p(H), from 0.01 to 0.99 in steps of 0.01 for each of these 900 *p*-values. This process yielded a total of 89,100 simulation experiments [100 steps of δ * 9 steps of sampling proportions * 99 levels of p(H)].

Both conditional probabilities of the data, p(D|H) and p(D|∼H), were independent of the prior probability of the hypothesis, p(H). The overall correlation observed between the two conditional probabilities was 0.200. Of central interest were the criterion correlations between the *p*-value and its inverse conditional, p(H|D), computed for each effect size using likelihood ratios. The mean of these correlations, after Fisher's r-Z-r transformation, was 0.571, mean linear $R^2 = 0.34$, mean polynomial $R^2 = 0.46$. **Figure 3A** plots this correlation, the two error ratios (False Alarm and Miss), and the phi correlations capturing overall categorical accuracy over variations in effect size.

We then returned to the issue of risky vs. safe research in contexts where the tested hypothesis is a statistical null. Researchers often know the difference between a good bet against the null hypothesis and a long shot. To model their inferences,

TABLE 1 | *Crossed proportions of conditional probability terms ($p < 0.05$).*

	p(H|D) ≤ 0.50	**p(H|D) > 0.50**
p(D|H) ≤ 0.05	8.080	1.937
p(D|H) > 0.05	42.030	47.953
Crossed proportions of conditional probability terms ($p < 0.01$).		
	p(H|D) ≤ 0.50	**p(H|D) > 0.50**
p(D|H) ≤ 0.01	1.89	0.14
p(D|H) > 0.01	48.38	49.59
Crossed proportions of conditional probability terms ($p < 0.001$).		
	p(H|D) ≤ 0.50	**p(H|D) > 0.50**
p(D|H) ≤ 0.001	0.22	0.00002
p(D|H) > 0.001	49.40	50.38

TABLE 2 | Positive correlation between p(H) and p(D|H).

***r*(p(H),p(D|H))**	***r*(p(D|H),p(H|D)**	**FA ratio**	**Miss ratio**	**Phi**
0	0.267	0.200	0.465	0.201
0.1	0.343	0.157	0.460	0.229
0.2	0.415	0.120	0.449	0.260
0.3	0.494	0.092	0.444	0.278
0.4	0.565	0.063	0.436	0.302
0.5	0.628	0.046	0.430	0.313
0.6	0.698	0.031	0.425	0.327
0.7	0.760	0.018	0.416	0.338
0.8	0.826	0.008	0.411	0.349
0.9	0.891	0.003	0.405	0.356

FA, false alarm.

we departed from assuming a uniform prior distribution of p(H). Instead, we assumed that researchers had learned enough to consider a bimodal distribution of priors, seeing some hypotheses as being either likely or unlikely to be true, while seeing few hypotheses as equally likely to be true and false.[12] We modeled their inference task by using the posterior probabilities of the hypothesis obtained after the first round of study (i.e., simulation) as the priors for the second round. We thereby obtained a revised value of p(H|D) for each of the 89,100 simulated experiments using the same diagnostic likelihood information as before. With this approach, the average criterion correlation increased to 0.634, mean linear $R^2 = 0.40$, mean polynomial $R^2 = 0.54$. **Figure 3B** shows the criterion correlations as well as the error ratios and the categorical accuracy correlation (phi) as a function of the effect size. Compared with the initial simulation, this second simulation, which granted some knowledge to the researcher, showed a clearer pattern. The criterion correlation increased earlier and more steeply as effect sizes increased and the false alarm ratio was lower for small effects.

[12]This bimodal distribution of p(H|D) can be seen against the Y-axis in **Figures 2A,B**.

Taken together, the two panels of **Figure 3** show that the *p*-values perform most poorly for small effects and best for medium effects. The prevalent type of error depends on the size of the effect. Small effects are easy to miss, whereas large effects are more likely to be falsely declared significant. The simulations reinforce the obvious point that small effects tend to yield higher *p*-values than large effects ($r = -0.642$, see **Table 4**). If a true effect is small and considered improbable *a priori* (p(H) > 0.5), the *p*-value may not be small enough to move p(H|D) below 0.5, thereby yielding an inferential Miss. Conversely, if a true effect is large and considered probable *a priori* (p(H) < 0.5), the *p*-value may be low enough to yield an inferential False Alarm (p(H|D) < 0.5). Significance testing is most efficient for medium effects ($\delta \approx 0.5$). Here, the risks of both types of error are low, and the phi coefficient between decisions based on the *p*-value (significant vs. not) and the estimated posterior probability of the null hypothesis (≤0.5 or >0.5) is high.

To conclude this section, we estimated the criterion correlations for the two rounds of simulation by computing them over the entire set of 89,100 settings. In the initial round of simulations, $r = 0.395$, with a linear prediction being p(H|D) as 0.936p(D|H) + 0.353, $R^2 = 0.156$. For *p*-values of 0.05, 0.01, and 0.001, the predicted probabilities of the null were

TABLE 3 | Negative correlation between p(D|H) and p(D|∼H).

***r*(p(D|H),p(D|∼H))**	***r*(p(D|H),p(H|D)**	**FA ratio**	**Miss ratio**	**Phi**
0	0.260	0.198	0.468	0.199
−0.1	0.287	0.181	0.464	0.213
−0.2	0.311	0.165	0.462	0.225
−0.3	0.345	0.144	0.462	0.236
−0.4	0.363	0.144	0.463	0.234
−0.5	0.390	0.135	0.461	0.242
−0.6	0.411	0.132	0.461	0.245
−0.7	0.437	0.126	0.459	0.249
−0.8	0.461	0.123	0.463	0.248
−0.9	0.492	0.125	0.456	0.253

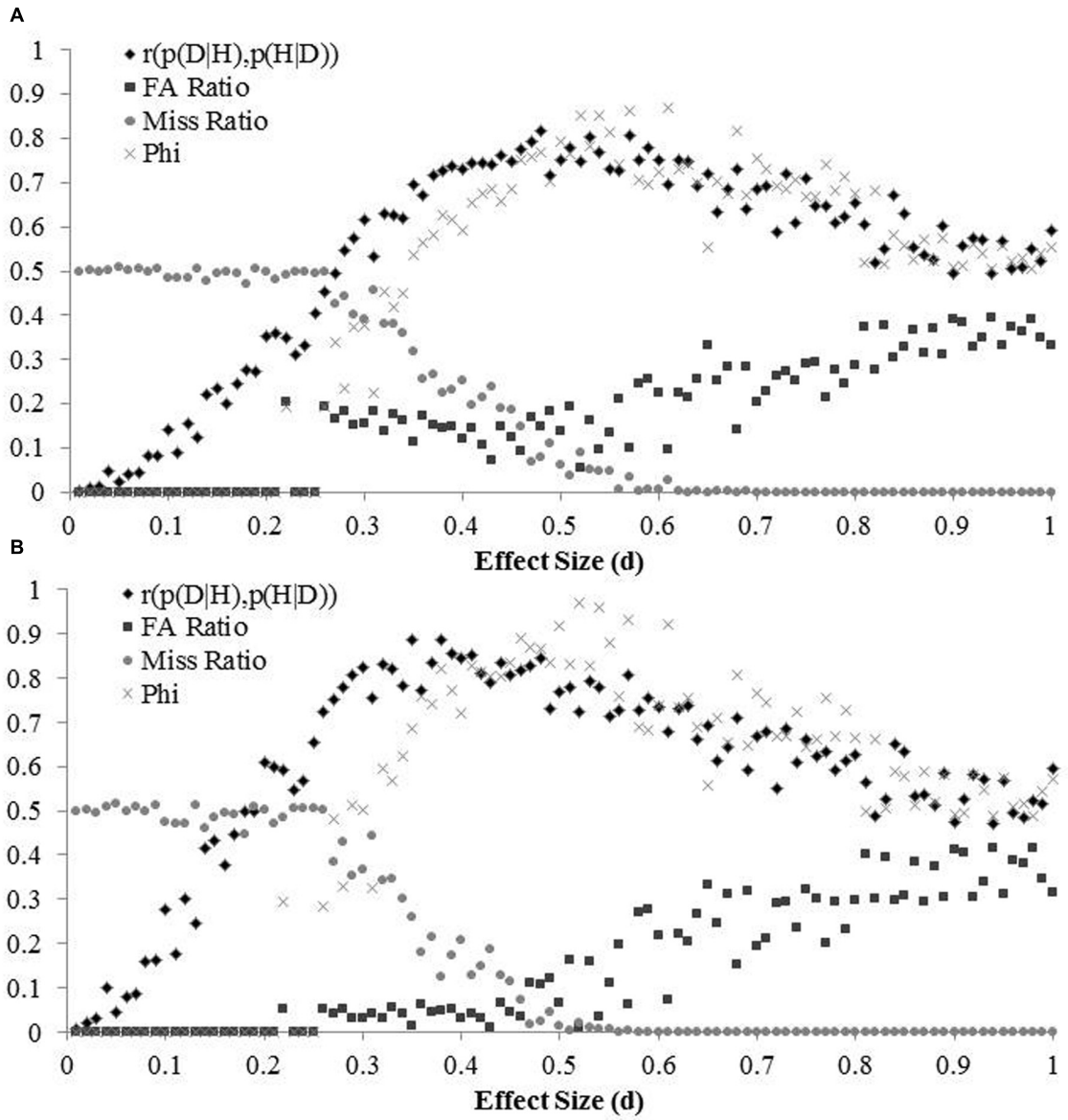

FIGURE 3 | Error rates, the criterion correlation, and the accuracy correlation (phi) over 100 effect sizes (0.01 to 1 in steps of 0.01). The sampling proportion varied from 0.1 (10% of samples from H) to 0.9 (90% of samples from H) in steps of 0.1. **(A)** p(H) varied from 0.01 to 0.99 in steps of 0.01 for each effect size. **(B)** Displays the same variables after revising p(H|D) using the posterior obtained under uniform assumptions.

TABLE 4 | Correlations for a simulation varying sampling proportion from 0.1 to 0.9, effect size from 0.01 to 1.0, and p(H) from 0.01 to 0.99.

	Sampling proportion	δ	p(H)	p(∼H)	p(D|H)	p(D|∼H)	p(H|D)	Updated p(H|D)
δ	0.000	–						
p(H)	0.000	0.000	–					
p(∼H)	0.000	0.000	−1.000	–				
p(D|H)	0.564	−0.642	0.000	0.000	–			
p(D|∼H)	−0.577	−0.636	0.000	0.000	0.200	–		
p(H|D)	0.713	−0.002	0.394	−0.394	*0.395*	−0.400	–	
Updated p(H|D)	0.767	0.000	0.279	−0.279	*0.435*	−0.444	0.969	–
Sample mean	−0.634	0.673	0.000	0.000	−0.800	−0.054	−0.593	−0.601

The criterion correlations are in italics.

0.400, 0.362, and 0.354, respectively. A non-linear fit resulted in $p(H|D) = -5.921p(D|H)^2 + 3.531p(D|H) + 0.258$, $R^2 = 0.273$, yielding posterior probabilities of 0.522, 0.297, and 0.262. The false alarm ratio was lower (25.22%) than the miss ratio (30.70%), although the difference was smaller than in previous simulations. Overall classification accuracy, phi, was 0.438.

In the secondary round of simulations, when assuming an informed researcher, *r* increased to 0.435, with a linear prediction of $1.104p(D|H) + 0.328$, $R^2 = 0.190$, and predicted values of p(H|D) of 0.383, 0.339, and 0.329 for the three benchmarks of *p*. The non-linear model is $-6.254p(D|H)^2 + 3.845p(D|H) + 0.228$, $R^2 = 0.304$, with benchmark predictions of 0.518, 0.2704, and 0.232. The overall false alarm ratio dropped slightly to 0.233 and the overall miss ratio decreased slightly to 0.290. Phi increased slightly to 0.474. **Table 4** shows the correlations among these simulated variables, including both the initial (uniform assumptions) and 'updated' p(H|D).

In these simulations, the *p*-value predicted the posterior probability of the tested (null) hypothesis, but the associations were far from perfect. Second-order (non-linear) models improved prediction, indicating that the linear modeling underestimated the contribution of the *p*-value to inductive inference. Going beyond intuition and back-of-the-envelope analysis, these simulations show lawful patterns in the size of the criterion correlation and the types of error attached to imperfect prediction. We suspect that researchers rarely ask about the criterion correlation between *p* and the posterior of the null. Seeking objectivity, they might hesitate to estimate unknown probabilities. Judging from informal observation, we surmise that researchers worry most about missing effects when planning and conducting a study, whereas they worry most about reporting false effects after having published their own work or when reviewing their colleagues' work.

ARE LARGE SAMPLES BETTER THAN SMALL SAMPLES?

In empirical research, samples vary in size. Limited resources or lack of will can keep samples below levels recommended by power analysis. Contrariwise, some samples exceed the needs of significance testing or parameter estimation (Gigerenzer and Marewski, 2015). Yet, the received wisdom is that large samples are always better, perhaps because large samples resemble what they are intended to represent, namely the population. Larger samples deliver greater statistical power and produce fewer Misses. However, the power perspective obscures the question of false alarm ratios. Much of the critical literature suggests that increases in sample size will protect researchers from making false positive inferences. We ask if this is so.

Building on the foregoing simulations, we chose three effect sizes ($\delta = 0.2$, 0.5, and 0.8), sampled observations, computed their means, and performed one-tailed *z*-tests on 20, 50, 100, or 200 of these means. We let the probability of the tested hypothesis, p(H), and the sampling parameter determine how many samples would be drawn from each distribution, ranging from 0.01 to 0.99 in steps of 0.01. As before, we assessed the criterion correlations between p(D|H) and p(H|D) and the R^2 for both the linear and the non-linear models. To assess the performance of the *p*-value, we again report the two error ratios and the phi coefficients. As before, we proceeded in two steps. In step 1, the prior probability of the hypothesis, p(H), varied independently of the *p*-value. In step 2, we allowed some prior knowledge so that there was a positive correlation between p(H) and p(D|H). To accomplish this, we again used the posterior probability of the null obtained in round 1 as the prior in round 2.

The results are displayed in **Tables 5**, **6** respectively for the first and the second round of simulations. The patterns were similar but clearer in the case of prior knowledge. Larger samples yielded lower *p*-values, and this effect was clearest when effect sizes were small. Importantly, the criterion correlations depended on both the size of the effect and the size of the sample. These correlations increased with sample size N for small effects, were fairly stable for medium effects, and *decreased* for large effects. This interactive pattern may violate intuition, but it highlights the need for caution when expecting large samples to be best. We see that when effects and samples are large, a low *p*-value is a poor predictor of the falsity of the hypothesis. The error ratios provide deeper insights. Perhaps surprisingly, false alarm ratios go up with sample size unless effects are small. Conversely, miss ratios are large for small effects and they decrease with sample size. The combined effects of the two types of error are seen in the phi coefficients. Phi generally tracks (as it has to) the criterion correlation, again showing that the *p*-value is at its diagnostic best for medium effects.

REPLICABILITY

Simulations of significance testing can help estimate the probability of certain errors, but it falls to additional research to help answer the question of whether an error has actually occurred. Additional research addresses the question of replicability. Meant to answer limitations of single studies or sets of studies, replication research reproduces the some of the inferential patterns and problems at a higher level. Mindful of this analogy, we adapted our simulations to see whether the *p*-value can predict the outcome of replication research.

The issue of replicability cuts to the core of empirical science. While conceptions of replicability vary considerably, most scholars seem to agree that the replicability of empirical findings reflects the reliability of method and measurement, which in turn enables and constrains the validity of the empirical results (Asendorpf et al., 2013; Stroebe, 2016). As our investigation targets the properties of the *p*-value, we focus on the probability of re-attaining a statistically significant result once one such a result has been observed. Doing so, we limit ourselves to attempts at exact replication, that is, studies that might yield different *p*-values because of sampling variation and no other reason.

When considering the question of whether their findings might replicate, many researchers look to power analysis. Power analysis is a feature of the Neyman–Pearson theory of

TABLE 5 | Varying sample size and effect size.

δ	*N*	Mdn p	r(p(D|H),p(H|D))	R^2 linear	R^2 poly	FA ratio	Miss ratio	Phi
0.2	20	0.321	0.156	0.024	0.025	0.000	0.503	0.000
	50	0.239	0.340	0.116	0.118	0.192	0.496	0.088
	100	0.157	0.552	0.305	0.319	0.162	0.429	0.316
	200	0.079	0.743	0.552	0.644	0.106	0.222	0.662
0.5	20	0.134	0.643	0.414	0.445	0.147	0.340	0.476
	50	0.032	0.761	0.579	0.747	0.134	0.078	0.786
	100	0.006	0.651	0.424	0.650	0.261	0.000	0.691
	200	0.000	0.519	0.270	0.400	0.340	0.000	0.557
0.8	20	0.032	0.759	0.577	0.742	0.172	0.052	0.764
	50	0.002	0.584	0.341	0.506	0.285	0.000	0.644
	100	0.000	0.482	0.232	0.331	0.369	0.000	0.507
	200	0.000	0.374	0.140	0.203	0.420	0.000	0.404

Round 1 – naïve investigator.

TABLE 6 | Varying sample size and effect size.

δ	*N*	Mdn p	r(p(D|H),p(H|D))	R^2 linear	R^2 poly	FA ratio	Miss ratio	Phi
0.2	20	0.321	0.300	0.090	0.091	0.000	0.507	0.000
	50	0.239	0.583	0.340	0.348	0.051	0.494	0.128
	100	0.157	0.785	0.617	0.655	0.035	0.403	0.433
	200	0.079	0.820	0.672	0.845	0.026	0.158	0.804
0.5	20	0.134	0.826	0.682	0.762	0.031	0.287	0.632
	50	0.032	0.772	0.595	0.840	0.079	0.020	0.899
	100	0.006	0.632	0.400	0.629	0.260	0.000	0.692
	200	0.000	0.507	0.257	0.382	0.344	0.000	0.554
0.8	20	0.032	0.767	0.588	0.817	0.132	0.009	0.846
	50	0.002	0.569	0.323	0.484	0.285	0.000	0.644
	100	0.000	0.478	0.228	0.325	0.364	0.000	0.511
	200	0.000	0.370	0.137	0.199	0.422	0.000	0.403

Round 2 – experienced investigator.

statistics. It is unknown in the Fisherian framework. Power analysis requires the stipulation of a second hypothesis, which is typically a non-null hypothesis or a 'real' difference. Assuming that this alternative hypothesis is true, that is, assuming that p(∼H) = 1, power analysis yields an estimate of the sample size needed to reject the hypothesis H with a desired probability (Cohen, 1988). Power analysis thereby shortcuts the question of *whether*, or *with what probability*, the alternative hypothesis might be true. Instead, it assumes the best possible case, namely p(∼H) = 1, i.e., p(H) = 0. It is also important to note that power analysis ignores the *p*-value of the original experiment. No matter if *p* was 0.05 or 0.00005, the researcher does the same power analysis, asking whether *p* will be at most 0.05 in the replication study. Thus, the *p*-value is not allowed to play any role in the power analysis approach to replicability. If we want to know if the *p*-value is associated with the probability of successful replication, we must modify the conventional power paradigm.

Whereas many researchers are naively optimistic that their findings will replicate, some scholars are staunchly pessimistic. Gigerenzer (in press, p. 11), for example, notes that "the chance of replicating a finding depends on many factors (e.g., [...], most of which the researcher cannot know for sure, such as whether the null or the alternative hypothesis is true.)." Our position is an intermediate one. We submit that researchers can use a two-step process to estimate the probability that a successful exact replication from the *p*-value of the original study (Krueger, 2001). Specifically, researchers can estimate the probability of re-attaining statistical significance by predicting p(∼H|D) from p(D|H) and then multiplying the result with the power index of 1 – β. They estimate p(H|D) by multiplying the observed *p*-value with a regression weight obtained from a simulated criterion correlation between p(D|H) and p(H|D) over a range of possibilities, take the complement of this estimate [i.e., p(∼H|D) = 1-p(H|D)], and multiply the result with the desired power coefficient. To illustrate this approach,

consider two criterion correlations from the initial round of simulations ('sampling probabilities'). The low estimate of the criterion correlation was 0.263, yielding the predicted values of 0.389, 0.365, and 0.360 for p(H|D) given the three benchmark values of *p*. The corresponding replication probabilities are 0.489, 0.508, and 0.512 if $1 - \beta = 0.8$ and 0.550, 0.572, and 0.576 if $1 - \beta = 0.9$. The more representative criterion correlation of 0.628, obtained under the assumption that researchers have some insight into the riskiness of their endeavor, suggests replication probabilities of 0.617, 0.662, and 0.672 for $1 - \beta = 0.8$ and 0.694, 0.744, and 0.756 for $1 - \beta = 0.9$. These probabilities increase inasmuch as researchers are knowledgeable before study (e.g., are able to predict effect sizes), have larger samples, and use non-linear models to predict the posterior probability of the null hypothesis. The data of replication studies then contribute to a cumulative updating of that probability (Moonsinghe et al., 2007).

The precision and the accuracy of these replicability estimates depend on judgment and experience (Miller, 2009). Some of the values we have reported may seem disappointing if researchers are naively optimistic regarding their chances to replicate a significant result (Stanley and Spence, 2014). This may be so because a study result is a recent, salient, and exciting stimulus that demands attention. As such stimuli generally compromise judgment under uncertainty (Dawes, 1988; Kahneman, 2011), misplaced optimism can be expected (Tversky and Kahneman, 1971; Moore and Healy, 2008). Commenting on his own approving summary of studies on social priming (Kahneman, 2011), Kahneman (2017) acknowledged he had "placed too much faith in underpowered studies." Many researchers do (Bakker et al., 2016). Moreover, asking to find $p < 0.05$ in a replication study is a stringent criterion. Finding $p = 0.055$ after having found $p = 0.045$ does not mean that a bold substantive claim has been refuted (Gelman and Stern, 2006). More lenient criteria may be more realistic (Braver et al., 2014). For example, when there is a large disutility in missing a true effect, researchers can ask whether the effect has the same sign (Meehl, 1998) or whether the pooled data yield a *p*-value smaller than the one obtained with the first sample alone (Goh et al., 2016).

To review, our simulations showed that replicability is high inasmuch as (a) the research hypothesis is safe, (b) the *p*-value of the original study is low, and (c) the power of the replication study is high. We also saw that statistical regression constrains replicability. The probability of a successful replication falls below power estimates and below the complement of the *p*-value. This pattern is evident in the report of the Open Science Collaboration (2015). Regression is a fact to be respected rather than an artifact to be fought (Fiedler and Krueger, 2012; Fiedler and Unkelbach, 2014). Even a researcher who shies away from simulation-based assumptions can heuristically predict a successful replication with a probability of about 2/3.[13]

[13]Incidentally, 2/3 is the probability Laplace derived for repeating "a successful" event when the first event emerged against a background of perfect ignorance (Dawes, 1989; Gigerenzer, 2008).

REVIEW AND DISCUSSION

Our goal was to learn how much the *p*-value reveals about the probability of the statistical hypothesis being true. We concur with Gelman (2013) that a casual inference from p(D|H) to p(H|D) has little justification. We found, however, that the two conditional probabilities are positively related. After replicating the criterion correlation of 0.38 in a baseline simulation, we found that the *p*-value and the posterior probability of the hypothesis are more closely linked under more realistic conditions. Many correlations were greater than 0.5, a value we considered necessary for an inferential cue to be useful. We also found that the probabilities of the two decision errors, False Alarms and Misses, depend on conditions other than the *p*-value itself. The size of the assumed effect and its prior probability are critical for the estimation of these errors. One intriguing result was that False Alarms pose a comparatively small problem. Consideration of sample size clarified this issue further. Unless effect sizes were small, larger samples invited more false positives. Large samples thereby *weakened* the *p*-value's predicted value.

Broad conclusions that the *p*-value has no evidentiary value seem overstated. One version of this argument is that a *p*-value, however high, cannot corroborate the tested hypothesis. Indeed, we found that the proportion of Misses was nearly as large as the proportion of Correct Retentions (i.e., correct decisions *not* to reject the null) for most settings. Yet, it is difficult to argue that there is no difference between $p = 0.8$ or 0.08. Meehl anticipated this difficulty when asking "if we were to scrupulously refrain from saying anything like that [that the hypothesis is probably true], why would we be doing a significance test in the pragmatic context" (Meehl, 1998, p. 395).

Meehl (1978) had another significant insight. Noting that significance testing is conventionally used in its weak form, where the hypothesis H is a null hypothesis of no effect, he suggested a stronger use, where it is a non-null (or non-nil) hypothesis, ∼H, that must be nullified, an argument anticipated by Fisher (1956). None of the statistical operations change with this reversal of the conventional frame, but the conceptual shift is considerable. Now a significant result is a strike *against* the hypothesis of interest. In other words, this shift puts significance testing in the service of a Popperian, falsificationist, approach to research (see also Mayo, 1996, for an epistemological treatise).

It is instructive to consider the implications of the present simulation experiments for this falsificationist approach. The *p*-value would be positively related to p(∼H|D), large samples would militate *against* the survival of a theoretical hypothesis, and false negatives would be perceived to be the greatest threat. Meehl deplored that few psychological theories are precise enough to provide hypotheses to be submitted for the strong use of significance testing. Today the situation is much the same. It is an epistemic and theoretical issue, not a limitation of significance testing or the *p*-value.

Finally, we explored the chances that significance will be re-attained. Most researchers eventually ask whether an effect that was statistically significant in an initial study will also be significant in a repeated experiment. Some researchers know

enough to cultivate a healthy skepticism and not assume that a significant result has proven their hypothesis. Clearly, a *p*-value of 0.05 does not mean that the probability of finding $p < 0.05$ again is 0.95.[14] But what is it? Our simulations show that once the posterior probability of the hypothesis is estimated and a power level has been selected, one may be guardedly optimistic about the recovery of a significant result, absent the ethical and educational concerns over questionable research practices.

In research practice, replications are rarely treated probabilistically, and there is a risk of placing too much emphasis on the outcome of a single replication study. The success or failure of a replication study is often treated as the input for another all-or-none decision as to whether an effect is 'real.' Yet, the outcome of a replication study is itself no more decisive than the outcome of the original study. Each additional study makes a smaller incremental contribution to the cumulative evidence. Stopping research after one failed or one successful replication study resembles the much-criticized practice of stopping data collection when significance is obtained (Simmons et al., 2011). Stopping after one failed replication and concluding that a claim has been refuted (i.e., debunked as a false positive) is as questionable as the claim that the initial result proved the case. Our simulations show that a non-significant result is almost as likely to be a Miss (Type II error) as a Correct Retention. Treating each experiment as one data point, one may wish to preset a satisfactory number of experiments, run these experiments, and plot the effect sizes and *p*-values (or use other meta-analytic tools). Individual investigators, however, may find this strategy unrealistic. They struggle with the opportunities and limitations of small-sample statistics, and trust the scientific community to eventually integrate the available data. This strikes us a reasonable mindset.

Current discussions surrounding the replicability of psychological research results are, in part, an outgrowth of the NHST culture.[15] Bayesians, who avoid categorical inferences about hypotheses, also avoid categorical inferences about the success or failure of a replication study. Bayesian methods model the gradual updating and refining of hypotheses, not their categorical acceptance or rejection. Likewise, parameter estimation methods are not concerned with testing and choosing, but with integrating the available evidence. Here, the weighted evidence of an original study and a follow-up provides the best window into nature. We conjecture that some of the skepticism about significance testing is motivated by the desire to overcome the replication crisis. If significance testing is replaced with "preferred methods," the replication crisis is not solved; it is defined away.

[14]However, Gigerenzer (in press) asserts that many researchers fail to muster even this minimal skepticism due to the learned and ritualistic nature of running a statistical test. Doing the dance of NHST as a ritual, they suffer the "crucial delusion that the *p*-value directly specifies the probability of a successful replication (1-p)" (p. 1).

[15]This is one reason for why we include an investigation of replicability in the report.

Though finding heuristic validity in the *p*-value, we do not advocate a protocol where *p*-values shoulder the full burden of inference (Gigerenzer and Marewski, 2015). The practice of statistics is best understood as the judicious use of a toolbox (Gigerenzer, 2004; Senn, 2011). A strategy of "exploring small" as Sakaluk (2016) recommends, while "confirming big," calls for the use of varying techniques whose strengths are best suited to the problem's constraints. Data analysis and inference require experience and judgment (Abelson, 1995; Krantz, 1999). An eclectic and prudent perspective highlights the need for shared ethical standards. Researchers need to be open and capable to analyze their data from a variety of perspectives, using diverse tools. At the same time, they need to ensure that they do not report whichever method yields the most rewarding or desirable outcome (Simmons et al., 2011; Fiedler and Schwarz, 2016).

THE *p*-VALUE IN A POST-HUMEAN WORLD

"Any rational evaluation of the significance test controversy must begin by clarifying the *aim* of inferential statistics." With these words, Meehl (1998, p. 393, italics are his) opened a chapter in which he claimed that the problem is epistemology, not statistics (see also Mayo, 1996). We concur that any discussion of quantitative methods must be informed by reflections on the role of theory in empirical research. Theory is always broader than the available data. Yet, theoretically driven science and hypothesis evaluation depend on evidence. Evidence is limited (there can always be more), whereas theories and hypotheses refer – by design – to a broader, even unlimited, world. The appeal of significance testing is that it honors the need for an inductive leap from the known (the sampled data) to the unknown (a hidden reality). That is, significance testing is embedded in an enterprise of making inferences with statistics. Inferences from data to theory are "risky bets" (Gigerenzer, 2008, p. 20), decisions made under uncertainty. The researcher who (tentatively) rejects a hypothesis bets that this hypothesis is more likely to be false than true. A bettor does not pretend to know for sure.

We have suggested that the *p*-value is a heuristic cue allowing the researcher to estimate the value of the probability of interest, namely p(H|D). A heuristic approach to the reduction of uncertainty is useful if normative methods are not available or computationally too expensive. An alternative to the *p*-value is the Bayesian likelihood ratio, which yields a Bayes factor when multiplied with the prior odds of the hypotheses. If use of the *p*-value is a heuristic, then a full Bayesian analysis may be, according to the Bayesians, the fully rational operation. With perfect subjective confidence, Lindley (1975, p. 106) asserted that "The only good statistics is Bayesian statistics." Setting aside the challenge of selecting a proper prior probability distribution, one may prefer likelihood ratios to *p*-values because they use information about both a hypothesis and its alternatives. Yet, when a specific alternative hypothesis is selected, the likelihood ratio adds surprisingly little – or nothing at all. Senn (2001, p. 200) noted that "the rank order correlation between *p*-values

and likelihood ratio can be perfect for tests based on continuous statistics." Consider the case in which theory predicts a large effect and the data fall between the hypothesis H and the alternative ∼H. Here, the likelihood ratio is confounded with the *p*-value. As the data drift toward ∼H, the *p*-value drops and so does the likelihood ratio. In simulation experiments, García-Pérez (2016) found perfect correlations between log-transformed *p*-values and likelihood ratios, concluding that this must be so because the latter is "only a transformation of the *p*-value, something that can be anticipated from the fact that, like the *p*-value, the Bayes factor [i.e., the likelihood ratio] is determined by the value of the *t*-statistic and the size *n* of the sample" (p. 11). We replicated this result in our own simulations.

Now consider a case in which theory predicts a small effect and the data lie beyond ∼H. Here, the *p*-value under H drops more gently than the probability of the data under ∼H. As a result, the likelihood ratio increases, providing growing relative support for a hypothesis that is becoming ever less likely. The correlation between the logged *p*-value and the likelihood ratio is perfectly negative.

The Bayesian default test also fails to provide much extra information. Wetzels et al. (2011) compared 855 empirical *p*-values with their corresponding default Bayes Factors [i.e., p(∼H|D)/p(H|D)]. The log-log correlation was negative and virtually perfect.[16] Wetzels et al. (2011, p. 295) claimed that "the main difference between default Bayes factors and *p*-values is one of calibration; *p*-values accord more evidence against the null than do Bayes factors. Consider the *p*-values between 0.01 and 0.05, values that correspond to "positive evidence" and that usually pass the bar for publishing in academia. According to the default Bayes factor, 70% of these experimental effects convey evidence in favor of the alternative hypothesis that is only "anecdotal." This difference in the assessment of the strength of evidence is dramatic and consequential." What appears to be a difference in calibration is a rather a difference in words. Most researchers using significance tests consider *p*-values between 0.01 and 0.05 to be significant, whereas most Bayesians view the corresponding Bayes factors as reflecting "anecdotal evidence." They use benchmarks and language suggested by Jeffreys (1961) that are no less heuristic than the benchmarks suggested by Fisher. If $p < 0.01$ were routinely required for significance, the calibration issue would be moot.[17]

Another alternative to significance testing is to abandon heuristic inferences about the probability of a hypothesis altogether. Instead, one may limit statistics to the calculation of descriptive indices such as effect size estimates, confidence intervals, or graphical displays (Tukey, 1977; Cumming, 2012; Stanley and Spence, 2014). These descriptive methods are useful tools in the statistical box, but they avoid making inferences about an uncertain future. We agree with the notion that computing such descriptive measures does little to change the epistemology (or: inference) drawn from a mean and its variability by undermining the researcher's ability to make predictions (Mayo and Spanos, 2011; Perezgonzalez, 2015a). If significance testing were abandoned, the implications would go beyond bidding farewell to the *p*-value. Researchers would be nudged away from thinking in terms of theories and hypotheses. They would be limited to thinking about the data they can see. Those who believe that the future belongs to big data may welcome this view (e.g., Button et al., 2013), but many laboratory experimenters will doubt the attainment of omniscience.

We believe that there is a need for inductive thinking and statistical tools to support inductive inferences.[18] Asking theoretical questions about latent populations enables the researcher to think about the processes that generate the data, which are then ready to be sampled (Fiedler, 2017). A rich psychological theory might describe the way in which the brain/mind produces measurable responses. It is the theorized psychological process that determines what kind of effect one may expect – if that alternative to the null hypothesis is true. For decades, the standard logic of inference has been that if the data are improbable under the null, they are probable under the substantive alternative. This logic appears to carry a grain of truth, the size of which varies.

Discontent with inductive inference is a recurring symptom of uncertainty aversion, which in turn can lead to contradictory complaints. Hearing that *p*-values are terrible and that, by the way, they are not low enough recalls the vacationer's complaint that "The food was horrible – and the portions were so small!" The two complaints nullify each other. We are not concerned with the possibility that some individuals hold both types of belief but with the fact that the field appears to be open to both types. Likewise, it is odd to categorically call for the abandonment of significance testing on the grounds that the method invites categorical inferences. Making strict distinctions between methods that make strict distinctions and methods that do not is an instance of the former method and thus self-contradictory (and perhaps an instance of Russell's 1902, paradox).

To be sure, contradictory critiques do not validate the method under investigation. Indeed, we confess an incoherence of our own. As we noted at the outset, we drew upon ideas from three discrete schools of statistical thought. The emphasis on exact *p*-values comes from the Fisherian school,

[16] See Figure 3 in Wetzels et al. (2011, p. 295). The authors did not compute a correlation coefficient for the plotted values.

[17] Wetzels et al. (2011) assert that "this problem would not be solved by opting for a stricter significance level, such as 0.01. It is well-known that the *p*-value decreases as the sample size, n, increases. Hence, if psychologists switch to a significance level of 0.01 but inevitably increase their sample sizes to compensate for the stricter statistical threshold, then the phenomenon of anecdotal evidence will start to plague *p*-values even when these *p*-values are lower than 0.01." This argument assumes that increasing sample size will lower the *p*-value while leaving the Bayes factor unchanged. How might this be the case if the *p*-value is needed for the computation of the Bayes factor? If some of the researchers had collected more data to lower *p*, then non-linearities should be seen Figure 3 in Wetzels et al.'s (2011). They are not, and neither are they seen in our simulations. It can be shown that raising N, *ceteris paribus*, lowers p(D|H) and p(D|∼H), but not at the same rate (unless the data fall precisely between H and ∼H). As a result, the ratio of the two also drops. To keep the ratio – and thus the Bayes factor – constant, ∼H would need to move away from the data. Moving the research hypothesis while collecting data in order to hold the Bayes factor constant hardly seems to be a recommendable intervention.

[18] This itself is an inductive inference based on past experience, and therefore tautologically true.

the use of power analysis and decision errors comes from the Neyman–Pearson school, and the estimation of posterior probabilities of hypotheses comes from the Bayesian school. Gigerenzer (2004, in press) warned that the tools offered by these schools ought to not be ritually combined, but he did not proscribe any mixing of methods under all circumstances. Hence, our admission is only a partial one. We think that an integration of statistical analysis tools can be attempted and gainfully employed (see Cohen, 1994, for an eloquent example), and we regard our integration as mindful rather than ritualistic.[19]

Our main concern is with the future of statistical practice and how our results might inform it. We submit that the use of significance testing in experimental work with small to medium-sized samples may remain beneficial, especially in cases involving new questions, and assuming that researchers will consider a variety of options from the statistical toolbox. This conclusion resembles Fisher's original advice (see also Cohen, 1990; Abelson, 1995; Wilkinson and The Task Force on Statistical Inference, 1999; Nuzzo, 2014; Sakaluk, 2016). In contrast, the eminent Bayesian Lindley (1975, p. 112) asserted that "all those methods that violate the likelihood principle" should be left to die. Later, one of us predicted that significance testing will be around because it has been around (Krueger, 2001). This prediction was an inductive one, and thus lacked logical force. But the data have supported it. Some critics of significance testing use *p*-values to support their arguments (e.g., Bakker et al., 2016; see Gigerenzer, in press, for a similar observation). We find this ironic but reassuring.

Much care is needed when it comes to a discussion of the limitations of significance testing and the traps they may set. One well-known concern is about the strict enforcement of the 0.05 threshold (which Fisher himself discouraged) and the all-or-none decision-making it begets. Bayesians lament the incoherence of significance testing, by which they mean – among other things – the intransitivity of inferences: if X is significantly greater than Z, but Y is not significantly greater than Z, it does not follow that X is greater than Y. We share these concerns, but regard them, as noted above, as a matter of education. Our principal concern belongs to the predictive validity of the *p*-value. We used a categorization scheme anchored on $p = 0.05$ to compute false alarm and miss ratios only for illustrative purposes.

Another concern is which types of hypothesis researchers select for study in the first place. Using prediction markets, Dreber et al. (2015) concluded that many researchers chase risky research hypotheses, which means that the statistical hypotheses they seek to reject are highly probable *a priori*.[20] Even when these risky hypotheses turn out to be true, their effect sizes are likely small. This conjecture matches the finding that in most natural and cultural fields, the size of a desired reward is inversely related to its probability (Pleskac and Hertwig, 2014). In the context of statistical effects it is easier to imagine how many forces conspire to create small differences or low correlations (i.e., effects) than it is to imagine forces strong enough – and operating unopposed – to create large effects. When seeking significance under such conditions, some researchers bemoan nature's uncooperativeness, while others invest resources to increase the size of their samples. Although this strong-effort strategy raises the probability of finding significance, our simulations suggest that it also raises the false alarm ratio.

Significance testers face a dilemma. In an idealized world, they find a significant result for a novel but risky hypothesis, replicate significance in the lab, publish in a high-impact journal, and see the results replicated by independent labs. Such is the journey of a hero who makes lasting discoveries. Alas, most researchers must accept reality and make a living by corroborating reasonably probable hypotheses. There is no shame in that.

AUTHOR CONTRIBUTIONS

JK and PH contributed equally to this article and author order was determined randomly. JK conducted literature review and theoretical analysis for this article, and drafted the main body of text. PH conducted the simulations and analyses, prepared the tables and figures, and drafted the results.

ACKNOWLEDGMENT

We thank Hilmar Brohmer, Michael Frank, David Freestone, Tim Pleskac, and Johannes Ullrich for helping us improve this manuscript significantly, $p < 0.05$.

[19] The reader may wonder why we do not endorse a full-fledged Bayesian approach. Following orthodox sample statistics, we have treated the data and not the hypotheses as random variables. Bayesians do the opposite. Throughout our treatment, we have assumed competitive testing for sets of two *specific* hypotheses. By contrast, Bayesians consider hypothetical density distributions. As Lindley (1975, p. 108) declared, Bayesian statistics does not only supersede significance testing, but also makes "problems of point estimation disappear: the 'estimate' is the probability distribution and any single value is nothing more than a convenient partial description of this distribution." See Koenderink (2016) for a more balanced view of the strengths and limitations of Bayesian statistics.

[20] The finding that *p*-values tend to be high and successful replications improbable when null hypotheses have high prior probabilities is consistent with our simulation results.

REFERENCES

Abelson, R. P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Erlbaum.

Asendorpf, J., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *Eur. J. Pers.* 27, 108–119. doi: 10.1002/per.1919

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., and van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychol. Sci.* 27, 1069–1077. doi: 10.1177/0956797616647519

Bakan, D. (1966). The test of significance in psychological research. *Psychol. Bull.* 66, 423–437. doi: 10.1037/h0020412

Barnes, L. R., Gruntfest, E. C., Hayden, M. H., Schultz, D. M., and Benight, C. (2009). Corrigendum: false alarm rate or false alarm ratio? *Weather Forecast.* 24, 1452–1454. doi: 10.1175/2009WAF2222300.1

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: winners, losers, and recommendations. *J. Exp. Soc. Psychol.* 66, 153–158. doi: 10.1016/j.jesp.2016.02.003

Berger, J. O., and Sellke, T. (1987). Testing a point null hypothesis: irreconcilability of p values and evidence. *J. Am. Statis. Assoc.* 82, 112–122. doi: 10.1080/01621459.1987.10478397

Braver, S. L., Thoemmes, F. J., and Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspect. Psychol. Sci.* 9, 333–342. doi: 10.1177/1745691614529796

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal social psychological research: a review. *J. Abnor. Soc. Psychol.* 65, 145–153. doi: 10.1037/h0045186

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

Cohen, J. (1990). Things I have learned (so far). *Am. Psychol.* 45, 1304–1312. doi: 10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round (p<.*05*). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. New York, NY: Routledge.

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Dawes, R. M. (1988). *Rational Choice in an Uncertain World*. San, Diego, CA: Harcourt, Brace and Jovanovich.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *J. Exp. Soc. Psychol.* 25, 1–17. doi: 10.1016/0022-1031(89)90036-X

Dayton, P. K. (1998). Reversal of the burden of proof in fisheries management. *Science* 279, 821–822. doi: 10.1126/science.279.5352.821

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15343–15347. doi: 10.1073/pnas.1516179112

Eich, E. (2014). Business not as usual. *Psychol. Sci.* 25, 3–6. doi: 10.1177/0956797613512465

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspect. Psychol. Sci.* 12, 46–61. doi: 10.1177/1745691616654458

Fiedler, K., and Krueger, J. I. (2012). "More than an artifact: regression as a theoretical construct," in *Social Judgment and Decision-Making*, ed. J. I. Krueger (New York, NY: Psychology Press), 171–189.

Fiedler, K., Kutzner, F., and Krueger, J. I. (2012). The long way from error control to validity proper: problems with a short-sighted false-positive debate. *Perspect. Psychol. Sci.* 7, 661–669. doi: 10.1177/1745691612462587

Fiedler, K., and Schwarz, N. (2016). Questionable research practices revisited. *Soc. Psychol. Pers. Sci.* 7, 45–52. doi: 10.1177/1948550615612150

Fiedler, K., and Unkelbach, C. (2014). Regressive judgment: implications of a universal property of the empirical world. *Curr. Dir. Psychol. Sci.* 23, 361–367. doi: 10.1177/0963721414546330

Fisher, R. A. (1935/1971). *The Design of Experiments*. 8th Edn. New York, NY: Hafner.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

García-Pérez, M. A. (2016). Thou shalt not bear false witness against null hypothesis significance testing. *Educ. Psychol. Measure.* 76, 1–32. doi: 10.1177/0013164416668232

Gelman, A. (2013). P values and statistical practice. *Epidemiology* 24, 69–72. doi: 10.1097/EDE.0b013e31827886f7

Gelman, A., and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *Am. Statist.* 60, 328–331. doi: 10.1198/000313006X152649

Gigerenzer, G. (2004). Mindless statistics. *J. Socio-Econ.* 33, 587–606. doi: 10.1016/j.socec.2004.09.033

Gigerenzer, G. (2008). Why heuristics work. *Perspect. Psychol. Sci.* 3, 20–29. doi: 10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G. (in press). *The End of Common Sense: Social Rituals and Surrogate Science*. Berlin: Max Planck Institute for Human Development.

Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* 102, 684–704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., and Marewski, J. (2015). Surrogate science: the idol of a universal method for scientific inference. *J. Manage.* 41, 421–440. doi: 10.1177/0149206314547522

Goh, J. X., Hall, J. A., and Rosenthal, R. (2016). Mini meta-analysis of your own studies: soe arguments no why and a primer on how. *Soc. Pers. Psychol. Compass* 10, 535–549. doi: 10.1111/spc3.12267

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Semin. Hematol.* 45, 135–140. doi: 10.1053/j.seminhematol.2008.04.003

Greenland, S., and Poole, C. (2013). Living with P values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* 24, 62–68. doi: 10.1097/EDE.0b013e3182785741

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., et al. (2016). Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31, 337–350. doi: 10.1007/s10654-016-0149-3

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychol. Bull.* 82, 1–20. doi: 10.1037/h0076157

Greenwald, A. G., Poehlman, T. A., Ulhmann, E. L., and Banaji, M. R. (2009). Understanding and using the implicit association test: III. meta-analysis of predictive validity. *J. Pers. Soc. Psychol.* 97, 17–41. doi: 10.1037/a0015575

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., Jennions, M. D., Barch, D., et al. (2015). The extent and consequences of p-hacking in science. *PLoS Biol.* 13:e1002106. doi: 10.1371/journal.pbio.1002106

Hoenig, J. M., and Helsey, D. M. (2001). The abuse of power. *Am. Statist.* 55, 19–24. doi: 10.1198/000313001300339897

Hooper, R. (2009). The Bayesian interpretation of a P-value depends weakly on statistical power in realistic situations. *J. Clin. Epidemiol.* 62, 1242–1247. doi: 10.1016/j.jclinepi.2009.02.004

Hume, D. (1739/1978). *A Treatise of Human Nature*. Glasgow: William Collins.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kahneman, D. (2017). *Response to Schimmack, Heene, and Kesavan (2017). Replicability-Index, Blog*. Availbale at: https://replicationindex.wordpress.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/

Koenderink, J. (2016). To bayes or not to bayes. *Perception* 45, 251–254. doi: 10.1177/0301006615619309

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *J. Am. Statist. Assoc.* 94, 1372–1381. doi: 10.1080/01621459.1999.10473888

Krueger, J. (2001). Null hypothesis significance testing: on the survival of a flawed method. *Am. Psychol.* 56, 16–26. doi: 10.1037//0003-066X.56.1.16

Krueger, J. I. (2017). "Reverse inference," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, eds S. O. Lilienfeld and I. D. Waldman (New York, NY: Wiley), 108–122. doi: 10.1002/9781119095910.ch7

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* 142, 573–603. doi: 10.1037/a0029146

Kruschke, J. K., and Lidell, T. M. (2017). The new Bayesian statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* 1–29. doi: 10.3758/s13423-016-1221-4

Lindley, D. V. (1957). A statistical paradox. *Biometrika* 44, 187–192. doi: 10.1093/biomet/44.1-2.187

Lindley, D. V. (1975). The future of statistics: a Bayesian 21st century. *Adv. Appl. Probab. (Suppl.)* 7, 106–115. doi: 10.2307/1426315

Lindley, D. V. (1983). Theory and practice of Bayesian statistics. *J. R. Statist. Soc. Ser. D (The Statistician)* 32, 1–11. doi: 10.1111/bmsp.12004

Kunert, J. (2016). Internal conceptual replications do not increase independent replication success. *Psychon. Bull. Rev.* 11:2016. doi: 10.3758/s13423-016-1030-9

Masicampo, E. J., and Lalande, D. R. (2012). A peculiar prevalence of p values just below.*05*. *Q. J. Exp. Psychol.* 65, 2271–2279. doi: 10.1080/17470218.2012.711335

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226511993.001.0001

Mayo, D., and Spanos, A. (2011). "Error statistics," in *Handbook of the Philosophy of Science: Philosophy of Statistics*, Vol. 7, eds P. S. Bandyopadhyay and M. R. Forster (London: Elsevier), 153–198.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46, 806–834. doi: 10.1037/0022-006X.46.4.806

Meehl, P. E. (1998). "The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions," in *What if There Were No Significance Tests?*, eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahway, NJ: Erlbaum), 393–425.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychon. Bull. Rev.* 16, 617–640. doi: 10.3758/PBR.16.4.617

Moonsinghe, R., Khoury, M. J., and Janssens, C. J. W. (2007). Most published research findings are false – but a little replication goes a long way. *PLoS Med.* 4:e28. doi: 10.1371/journal.pmed.0040028.g002

Moore, D. A., and Healy, P. J. (2008). The trouble with overconfidence. *Psychol. Rev.* 115, 502–517. doi: 10.1037/0033-295X.115.2.502

Murayama, K., Pekrun, R., and Fiedler, K. (2014). Resarch practices that can prevent an inflation of false-positive rates. *Pers. Soc. Psychol. Rev.* 18, 107–118. doi: 10.1177/1088868313496330

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). P-values are random variables. *Am. Statist.* 62, 242–245. doi: 10.1198/000313008X332421

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989X.5.2.241

Nuzzo, R. (2014). Statistical errors. *Nature* 506, 150–152. doi: 10.1038/506150a

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Perezgonzalez, J. D. (2015a). Confidence intervals and tests are two sides of the same research question. *Front. Psychol.* 6:34. doi: 10.3389/fpsyg.2015.00034

Perezgonzalez, J. D. (2015b). The meaning of significance in data testing. *Front. Psychol.* 6:1293. doi: 10.3389/fpsyg.2015.01293

Pleskac, T. J., and Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *J. Exp. Psychol. Gen.* 143, 2000–2019. doi: 10.1037/xge0000013

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., and Wagenmakers, E.-J. (2017). Is there a free lunch in inference? *Topics Cogn. Sci.* 8, 520–547. doi: 10.1111/tops.12214

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225

Russell, B. (1902). "Letter to frege," in *From Frege to Gödel*, ed. J. V. Heijenoort (Cambridge, MA: Harvard University Press), 124–125.

Sakaluk, J. K. (2016). Exploring small, confirming big: an alternative system to the new statistics for advancing cumulative and replicable psychological research. *J. Exp. Soc. Psychol.* 66, 47–54. doi: 10.1016/j.jesp.2015.09.013

Schmidt, F. L., and Hunter, J. E. (1997). "Eight common but false objections to the discontinuation of significance testing in the analysis of research data," in *What if There Were No Significance Tests?*, eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah, NJ: Erlbaum), 37–64.

Sedlmeier, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105, 309–316. doi: 10.1037//0033-2909.105.2.309

Senn, S. (2001). Two cheers for P-values? *J. Epidemiol. Biostat.* 6, 193–204. doi: 10.1080/135952201753172953

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *RMM* 2, 48–66.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2013). P-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* 143, 534–547. doi: 10.1037/a0033242

Stanley, D. J., and Spence, J. R. (2014). Expectations for replications: are yours realistic? *Perspect. Psychol. Sci.* 9, 305–318. doi: 10.1177/1745691614528518

Stroebe, W. (2016). Are most published social psychological findings false? *J. Exp. Soc. Psychol.* 66, 134–144. doi: 10.1016/j.jesp.2015.09.017

Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychol. Sci. Public Interest* 1, 1–26. doi: 10.1111/1529-1006.001

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. *Psychol. Rev.* 110, 526–535. doi: 10.1037/0033-295X.110.3.526

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psychol.* 37, 1–2. doi: 10.1080/01973533.2015.1012991

Trafimow, D., and Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *J. Gen. Psychol.* 136, 261–269. doi: 10.3200/GENP.136.3.261-270

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. *Psychol. Bull.* 76, 105–110. doi: 10.1037/h0031322

Vadillo, M. A., Konstantinidis, E., and Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychon. Bull. Rev.* 23, 87–102. doi: 10.3758/s13423-015-0892-6

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* 14, 779–804. doi: 10.3758/BF03194105

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2016). "The need for Bayesian hypothesis testing in psychological science," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, eds S. O. Lilienfeld and I. D. Waldman (New York, NY: Wiley).

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *Am. Statist.* 70, 129–133. doi: 10.1080/00031305.2016.1154108

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison of 855 t tests. *Perspect. Psychol. Sci.* 6, 291–298. doi: 10.1177/1745691611406923

Wilkinson, L., and The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: guidelines and explanations. *Am. Psychol.* 54, 594–604. doi: 10.1037/0003-066X.54.8.594

Conflict of Interest Statement: 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.

The reviewer TW and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Psychology

GENERAL COMMENTARY
published: 23 August 2017
doi: 10.3389/fpsyg.2017.01434

Commentary: The Need for Bayesian Hypothesis Testing in Psychological Science

Jose D. Perezgonzalez *

Business School, Massey University, Palmerston North, New Zealand

Keywords: *p*-value, logic, reductio argument, modus tollens, data testing, statistics

A commentary on

The Need for Bayesian Hypothesis Testing in Psychological Science
by Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions, eds S. O. Lilienfeld and I. D. Waldman (Chichester: John Wiley & Sons), 123–138.

OPEN ACCESS

Edited by:
Ulrich Dettweiler,
University of Stavanger, Norway

Reviewed by:
Kathy Ellen Green,
University of Denver, United States

***Correspondence:**
Jose D. Perezgonzalez
j.d.perezgonzalez@massey.ac.nz

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *09 June 2017*
Accepted: *08 August 2017*
Published: *23 August 2017*

Citation:
Perezgonzalez JD (2017) Commentary: The Need for Bayesian Hypothesis Testing in Psychological Science. Front. Psychol. 8:1434. doi: 10.3389/fpsyg.2017.01434

Wagenmakers et al. (2017) argued the need for a Bayesian approach to inferential statistics in *Psychological Science under Scrutiny*. Their primary goal was to demonstrate the illogical nature of *p*-values, while, secondarily, they would also defend the philosophical consistency of the Bayesian alternative. In my opinion, they achieved their secondary goal but failed their primary one, thereby this contribution. I will, thus, comment on their interpretation of the logic underlying *p*-values without necessarily invalidating their Bayesian arguments.

Historical criticisms (e.g., Harshbarger, 1977, onwards) have already delved in the illogical nature of null hypothesis significance testing (NHST)—a mishmash of Fisher's, Neyman-Pearson's, and Bayes's ideas (e.g., Gigerenzer, 2004; Perezgonzalez, 2015a). Wagenmakers et al.'s original contribution is to generalize similar criticisms to the *p*-value itself, the statistic used by frequentists when testing research data.

Wagenmakers et al. assert that Fisher's disjunction upon obtaining a significant result—i.e., either a rare event occurred or H_0 is not true (Fisher, 1959)—follows from a logically consistent *modus tollens* (also Sober, 2008): If P, then Q; not Q; therefore not P, which the authors parsed as, If H_0, then not y; y; therefore not H_0.

"Y" is defined as "the observed data... [summarized by] the *p*-value" (p. 126). Therefore, their first premise proposes that, if H_0 is true, the observed *p*-values cannot occur (also Cohen, 1994; Beck-Bornholdt and Dubben, 1996). This seems incongruent, as the first premise of a correct *modus tollens* states a general rule—H_0 implies "not y"—while the second premise states a specific test to such rule—"this y" has been observed. If the authors meant for "y" to represent "significant data" as a general category in the first premise and as a specific realization in the second, a congruent *modus tollens* would ensue, as follows (also Pollard and Richardson, 1987):

If H_0, then not $p <$ sig; $p <$ sig (observed); therefore not H_0 (1)

Wagenmakers et al.'s (also Pollard and Richardson, 1987; Cohen, 1994; Falk, 1998) main argument is that a correct *modus tollens* is rendered inconsistent when made probabilistic, as follows:

If H_0, then $p <$ sig very unlikely; $p <$ sig; therefore probably not H_0 (2)

There are, however, three problems with (2), problems which I would like to comment upon. One problem is stylistic: The first premise states a redundant probability; that is, that a significant result—which already implies an unlikely or improbable event under H_0—is unlikely. Therefore, the syllogism could be simplified as follows:

If H_0, then $p <$ sig; $p <$ sig; therefore probably not H_0 (3)

Correction (3) now highlights another of the problems: The second premise simply affirms that an unlikely result just happened (also Cortina and Dunlap, 1997), something which is neither precluded by the first premise (no contrapositive ensues; Adams, 1988) nor formally conducive to a logical conclusion under *modus tollens* (Evans, 1982). Indeed, in the examples given (also by Cohen, 1994; Beck-Bornholdt and Dubben, 1996; Cortina and Dunlap, 1997; Krämer and Gigerenzer, 2005; Rouder et al., 2016), Tracy is a US congresswoman, Francis is the Pope, and John made money at the casino, each despite their odds against. Yet, none of those realizations deny the consequents. A correction, following Harshbarger (1977) and Falk (1998), would state:

If H_0, then not $p <$ sig; $p <$ sig; therefore probably not H_0 (4)

Correction (4) brings to light the most important problem: *Modus tollens* is in the form, If P, then Q; not Q; therefore not P. Thus, whenever the consequent (Q) gets denied in the second premise, it leads to denying the antecedent (P) in the conclusion. Such operation ought to prevail with probabilistic premises, as well (e.g., Oaksford and Chater, 2001, 2009; Evans et al., 2015), whereby a probable consequent (Q_p) may be denied without its probability warranting transposition onto a non-probabilistic antecedent (P). For example, if all red cars (P) have a 95% chance of getting stolen ($Q \geq 0.95$) and we learn of a Lamborghini with a lesser probability of so disappearing (not $Q \geq 0.95$), it is logical to conclude that the Lamborghini is not red (not P).

In comparison, Bayesian logic allows for the antecedent to be probable. For example, if John always submits to Nature (Q) whenever his subjective probability of getting published soars above 20% ($P > 0.2$), yet he is not submitting his latest article (not Q), it is logical to conclude that he probably expects no publication (not $P > 0.2$).

We can, thus, envisage P or Q, or both, as probable without either warranting inter-transposition of their probabilities, which brings us back to a valid *modus tollens* (1). Said otherwise, while Bayesian statistics allow for the antecedent to be probable (P_p), Fisher's and Neyman-Pearson's tests assume exact antecedents (P); therefore, a probabilistic conclusion does not hold with frequentist tests (Mayo, 2017).

It ought to be noted that the *p*-value is a statistic descriptive of the probability of the data under H_0 [$p(D|H_0)$] (Perezgonzalez, 2015b). The *reductio ad absurdum* argument may be informed by, but it is not dependent on, such *p*-value, the *reductio* being determined exclusively by the chosen level of significance. For "it is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him" (Fisher, 1960, p.13).

In conclusion, the technology of frequentist testing holds their *modus tollens* logically. Wagenmakers et al.'s criticism of the *p*-value is faulty in that they allow for a probability transposition not warranted either by *modus tollens* or by the technical apparatus of Fisher's and of Neyman-Pearson's tests. This critique, however, does not extend to their Bayesian argumentation, an approach much needed for testing hypotheses—rather than just testing data—in contemporary science.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Adams, E. W. (1988). Modus tollens revisited. *Analysis* 48, 122–128. doi: 10.1093/analys/48.3.122

Beck-Bornholdt, H. P., and Dubben, H. H. (1996). Is the Pope an alien? *Nature* 381:730. doi: 10.1038/381730d0

Cohen, J. (1994). The earth is round (p < 0.05). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing. Psychol. *Methods* 2, 161–172. doi: 10.1037/1082-989X.2.2.161

Evans, J. St. B. T. (1982). *The Psychology of Deductive Reasoning*. London: Routledge & Kegan Paul.

Evans, J. St. B. T., Thompson, V. A., and Over, D. E. (2015). Uncertain deduction and conditional reasoning. Front. *Psychol.* 6:398. doi: 10.3389/fpsyg.2015.00398

Falk, R. (1998). In criticism of the null hypothesis statistical test. Am. *Psychol.* 53, 798–799. doi: 10.1037/0003-066X.53.7.798

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference, 2nd Edn*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1960). *The Design of Experiments, 7th Edn*. Edinburgh: Oliver and Boyd.

Gigerenzer, G. (2004). Mindless statistics. J. Soc. *Econ.* 33, 587–606. doi: 10.1016/j.socec.2004.09.033

Harshbarger, T. R. (1977). *Introductory Statistics: A Decision Map, 2nd Edn*. New York, NY: Macmillan.

Krämer, W., and Gigerenzer, G. (2005). How to confuse with statistics or: the use of misuse of conditional probabilities. Stat. *Sci.* 20, 223–230. doi: 10.1214/088342305000000296

Mayo, D. G. (2017). If You're Seeing Limb-Sawing in p-value Logic, You're Sawing off the Limbs of Reductio Arguments [Web Log Post]. Available online at: https://errorstatistics.com/2017/04/15/if-youre-seeing-limb-sawing-in-p-value-logic-youre-sawing-off-the-limbs-of-reductio-arguments/

Oaksford, M., and Chater, N. (2001). The probabilistic approach to human reasoning. Trends Cogn. *Sci.* 5, 349–357. doi: 10.1016/S1364-6613(00)01699-5

Oaksford, M., and Chater, N. (2009). Précis of bayesian rationality: the probabilistic approach to human reasoning. Behav. *Brain Sci.* 32, 69–84. doi: 10.1017/S0140525X09000284

Perezgonzalez, J. D. (2015a). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front. *Psychol.* 6:223. doi: 10.3389/fpsyg.2015.00223

Perezgonzalez, J. D. (2015b). P-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations." Front. *Psychol.* 6:341. doi: 10.3389/fpsyg.2015.00341

Pollard, P., and Richardson, J. T. E. (1987). On the probability of making type I errors. Psychol. *Bull.* 102, 159–163. doi: 10.1037/0033-2909.102.1.159

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., and Wagenmakers, E. J. (2016). Is there a free lunch in inference? Top. Cogn. *Sci.* 8, 520–547. doi: 10.1111/tops.12214

Sober, E. (2008). *Evidence and Evolution. The Logic Behind the Science*. Cambridge: Cambridge University Press.

Wagenmakers, E. J., , Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). "The need for Bayesian hypothesis testing in psychological science," in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, eds S. O. Lilienfeld and I. D. Waldman (Chichester: John Wiley & Sons), 123–138.

Conflict of Interest Statement: 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.

frontiers in Psychology

GENERAL COMMENTARY
published: 27 September 2017
doi: 10.3389/fpsyg.2017.01715

Commentary: Psychological Science's Aversion to the Null

Jose D. Perezgonzalez[1]*, *Dolores Frías-Navarro*[2] and *Juan Pascual-Llobell*[2]

[1] Business School, Massey University, Palmerston North, New Zealand, [2] Department of Methodology of the Behavioral Sciences, Universitat de València, Valencia, Spain

Keywords: data testing, hypothesis testing, null hypothesis significance testing, effect size, falsificationism, statistics

A commentary on

Psychological Science's Aversion to the Null
by Heene, M., and Ferguson, C. J. (2017). Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions, eds S. O. Lilienfeld and I. D. Waldman (Chichester: John Wiley & Sons), 34–52.

Heene and Ferguson (2017) contributed important epistemological, ethical and didactical ideas to the debate on null hypothesis significance testing, chief among them ideas about falsificationism, statistical power, dubious statistical practices, and publication bias. Important as those contributions are, the authors do not fully resolve four confusions which we would like to clarify.

One confusion is equating the null hypothesis (H_0) with randomness when "chance" actually resides in the sample. We can, indeed, read three different instances of randomness in the text: associated with the sample on pages 36 (trial performance) and 37; associated with the alternative hypothesis (H_A) on page 41 ("less likely to observe mean differences…far off the true…mean difference of 0.7"); and associated with H_0 throughout the text, starting on page 36. In reality, H_0 simply claims a population non-effect (H_0: $\Delta = 0$) while H_A claims a constant effect (e.g., H_A: $\Delta = 0.7$), their corresponding distributions assuming random sampling variation in both cases. It is in the (random) sample where "chance" resides, as by chance we may pick a sample which shows a given effect (e.g., $\delta = 0.3$) when the true effect in the population is either "0" (H_0) or "0.7" (H_A). Frequentist tests only assess the probability of getting the observed sample effect under H_0 while Bayesian statistics also assesses the probability of such effect under H_A (e.g., Rouder et al., 2009). Therefore, the *p*-value does not inform about a hypothesis of chance but about the probability of the data under H_0 (Fisher, 1954).

A second issue confuses power with missing true effects, something explicitly expressed on page 42 but also suggested when discussing sample sizes throughout the text (p. 36 onwards). The underlying argument is that larger sample sizes allow for achieving statistical significance so that a true effect may not be missed—something which is, at the same time, portrayed as unethical, e.g., p. 36, and ludicrous, e.g., p. 44. In reality, "we cannot manipulate population effect sizes" (p. 41), as they are deemed constant in the population (e.g., H_A: $\Delta = 0.7$), and a significant result at 50% power will not be missed at 80% power. As Heene and Ferguson's Figures 3.1A,C show, power simply moves the goalposts on the real line, reducing the Type II error (β), while the larger sample size also reduces the standard error. By moving the goalposts, smaller (by chance) sample effects get associated with H_A, which is a correct association as long as there is a true population effect. Thus, power is there not to prevent missing effects due to small sample sizes but to be able to justify whether we could plausibly accept H_0 when results are not significant (Neyman, 1955; Cohen, 1988).

OPEN ACCESS

Edited by:
Hannes Schröter, German Institute for Adult Education (LG), Germany

Reviewed by:
Daniel Bratzke, Universität Tübingen, Germany

***Correspondence:**
Jose D. Perezgonzalez
j.d.perezgonzalez@massey.ac.nz

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *30 May 2017*
Accepted: *19 September 2017*
Published: *27 September 2017*

Citation:
Perezgonzalez JD, Frías-Navarro D and Pascual-Llobell J (2017) Commentary: Psychological Science's Aversion to the Null. Front. Psychol. 8:1715. doi: 10.3389/fpsyg.2017.01715

A third issue is about falsificationism (pp. 35–37), which the authors argue cannot happen in psychology because we never accept H_0, only reject it or fail to reject it. In reality, frequentist tests are logically based on *modus tollens*, the valid argument form for the falsification of statements (Perezgonzalez, 2017a). H_0 is simply the contrapositive of our research hypothesis, and denying H_0 allows us to affirm the latter. Therefore, frequentist tests are eminently falsificationist, attempting to disprove H_0 via *reductio* arguments (p, α; Mayo, 2017). Indeed, H_0 does not even need to be "zero" in the population: We could perfectly substitute the actual value of our H_A, so that we may prove the theory false with a significant result (the "strong" test purported by Meehl, 1997).

A fourth issue is whether we always need to be in the position of accepting H_0 (something argued on pages 36–37). This is not necessarily so. Just testing H_0 as for rejecting it is suitable when we are only interested in learning about our research hypothesis (e.g., does the treatment have an effect?—Perezgonzalez, 2016). In such context, H_0 provides a precise statistical hypothesis for carrying out the test and, because the actual parameter (Δ) is unknown, it only provides informative value via its rejection (Fisher, 1954), H_0 acting merely as a "straw man" (Cortina and Dunlap, 1997). This testing procedure was not only developed in the context of small samples (Fisher, 1954) but the lack of a specific H_A precludes the control of Type II errors and of power. (A way forward would be to assess the effects warranted under H_0—Mayo and Spanos, 2006—or to control sample size via a sensitiveness analysis—Perezgonzalez, 2017b).

If we wish to be able to accept H_0, then we are stating that we are also interested in the potential demise of our intervention (i.e., if the treatment has no effect, we want to make sure it is akin to placebo; Perezgonzalez, 2016). This testing seems similar to Fisher's, but it requires active control of the severity with which the alternative hypothesis is to be tested (ideally, ≥80% power; Neyman, 1955; Cohen, 1988). Such control necessarily means more information—a precise alternative hypothesis (e.g., H_A: $\mu_1 - \mu_2 = 0.7$, vs. H_0: $\mu_1 - \mu_2 = 0$) and a specified Type II error for H_A (e.g., $\beta = 0.20$)—so that the power of the test can be managed (given α, β, and N). This approach not only allows for accepting H_0 but also illustrates that power is only relevant for such purpose, not for rejecting H_0. Such approach, and similar ones, have also been available since Fisher's tests of significance (e.g., Neyman and Pearson, 1928; Jeffreys, 1939).

As final note, frequentist approaches only deal with the probability of data under H_0 [$p(D|H_0)$]. If we want to say anything about the (posterior) probability of the hypotheses, then a Bayesian approach is needed in order to confirm which hypothesis is most likely given both the likelihood of the data and the prior probabilities of the hypotheses themselves (Jeffreys, 1961; Gelman et al., 2013).

AUTHOR CONTRIBUTIONS

JDP initiated and drafted the general commentary. DF and JP contributed theoretical background and feedback. All authors approved the final version of the manuscript for submission.

REFERENCES

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* New York, NY: Psychology Press.

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychol. Methods* 2, 161–172. doi: 10.1037/1082-989X.2.2.161

Fisher, R. A. (1954). *Statistical Methods for Research Workers, 12th Edn.* Edinburgh: Oliver and Boyd.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, 3rd Edn.* Boca Raton, FL: CRC Press.

Heene, M., and Ferguson, C. J. (2017). "Psychological science's aversion to the null, and why many of the things you think are true, aren't," in *Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions*, eds S. O. Lilienfeld and I. D. Waldman (Chichester: John Wiley & Sons), 34–52.

Jeffreys, H. (1939). *Theory of Probability.* Oxford: Clarendon Press.

Jeffreys, H. (1961). *Theory of Probability, 3rd Edn.* Oxford: Clarendon Press.

Mayo, D. G. (2017). *If you're Seeing Limb-Sawing in p-Value Logic, You're Sawing Off the Limbs of Reductio Arguments [Web log post].* Available online at: https://errorstatistics.com/2017/04/15/if-youre-seeing-limb-sawing-in-p-value-logic-youre-sawing-off-the-limbs-of-reductio-arguments/.

Mayo, D. G., and Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *Br. J. Philos. Sci.* 57, 323–357. doi: 10.1093/bjps/axl003

Meehl, P.E. (1997). "The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions," in *What If There Were No Significance Tests?* eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah: Erlbaum), 393–425.

Neyman, J. (1955). The problem of inductive inference. *Commun. Pure Appl. Math.* 8, 13–45. doi: 10.1002/cpa.3160080103

Neyman, J., and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 20A, 175–240. doi: 10.2307/2331945

Perezgonzalez, J. D. (2016). Commentary: how Bayes factors change scientific practice. *Front. Psychol.* 7:1504. doi: 10.3389/fpsyg.2016.01504

Perezgonzalez, J. D. (2017a). Commentary: the need for Bayesian hypothesis testing in psychological science. *Front. Psychol.* 8:1434. doi: 10.3389/fpsyg.2017.01434

Perezgonzalez, J. D. (2017b). *Statistical Sensitiveness for the Behavioral Sciences.* Available online at: https://osf.io/preprints/psyarxiv/qd3gu.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237. doi: 10.3758/PBR.16.2.225

Conflict of Interest Statement: 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.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 27 August 2019
doi: 10.3389/fpsyg.2019.01949

Should We Say Goodbye to Latent Constructs to Overcome Replication Crisis or Should We Take Into Account Epistemological Considerations?

*Barbara Hanfstingl**

Institute of Instructional and School Development, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

This paper discusses theoretical and epistemological problems concerning validity of psychological science in the context of latent constructs. I consider the use of latent constructs as one reason for the replicability crisis. At the moment, there exist different constructs describing the same psychological phenomena side by side, and different psychological phenomena that are reflected by the same latent construct. Hagger called them déjà-variables, which lead to a decreasing validity of measurements and inhibit a deeper understanding of psychological phenomena. To overcome this problem, I suggest a shift of theoretical and epistemological perspective on latent constructs. One main point is the explicit consideration of latent constructs as mental representations, which change objects and are changed by objects via assimilative and accommodative processes. The explicit orientation toward assimilation and accommodation allows the control of normally automatized processes that influence our understanding of psychological phenomena and their corresponding latent constructs. I argue that assimilation and accommodation are part of our research practice anyway and cause the mentioned problems. For example, taking a measurement is an assimilative process, and thus a high measurement error should lead to an increase of accommodative processes. Taking into account these considerations, I suggest consequences for research practices, for individual researchers and for the philosophy of science.

Keywords: epistemology, latent constructs, assimilation, accommodation, déjà-variables, assimilation bias, over-accommodation, over-assimilation

OPEN ACCESS

Edited by:
Cesar Merino-Soto, University of San Martín de Porres, Peru

Reviewed by:
Fernando Marmolejo-Ramos, University of South Australia, Australia
Stephen Humphry, University of Western Australia, Australia

***Correspondence:**
Barbara Hanfstingl
barbara.hanfstingl@aau.at

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *01 February 2019*
Accepted: *08 August 2019*
Published: *27 August 2019*

Citation:
Hanfstingl B (2019) Should We Say Goodbye to Latent Constructs to Overcome Replication Crisis or Should We Take Into Account Epistemological Considerations? Front. Psychol. 10:1949. doi: 10.3389/fpsyg.2019.01949

INTRODUCTION

In this paper, I argue that replication problems in empirical psychology are not only due to statistical and methodological artifacts but also due to a lack of epistemological clarity. I structure my argument around the following four points: First, I state validity problems that can emerge when latent constructs are used to explain psychological phenomena when research is oriented mainly toward positivism. Second, I show how these problems can be seen through a different epistemological perspective, namely an adaption of Piaget's psychogenesis with a focus on assimilation and accommodation. Third, I describe examples from psychological research where the concept of assimilation and accommodation helps to understand phenomena where over-assimilation and over-accommodation disturb the achievement of equilibration. And fourth, I delineate consequences at the level of research methods, researchers and for philosophy of science.

The development, description, and investigation of latent constructs (e.g., personality constructs) is a core focus in psychological research. Despite the high development of statistics, the effective and sustainable validation of latent constructs still remains a huge challenge. The call for a higher validity of latent constructs and their generalizability is an issue that has been discussed over many decades in psychological research and adjacent disciplines **(e.g.,** Sackman, 1974**;** Skinner, 2007**;** Rossiter, 2008**;** Johnston et al., 2014**;** Fiedler, 2017**;** Swami et al., 2017**)**. There are two interrelated reasons why the striving for validity of latent constructs is still one of the main challenges. The first lies in the general tradition of science and scientific practice. According to Bickhard and Campbell (2005), psychology has a strong positivistic tradition which was influenced by Ernst Mach. This does not seem to be a problem at first glance. Burrhus Frederic Skinner, for example, took Mach's approach "as chief basis for his own positivistic views of science" (Smith, 1986, p. 264). However, a pure positivistic perspective on scientific issues can lead to severe validity problems. As Popper (2005) pointed out the weaknesses of positivism for all scientific disciplines, psychological researchers are affected by this issue in a special way. Bickhard and Campbell (2005) still attribute a high influential power in psychological research to neo-Machian positivism because the idea of operationism fosters a positivistic perspective on psychological issues.

The second reason for the validity problem is the use of latent constructs. Their application to make psychology evidence-based even for non-observable phenomena bears boon and bane at the same time. The boon is that now we can investigate non-observable phenomena empirically. The bane is that when researchers started to focus on statistical procedures to calculate latent constructs, they also started to ignore epistemological rules and knowledge that can be drawn solely by theoretical and logical conclusions. For example, Michell (2013) sees the problem in the focus on constructs and differentiates between psychological scientists and scientists in traditional sciences like physics or chemistry. The first concentrate on constructs while the latter concentrate on theoretical concepts. In other words, a psychological researcher thinks in constructs rather than in theories. However, Sherry (2011) used the example of the development of thermometers, which began with the observation of qualitative temperature observations, to argue that the difference between physics and psychology is not given by the difference between construct and theory. I agree with this argument, with one limitation. There is no doubt that the development process of thermometers and psychometric scales is very similar; the younger the process, the more similarities there seem to be. In the meantime, however, physicists have found theoretical foundations – be it the absolute zero point, Brownian motion or gas laws – which all influence the temperature calculably and thus create a basis for temperature beyond thermometers. In psychology we can at best only inaccurately deduce measurements from theories or theories from measurements.

There is no doubt that methodological rules which dominate construct-based research are mandatory but ultimately insufficient for scientific progress and cannot compensate for epistemological or even theoretical considerations (e.g., Fiedler, 2017). Edelsbrunner and Dablander (2018) could show that psychological modeling and scientific reasoning do not always follow a logical procedure. Heene (2013) describes in a very restrictive way why no approach, neither additive conjoint measurement nor modeling of structural equations or item-response theory, can solve the problem of measurement from a purely mathematical point of view and concludes that perhaps "human cognitive abilities and personality traits are simply not quantitative" (p. 3). Here, I would add the idea that cognitive abilities and personality traits might not solely be quantitative. From a metrological perspective, Uher (2018) shows us which epistemological and methodological aspects in most psychological studies are ignored, with a marked reduction of the validity of those studies as a consequence. Recently, Trendler (2019a) revisited an ongoing debate about the justified use of conjoined measurement in psychological research (see also Krantz and Wallsten, 2019; Michell, 2019; Trendler, 2019b).

To summarize, there is a tendency toward (a) positivism and (b) statistic methodical orientation with a coincident lack of theoretical and epistemological orientation. In this paper, I argue that these two reasons bear one of the main responsibilities for the replication crisis. They account for the problem of many overlapping psychological findings that exist side by side, validated within one methodological approach, but bringing them together on a theoretical level fails plenty of times. In psychological literature, we often find the statement that there is "no single," "no distinct," or "no homogeneous" definition when a latent construct is introduced or investigated. In fact, many researchers report different definitions of a single concept and finally elaborate their own view and their own definition. To say it more provocatively, for some latent constructs, there are nearly as many definitions as there are researchers working on them. Hagger (2014) spotlighted the problem of the many overlapping constructs in psychological research and claimed that more guides to constructs are needed, as Skinner (1996) presented for constructs addressing issues of control. Mentioning the term-mingling problem, Skinner says that "when the same term is used to refer to different constructs, reviewers may conclude that findings are inconsistent or even contradictory, when in fact it is definitions that are inconsistent and contradictory" (Skinner, 1996, p. 550). Later, Skinner explicates this problem with the term "secondary control" (Skinner, 1996). In psychological science, the existence of different terms with an implicitly overlapping meaning and the existence of a single term with different meanings both entail difficulties for empirical research.

Due to a dominant focus on statistical methodology, psychologists tend to concentrate more on the inner consistency and congruency of latent constructs than on the valid description of psychological phenomena, as Michell (2013) already argued. Dealing with latent constructs, epistemology seems to be reduced to a halfhearted demand for generalizability, the demand for objectivity and simultaneously the ignoring of the researcher's subjectivity and perspective, respectively. Theory, sometimes, seems to be reduced to considerations about the constructs that were measured in the study. This neglect

leads to the problem of overlapping constructs, concepts and approaches in psychological research and makes it redundant and uncontrollably inexact. If objectivity and generalizability really would work with latent constructs, there would be no problem with overlap, redundancy and, last but not least, the replication crisis. So, should we say goodbye to latent constructs, objectivity or generalizability? Can we overcome the replication crisis taking into account epistemological considerations? Maybe we can solve some parts of it.

The argument is that for a capable, process-oriented and updatable validation of latent constructs that protects us against redundant concepts of psychological phenomena, we have to replace generalization, induction and deduction with a more natural and efficient approach to learning: assimilation and accommodation. In other words, the idea is to consider always that meeting statistical objectivity and generalizability of a construct does not mean that a theory is really true (Meehl, 1992). Statistically perfectly verified constructs also should be handled with theoretical and phenomenological reflection and perspective-taking. As soon as latent constructs depend on personal perspectives, objectivity and generalizability a strict induction-deduction-logic is excluded. However, latent constructs always depend on a perspective: In the best case, they depend on the perspective of an approach or a theory, but they are also influenced by the strategy to calculate them, by a single scientist or a group of scientists. In fact, we need a more honest approach to our latent constructs that explicate our automatized perspective-taking. With perspective-taking, I do not mean to let personal, subjective or political aspects influence a certain construct. It is meant, as a first step, to identify the potential influences that create the understanding of a construct. Not to identify the influences does not mean that the influences do not take place.

I argue that latent constructs should be handled flexibly like mental representations as Piaget (1976) and Baldwin (1906) before him proposed and investigated. Assimilation in the original psychological sense means that an outside object is adapted to an already existing mental representation. Accommodation, in contrast, means that a mental representation is adapted (newly created or actualized) to an object. In other words, assimilation means that a mental representation changes (via perception or action) an object, whereas accommodation means that an object changes a mental representation. Transferred to latent constructs, you can say that assimilating mental representations and latent constructs (e.g., via expertise or a measuring procedure) adapts psychological phenomena to the mental representation or the questionnaire's concept. It is similar to Edwards and Bagozzi's (2000) distinction between reflective and formative measurements of latent constructs. The authors describe reflective (assimilative) measures as something where "constructs are usually viewed as causes of measures" (Edwards and Bagozzi, 2000, p. 155), and further "[i]n some situations, measures are viewed as causes of constructs" (ibid). The latter they call formative (accommodative) measurement, which occurs especially when we "know" that the construct is not one-dimensional, such as socioeconomic status. **Figure 1** shows the difference between the two approaches.

Assimilating here implies the ignoring of potential changes of phenomena, because latent constructs or mental representations cause the measurement. Accommodating latent constructs means that psychological phenomena are perceived less dependent from existing constructs or mental representations. Potential changes of psychological phenomena are not ignored, but they foster an actualization of the mental representation and, consequently, an actualization of the corresponding latent construct. It seems that this is one of the basics of good research practice that is applied anyhow.

OUR PROBLEM WITH INDUCTION, DEDUCTION AND GENERALIZABILITY AND ELLEN SKINNER'S WORK

Why could it be an advantage to apply assimilation and accommodation instead of induction and deduction? In philosophy of science, induction means to infer generalized principles from specific empirical observations. Deduction means to infer the validity of specific empirical observations from a generalized principle. In contrast, the distinction of assimilation versus accommodation describes a different, very basic and adaptive strategy to generate knowledge. Induction and deduction describe logic-based inferring procedures, coming from a highly sophisticated epistemological tradition. However, assimilation and accommodation are closer to our natural knowledge-generating functions, or as Caligiore et al. (2014) would say, the processes have an intuitive power and occur automatically. What is most important here, in the context of assimilation and accommodation, generalizability does not have this absolute understanding of generalizability. Knowledge can formally depend on perspective, time, or place.

Applying the scientific induction concept, we falsely assume the generalized validity of mental representations without the possibility of testing them empirically. This was Popper's main critique point of positivism and induction (Popper, 2005). However, even the deduction concept has a similar problem, because it assumes that a valid generalized mental representation already exists. This is the reason why Popper suggested to speak of tentative knowledge and not of secure scientific knowledge. Maybe the second problem is not so virulent for observable facts because exceptions are very obvious sooner or later. But it becomes difficult when we think, for example, about a mental representation of a personality trait which is non-observable but should be valid interculturally or over a time period of fifty or a hundred years. So, a proven theory fosters the assimilative mode. We apply "already existing" knowledge, also, for example, after a sophisticated inductive process. A too distinctive assimilative mode means that we tend to assimilate even if we should have received empirical hints to rethink – accommodate – our mental representations.

There is a further problem in developing latent constructs. I postulate that the development of latent constructs is based on very similar assimilative and accommodative processes such as children's development of mental representations – for example, how a child develops his or her mental representation of a cat.

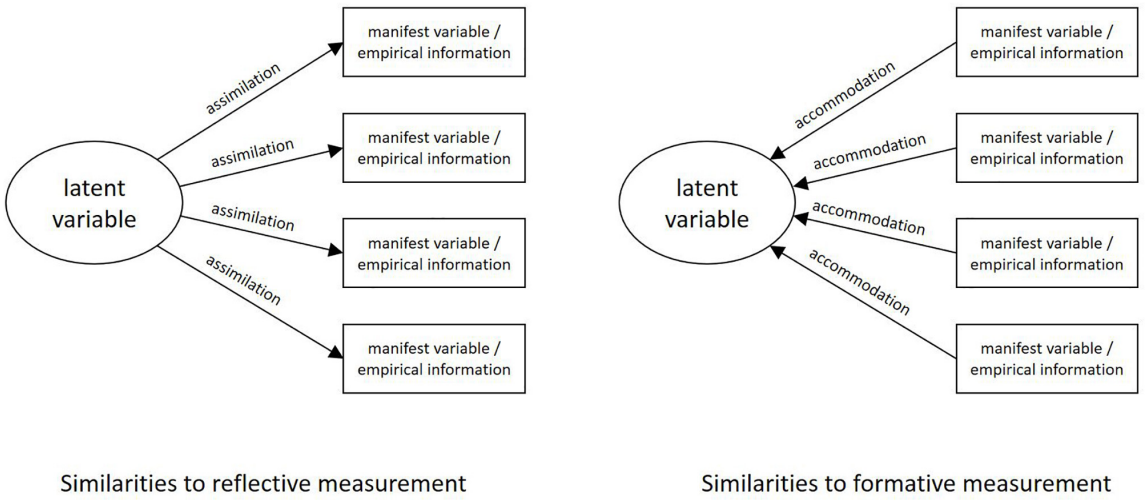

FIGURE 1 | Similarities to reflective and formative measurements according to Edwards and Bagozzi (2000).

He or she learns to say "cat" when he or she sees a cat. The child also says "cat" if he or she sees a dog or another animal that bears analogies to a cat, such as a marten. In this case, the child does not yet have a well-developed representation of a cat. In spite of this fact, the child tends to assimilate all objects which are more or less analogous to a cat. In order to fit the objects "dog" and "marten" to the scheme of a cat, micro-accommodative processes, as Piaget (1976) described them, are necessary to handle the discrepancies between the animals in an automatized process. Only when the child conducts non-automatic (and non-micro-) accommodative processes aimed at developing two new mental representations, one called "dog" and another called "marten," is he or she able to distinguish the three objects "cat," "dog," and "marten" correctly. Using three schemes instead of one also implies that the automatically running micro-accommodative processes which accompany the perception of one of the three animals are no longer as extensive.

Skinner et al. (2003) show us how we can reduce micro-accommodative processes in psychological research. Investigating the different meanings of the term "coping," the authors say: "[W]e focused on how these category systems were created. We considered about 100 schemes used during the past 20 years" (p. 218). The authors then make a lot of distinctions within the concept of coping. For example, they identified different functions of coping, topological distinctions as higher order categories of coping, effortful versus involuntary responses to stress, and so on. In sum, Skinner et al. (2003) reconsider the term "coping," suggesting new understandings of the association of the different definitions, hierarchical connections and theoretical implications. As mentioned above, Skinner (1996) provided a similar "guide to constructs" for the term "control": "The goal of this article is to collect control-related constructs and to organize them according to their definitions" (p. 550). In her article, Skinner differs between subjective and objective control, the experiences of control, motivations for control, agents, means and ends of control and means-ends, agent-means and agent-ends relations, respectively, and so forth. In fact, Skinner provides a theoretical integration of many independently developed constructs in order to enhance the validity of psychological research and to reduce replication problems due to definition fuzziness.

RECENT CONCEPTS OF ASSIMILATION AND ACCOMMODATION

Several approaches discuss assimilative and accommodative processes in different psychological contexts. The most basic one investigates them on a physiological information processing level. Fiedler (2001) and Fiedler et al. (2010) describe assimilation as a top-down process which is knowledge driven. In contrast, accommodation can be seen as a stimulus-driven bottom-up process.

The assimilative style in positive mood is by definition less contingent on large amounts of stimulus input than the accommodative style in negative mood. Conversely, assimilation includes the ability to enrich and elaborate a limited stimulus input through self-generated inferences, by going actively beyond the information given (Fiedler et al., 2010, p. 484).

These considerations go in line with further ideas, for example the so-called assimilation bias (Lord et al., 1979; Lord and Taylor, 2009). Lord and Taylor (2009) associate the assimilation bias with a tendency to over-generalize information that "allows people to develop assumptions and expectations even for specific objects that they have never encountered before" (p. 828). To some degree, the assimilation bias can be associated or even identified as an overlapping phenomenon with other biases, like the confirmation bias (Nickerson, 1998), which recently has been associated with the replication crisis (e.g., Lilienfeld, 2017). At the recent level of concretization, assimilation bias and confirmation bias reflect very similar phenomena: "As the term is used in this article and, I believe, generally by psychologists, confirmation

bias connotes a less explicit, less consciously one-sided case-building process. It refers usually to unwitting selectivity in the acquisition and use of evidence" (Nickerson, 1998, p. 175). Analogical, "[b]iased assimilation occurs when perceptions of new evidence are interpreted in such a way as to be assimilated into preexisting assumptions and expectations" (Lord and Taylor, 2009, p. 827). Similar to both of these biases, the Einstellung effect, first investigated by Luchins (1942), plays a role when once found problem solutions are preferred to faster or easier solutions. Bilalić et al. (2010) showed that this effect takes place when experts are using their expertise (see also Bilalić, 2017), and researchers and scientists are assumed to be experts in using and applying theories and concepts.

Proulx and Heine (2010) discussed these phenomena on the level of philosophy and psychological research. They suppose the meaning maintenance model as an integrative framework, which focus on forced assimilative processes when a meaning making system is violated by external stimuli and argue this effect in the context of threat-compensation literature. In this context, "Assimilation is a common response to meaning threats because it's fast and requires little in the way of cognitive resources" (Proulx and Heine, 2010, p. 894). Conversely, "accommodation is such a resource-heavy process, in the face of an anomaly people often do not have the wherewithal to begin to make any sense of what they've encountered" (ibid). For similar differences between assimilation and accommodation see Labouvie-Vief et al. (2010), where they focus on life-long-learning and the role of emotions:

Assimilation represents a low effort, automatic, and schematic processing mode, in which judgments are framed in a binary fashion of good or bad, right or wrong, and positive or negative. Regulation is oriented at dampening deviations from these binary evaluations. Accommodation, in contrast, involves a conscious and effortful unfolding, elaboration, and coordination of emotional schemas into complex knowledge structures (p. 87).

Looking at Piaget, how can it be that assimilation and accommodation are not always equilibrated but biased? Maybe because the two antagonists are not always balanced. Block (1982) was the first who proposed a different concept of equilibration that allows the idea of prolonged assimilative or accommodative forces, with appropriate consequences for our knowledge generation. In several later approaches, like in the assimilation bias and similar biases, assimilation and accommodation are not forced balanced (e.g., Hollon and Garber, 1988; Bosma and Kunnen, 2001; Fiedler, 2001).

BLOCK'S APPROACH OF EQUILIBRATION, THE POSSIBILITY TO OVER-ACCOMMODATE AND OVER-ASSIMILATE AND WHAT WE CAN LEARN FROM NEURO-ROBOTICS

Piaget (1976) described equilibration as something that is reached automatically due to the subject's adaptation to the world via assimilation and accommodation. Assimilation and accommodation, in their understanding, are balanced out and occur equally distributed. Coming from a personality-oriented perspective, Block (1982) suggested a different understanding of equilibration. Here, people differ in their way to approach equilibration with their environment: Whether a person tends to assimilate to reach equilibration with his or her environment, or he or she tends to accommodate to reach equilibration. Block associated people with prolonged assimilative efforts with the absence of the registration of discrepancies, with being too enthusiastic in the application of schemes, and with intolerance of ambiguity. In contrast, he characterizes people with prolonged accommodative efforts through their behavioral fluctuations, ever-changing perceptual-cognitive-action recognitions of possibilities, and intolerance of simplicity (Block, 1982, p. 292). Block's suggestion to perceive assimilation, accommodation and equilibration differently is a helpful foundation to explain the phenomenon of over-assimilation or over-accommodation, which have been investigated in clinical research.

In clinical research, for example, a different understanding of equilibration is part of trauma research. Littleton and Grills-Taquechel (2011, p. 421), for example, describe the sub-optimal strategy of over-accommodation when dealing with traumas, which means a "maladaptive or extreme schema change" (see also Hollon and Garber, 1988; Krawczyk et al., 2017). Taking the definition of assimilation and accommodation by Aguilar and Pérez y Pérez (2015), over-accommodation clearly should be considered seriously as a relevant source of the replication crisis. They define assimilation processes in their neurorobotic system as "search of schemas in memory representing similar situations to the one described in the current-context." (Aguilar and Pérez y Pérez, 2015, p. 31). Conversely, they see accommodation processes "as creation of new schemas and the modification of the existing ones as a result of dealing with unknown situations in the world" (Aguilar and Pérez y Pérez, 2015, p. 29).

Given that psychological phenomena are already described and investigated in literature, sometimes it would be better to read more before creating a new latent construct. Constructing new psychological constructs without a systematical scan of literature comes very close to a scientific over-accommodation, with many overlapping constructs and replication problems as a result. In contrast, if we transfer Block's approach of equilibration to a fixed, generalized latent construct, which is measured by a questionnaire or test, the measurement is clearly associated with an assimilative mode and fosters the ignorance of discrepancies. Even if we can identify the quantity of measurement error, we do not know the quality of measurement error or unexplained variance. In many models, the amount of unexplained variance is higher than the amount of explained variance. Even effect sizes are no reliable identifier of the amount of measurement error (Loken and Gelman, 2017). Here, an accommodative process is needed, not only on a statistical level but also on conceptual, theoretical and epistemological levels.

To conclude, the development of latent constructs in a positivistic tradition via induction and deduction implies the demand of their generalizability. However, empirically, this is a status that hardly can be reached. Even more, it leads to many uncontrollably over-assimilated or over-accommodated and therefore overlapping latent constructs. Latent constructs are particularly affected by this problem because they are (1) non-observable, (2) only weakly dependent

upon concrete behavior and therefore difficult to validate, (3) individually abstracted by us and therefore (4) more vulnerable to implicit subjectivity and assimilation and similar biases. I assume that the explication of assimilative and accommodative processes in empirical research methods helps to enhance the validity of latent constructs and to reduce the déjà-variable phenomenon as well as the poor replicability of our research; according to Block, "Assimilate if you can, accommodate if you must" (Block, 1982, p. 286). Following Block (1982), I summarized some causes and consequences of over-assimilation and over-accommodation when doing research with latent constructs (**Table 1**).

Speaking about these challenges seems to imply that many aspects of research practice need to be changed, but this is not the case. In fact, most field-tested research methods which are currently in use do a very good job: They are doubtless the most highly elaborated perspectives to refine and actualize constructs or theories. They allow the necessary professional distance to mental representations and therefore the potential to reduce over-assimilation or over-accommodation. However, research methods alone do not protect a researcher against over-assimilation and over-accommodation automatically. In the following, I set out some consequences and implications on the research method level, on the individual level and on the level of philosophy of science which could enhance the explication of well-balanced assimilative and accommodative research processes.

TABLE 1 | Causes and consequences of over-assimilation and over-accommodation.

	Over-assimilation	Over-accommodation
Development and application of latent constructs	Ignorance of discrepancies and (e.g., societal or cultural) changes of constructs	Ignorance of already existing constructs
Problems concerning the validity of latent constructs	Implicit (unidentified) overlaps of constructs because one construct describes different phenomena	"Invention" of "new" constructs which describe already known phenomena without additional information (déjà-variables)
Scientists' personal tendency	Intolerance of ambiguity	Intolerance of simplicity
Scientists' needs	Need to defend their "own" construct	Need to develop their "own" construct

CONSEQUENCES AT RESEARCH METHOD LEVEL

The explication of assimilation and accommodation in research processes is accompanied by perspective-taking. Research methods should be seen as perspectives which provide a view on constructs or psychological phenomena. Sometimes, there seems to be a kind of confusion between research method and psychological phenomenon, especially when a phenomenon can be made evident by only one research method. This confusion of research methods with the construct itself plays a role in the emergence of overlapping constructs. Even more, as discussed above, scientific over-accommodation takes place when a new construct is introduced without scanning existing literature and the new construct can be measured by one research method. Measuring a construct as assimilative process again ignores discrepancies between the new construct, already existing constructs and the phenomenon itself. Given a validity study with a correlation of $r = 0.7$, still 51% of the variance remains unexplained, without any idea what this 51% could be (see Loken and Gelman, 2017). Many constructs overlap because they are each "found" by one research method, one style of thinking or one view of different disciplines. The real problem comes when the many overlapping constructs are not compared with each other on a theoretical level. Here, an exact differentiation between the construct, the view of the construct and the phenomenon is needed. Uher (2018) provides a highly elaborated guide which should be considered when measuring psychological phenomena.

CONSEQUENCES AT THE LEVEL OF STUDENTS AND SCIENTISTS

One implication at the level of scientists is that they need well elaborated mental representations of theories and constructs to ensure a differentiation between theories, constructs and different views on a construct. The recent need to conduct reviews is a consequence of not having the same concepts in mind when talking about theories. Systematical reviews are one good solution to meet this growing problem in empirical research. For example, Morling and Evered (2006) brought some clarity into the research about secondary control. They reviewed 53 empirical articles which were published between 1985 and 2005, compared the definitions of secondary control which were used in the studies and proposed a definition of secondary control that should comprise all relevant aspects of the empirical work on the construct. Skinner (2007) could, based on Morling's and Evered's challenging but necessary work, provide even more theoretical clarity about the concept of secondary control. If Morling, Evered and Skinner had not done this work, many different perspectives on secondary control would still stand side by side and reinforce the problem of Hagger's déjà variable. However, this is only one construct which was only used in a manageable area of research. Thus, the strategy is to be as well informed as possible about theories and constructs in psychology and neighboring disciplines. This implies an intensive theory-based education for students, which ensures a well elaborated development of mental representations of theories. Furthermore, students need trainings to develop the competence to consider and clearly discuss constructs, their interconnections and their connections to theories as well as the competence to identify their own relation to psychological phenomena. In order to foster the connection between the students' mental representations and psychological phenomena, we must teach them the competence to distinguish between their own assimilative and accommodative modes. Additionally, well developed mental representations ensure a

higher quality of their application in research practice, but also in clinical and other practices.

One of the most important but perhaps underestimated consequences for scientists is the point that responsibility for the validity of a construct or theory cannot be delegated to empirical research methods. They can help a scientist to accommodate his or her cognitive schemes when they facilitate a better or more specific view of a construct or theory. We have to take the consequences that Meehl (1992) already articulated: "No statistical procedure should be treated as a mechanical truth generator" (p. 152). The validity of a construct still depends on a scientist's or a scientific community's conclusions (see also Edelsbrunner and Dablander, 2018). The validity is maximized when they minimize over-accommodation or over-assimilation and other biases. One example comes from the psychotherapeutic research practice. There is a more than 20-year-old debate about how to integrate knowledge from different therapeutic schools. In this context, Wolfe (2001, 2008) suggested to develop this integration on both assimilative and accommodative integration, not only assimilative integration.

CONSEQUENCES FOR PHILOSOPHY OF SCIENCE

All in all, there are not as many inconsistencies between positivistic thinking, critical rationalism, and systemic and constructivist epistemologies as often discussed. Rather, I assume, they describe different phases of a knowledge-generating process. Positivism describes the determination of a cognitive scheme on the basis of verification. It also justifies the assimilative-oriented process of induction or description of the world on the base of logico-mathematical principles, according to Block (1982, p. 286): "assimilate if you can." Critical rationalism, besides preferring deduction, draws attention to the point that sometimes it is better to be in an accommodative mode in order to realize that a mental representation (theory, construct) does not necessarily fit the outside world: "accommodate if you must" (ibid). Systemic approaches show us the relevance of perspective and that our own perspective and our own behavior are part of and influencing those systems; or that some principles perhaps do not follow a logico-mathematical order. Ignoring it does not mean that it does not take place. Constructivism reminds us that, in fact, we cannot slip out of ego- and anthropocentrism, a point that should also not be ignored anymore. Thus, there is no need to take sides with any one of those ideas because all of them describe important aspects of the process of knowledge-generation. However, maybe we should take into account more the psychology of science (e.g., Gholson et al., 1989; Feist, 2006).

CONCLUSION

In this paper, I want to show why I assume that replication problems in empirical psychology are not only due to statistical artifacts and methodological errors and why they are also caused by a lack of epistemological and theoretical clarity. I refer to validity problems that can arise from the use of latent constructs with a simultaneous positivist scientific orientation. As long as latent constructs are evaluated predominantly with regard to their calculation quality and too little with regard to their theoretical embeddedness in a coherent theory system, there is the potential that once calculated constructs are hardly falsified. Phenomena like Martin Hagger's "déjà-variables" point to this problem.

I argue to meet this problem by taking a different epistemological perspective and propose why the use of assimilation and accommodation could be quite appropriate. Assimilation and accommodation are specific adaptation processes that describe and explain the development of cognitive and behavioral processes. They should therefore also be suitable for formalizing the further development of latent constructs. One important point is that there are already several examples from psychological research, such as Block's personality approach or clinical work, where the concept of assimilation and accommodation helps to understand phenomena where over-assimilation and over-accommodation hinder the achievement of equilibrium and thus validity. The explicit integration of assimilation and accommodation in epistemology changes the perspective on theory development at the level of research methods, researchers and philosophy of science.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

I would like to thank the Research Council, the Publication Fund of the Alpen-Adria-Universität Klagenfurt and the Institute for Teaching and School Development for their financial support.

ACKNOWLEDGMENTS

Thanks to all friends and colleagues who discussed the idea of this contribution with me and helped me to bring it further. Many thanks go to the reviewers, who contributed significantly to the quality of the article with their feedback.

REFERENCES

Aguilar, W., and Pérez y Pérez, R. (2015). Dev E-R: a computational model of early cognitive development as a creative process. *Cogn. Syst. Res.* 33, 17–41. doi: 10.1016/j.cogsys.2014.09.002

Baldwin, J. M. (1906). *Mental Development in The Child and The Race: Methods and Processes*. New York, NY: The Macmillan Company.

Bickhard, M. H., and Campbell, R. L. (2005). New ideas in psychology. *New Ideas Psychol.* 23, 1–4. doi: 10.1016/j.newideapsych.2005.09.002

Bilalić, M. (2017). *The Neuroscience of Expertise*. Cambridge: Cambridge University Press.

Bilalić, M., McLeod, P., and Gobet, F. (2010). The mechanism of the einstellung (set) effect. *Curr. Dir. Psychol. Sci.* 19, 111–115. doi: 10.1177/0963721410363571

Block, J. (1982). Assimilation, accommodation, and the dynamics of personality development. *Child Dev.* 53, 281–295. doi: 10.2307/1128971

Bosma, H. A., and Kunnen, E. S. (2001). Determinants and mechanisms in ego identity development: a review and synthesis. *Dev. Rev.* 21, 39–66. doi: 10.1006/drev.2000.0514

Caligiore, D., Tommasino, P., Sperati, V., and Baldassarre, G. (2014). Modular and hierarchical brain organization to understand assimilation, accommodation and their relation to autism in reaching tasks: a developmental robotics hypothesis. *Adapt. Behav.* 22, 304–329. doi: 10.1177/1059712314539710

Edelsbrunner, P. A., and Dablander, F. (2018). The psychometric modeling of scientific reasoning: a review and recommendations for future avenues. *Educ. Psychol. Rev.* 31, 1–34. doi: 10.1007/s10648-018-9455-5

Edwards, J. R., and Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychol. Methods* 5, 155–174. doi: 10.1037//1082-989x.5.2.155

Feist, G. J. (2006). *The Psychology of Science and the Origins of the Scientific Mind*. New Haven: Yale University Press.

Fiedler, K. (2001). "Affective states trigger processes of assimilation and accommodation," in *Theories of Mood and Cognition: A User's Guidebook*, eds L. L. Martin and G. L. Clore (Mahwah, NJ: Lawrence Erlbaum Associates).

Fiedler, K. (2017). What constitutes strong psychological science? the (neglected) role of diagnosticity and a priori theorizing. *Perspect. Psychol. Sci.* 12, 46–61. doi: 10.1177/1745691616654458

Fiedler, K., Renn, S.-Y., and Kareev, Y. (2010). Mood and judgments based on sequential sampling. *J. Behav. Decis. Mak.* 23, 483–495. doi: 10.1002/bdm.669

Gholson, B., William, R. S. Jr., Neimeyer, R. A., and Houts, A. C. (eds) (1989). *Psychology of Science: Contributions to Metascience*. Cambridge: Cambridge University Press.

Hagger, M. S. (2014). Avoiding the "déjà-variable" phenomenon: social psychology needs more guides to constructs. *Front. Psychol.* 5:52. doi: 10.3389/fpsyg.2014.00052

Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Front. Psychol.* 4:246. doi: 10.3389/fpsyg.2013.00246

Hollon, S. D., and Garber, J. (1988). "Cognitive therapy," in *Social Cognition and Clinical Psychology: A Synthesis*, eds L. Y. Abramson and L. Y. Abramson (New York, NY: Guilford Press).

Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., and Pollard, B. (2014). Discriminant content validity: a quantitative methodology for assessing content of theory-based measures, with illustrative applications. *Br. J. Health Psychol.* 19, 240–257. doi: 10.1111/bjhp.12095

Krantz, D. H., and Wallsten, T. S. (2019). Comment on trendler's (2019) "conjoint measurement undone". *Theory Psychol.* 29, 129–137. doi: 10.1177/0959354318815767

Krawczyk, M. C., Fernández, R. S., Pedreira, M. E., and Boccia, M. M. (2017). Toward a better understanding on the role of prediction error on memory processes: from bench to clinic. *Neurobiol. Learn. Mem.* 142(Pt. A), 13–20. doi: 10.1016/j.nlm.2016.12.011

Labouvie-Vief, G., Grühn, D., and Studer, J. (2010). "Dynamic integration of emotion and cognition: equilibrium regulation in development and aging," in *The Handbook of Life-Span Development Social and Emotional Development*, eds R. M. Lerner, M. E. Lamb, and A. M. Freund (Hoboken, NJ: Wiley).

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: righting the ship. *Perspect. Psychol. Sci.* 12, 660–664. doi: 10.1177/1745691616687745

Littleton, H. L., and Grills-Taquechel, A. (2011). Evaluation of an information-processing model following sexual assault. *Psychol. Trauma* 3, 421–429. doi: 10.1037/a0021381

Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science* 355, 584–585. doi: 10.1126/science.aal3618

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. Personal. Soc. Psychol.* 37, 2098–2109. doi: 10.1037//0022-3514.37.11.2098

Lord, C. G., and Taylor, C. A. (2009). Biased assimilation: effects of assumptions and expectations on the interpretation of new evidence. *Soc. Personal. Psychol. Compass* 3, 827–841. doi: 10.1111/j.1751-9004.2009.00203.x

Luchins, A. S. (1942). Mechanization in problem solving: the effect of einstellung. *Psychol. Monogr.* 5:6.

Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *J. Pers.* 60, 117–174. doi: 10.1111/j.1467-6494.1992.tb00269.x

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas Psychol.* 31, 13–21. doi: 10.1016/j.newideapsych.2011.02.004

Michell, J. (2019). Conjoint measurement underdone: comment on günter trendler (2019). *Theory Psychol.* 29, 138–143. doi: 10.1177/0959354318814962

Morling, B., and Evered, S. (2006). Secondary control reviewed and defined. *Psychol. Bull.* 132, 269–296. doi: 10.1037/0033-2909.132.2.269

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

Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen. Konzepte der Humanwissenschaften [The equilibration of cognitive structures. Concepts of human sciences]*. Stuttgart: Ernst Klett.

Popper, K. R. (2005). *Logik Der Forschung. [The Logic of Scientific Discovery]*. Tübingen: Mohr Siebeck.

Proulx, T., and Heine, S. J. (2010). The frog in kierkegaard's beer: finding meaning in the threat-compensation literature. *Soc. Personal. Psychol. Compass* 4, 889–905. doi: 10.1111/j.1751-9004.2010.00304.x

Rossiter, J. R. (2008). Content validity of measures of abstract constructs in management and organizational research. *Br. J. Manag.* 19, 380–388. doi: 10.1111/j.1467-8551.2008.00587.x

Sackman, H. (1974). *Delphi Assessment: Expert Opinion, Forecasting, and Group Process. A Report Prepared for United States Air Force Project*. Santa Monica, CA: RAND Corporation.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Stud. Hist. Philos. Sci.* 42, 509–524. doi: 10.1016/j.shpsa.2011.07.001

Skinner, E. A. (1996). A guide to constructs of control. *J. Pers. Soc. Psychol.* 71, 549–570. doi: 10.1037/0022-3514.61.3.549

Skinner, E. A. (2007). Secondary control critiqued: is it secondary? is it control? comment on morling and evered (2006). *Psychol. Bull.* 133, 911–916. doi: 10.1037/0033-2909.133.6.911

Skinner, E. A., Edge, K., Altman, J., and Sherwood, H. (2003). Searching for the structure of coping: a review and critique of category systems for classifying ways of coping. *Psychol. Bull.* 129, 216–269. doi: 10.1037/0033-2909.129.2.216

Smith, L. D. (1986). *Behaviorism and Logical Positivism: A Reassessment of the Alliance*. Stanford: Stanford University Press.

Swami, V., Barron, D., Weis, L., Voracek, M., Stieger, S., Furnham, A., et al. (2017). An examination of the factorial and convergent validity of four measures of conspiracist ideation, with recommendations for researchers. *PLoS One* 12:e0172617. doi: 10.1371/journal.pone.0172617

Trendler, G. (2019a). Conjoint measurement undone. *Theory Psychol.* 29, 100–128. doi: 10.1177/0959354318788729

Trendler, G. (2019b). Measurability, systematic error, and the replication crisis: a reply to Michell (2019) and Krantz and Wallsten (2019). *Theory Psychol.* 29, 144–151. doi: 10.1177/0959354318824414

Uher, J. (2018). Quantitative data from rating scales: an epistemological and methodological enquiry. *Front. Psychol.* 9:585. doi: 10.3389/fpsyg.2018.02599

Wolfe, B. E. (2001). A message to assimilative integrationists: it's time to become accommodative integrationists: a commentary. *J. Psychother. Integr.* 11, 123–131.

Wolfe, B. E. (2008). Toward a unified conceptual framework of psychotherapy. *J. Psychother. Integr.* 18, 292–300. doi: 10.1037/1053-0479.18.3.292

Conflict of Interest Statement: 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.

frontiers in Psychology

METHODS
published: 20 November 2018
doi: 10.3389/fpsyg.2018.02245

On the Development of a Computer-Based Tool for Formative Student Assessment: Epistemological, Methodological, and Practical Issues

Martin J. Tomasik[1,2], Stéphanie Berger[1,3] and Urs Moser[1]*

[1] *Institute for Educational Evaluation, University of Zurich, Zurich, Switzerland,* [2] *Department of Developmental and Educational Psychology, University of Witten/Herdecke, Witten, Germany,* [3] *Research Centre for Examinations and Certification, University of Twente, Enschede, Netherlands*

Formative assessments in schools have the potential to improve students' learning outcomes and self-regulation skills; they make learning visible and provide evidence-based guidelines for setting up and pursuing individual learning goals. With the recent introduction of the computer-based formative assessment systems for the educational contexts, there is much hope that such systems will provide teachers and students with valuable information to guide the learning process without taking much time from teaching and learning to spend on generating, evaluating and interpreting assessments. In this paper, we combine the theoretical and applied perspectives by addressing (a) the epistemological aspects of the formative assessment, with an emphasis on data collection, model building, and interpretation; (b) the methodological challenges of providing feedback in the context of instruction in the classroom; and (c) practical requirements for and related challenges of setting up and delivering the assessment system to a large number of students. In the epistemological section, we develop and explicate the interpretive argument of formative assessment and discuss the challenges of obtaining data with high validity. From the methodological perspective, we argue that computer-based formative assessment systems are generally superior to the traditional methods of providing feedback in the classroom, as they better allow supporting inferences of the interpretive argument. In the section on practical requirements, we first introduce an existing computer-based formative assessment system, as a case in point, for discussing related practical challenges. Topics covered in this section comprise the specifications of assessment content, the calibration and maintenance of the item bank, challenges concerning teachers' and students' assessment literacy, as well as ethical and data-protection requirements. We conclude with an outlook on possible future directions for computer-based formative assessment systems and the field in general.

Keywords: abilities, adaptive testing, competencies, computer-based assessment, education, epistemology, formative assessment

OPEN ACCESS

Edited by:
Barbara Hanfstingl, Alpen-Adria-Universität Klagenfurt, Austria

Reviewed by:
Maria Tulis, University of Salzburg, Austria
Peter Adriaan Edelsbrunner, ETH Zürich, Switzerland

****Correspondence:***
Martin J. Tomasik
martin.tomasik@ibe.uzh.ch

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *17 May 2018*
Accepted: *29 October 2018*
Published: *20 November 2018*

Citation:
Tomasik MJ, Berger S and Moser U (2018) On the Development of a Computer-Based Tool for Formative Student Assessment: Epistemological, Methodological, and Practical Issues. Front. Psychol. 9:2245. doi: 10.3389/fpsyg.2018.02245

INTRODUCTION

Educational research has experienced a remarkable progress in the past 20 years. This is reflected in the creation of new institutional structures, a massive expansion in funding, and an increase in the public interest and recognition (Köller, 2014). These successful developments can partly be attributed to methodological shifts toward quantitative method. This method has allowed measuring the outputs and outcomes of entire educational systems—a process often referred to as 'educational monitoring' (Scheerens et al., 2003). Although educational evaluation results were initially prepared for the use of teachers, principals, and school administrators, it soon became clear that the formative assessment could have a substantial impact on students' learning and performance (e.g., Hattie and Timperley, 2007). Formative assessments provide feedback on students' learning progress, encouraging a systematic use of data. The expansion of information technologies has given schools the opportunity to develop an efficient and user-friendly culture of formative assessment for teachers who may not be experts in rigorous test analyses (Brown, 2013), allowing them to focus on teaching. Experts have even argued that an automated formative assessment is the most effective use of digital technologies in the classroom, compared with the other cases of computer-assisted instruction, such as drill-and-practice applications (e.g., Moser, 2016). Technological assessment systems have several advantages for everyday use that make learning visible to students and teachers. Computer-assisted formative assessment helps teachers to focus their attention on instruction and grade data objectively with minimal time and effort expended in data collection and analysis. In addition to assessment for learning and diagnostic testing (see van der Kleij et al., 2015), this data-based decision making in education (see Schildkamp et al., 2013) is considered one of the three most important approaches to the formative assessment. Decisions based on objective data can also increase teaching effectiveness and minimize bias (see Lai and Schildkamp, 2013; Schildkamp and Ehren, 2013).

This paper discusses the core aspects of data-based formative assessment technology. It comprises five parts. In the first part, we provide an overview of the theoretical foundations of the formative assessment, along with some empirical evidence on its benefits for learning. In the second part, we focus on the epistemological aspects of the formative assessment systems and develop an interpretive argument about scoring, generalization, extrapolation, and implication in the formative assessment. In the third part, we examine the methodological challenges of such systems and argue that computer-based technology can provide more effective solutions than the traditional methods. In the fourth part, we introduce a sample case of a computer-based formative assessment system and discuss some fundamental practical requirements related to its development and operation. We conclude with a discussion of possible further developments in computer-based formative assessment and examine some ideas on how it could evolve.

AN OVERVIEW OF THE FORMATIVE ASSESSMENT BENEFITS

From a theoretical perspective, formative assessments pursue several purposes. They can 'provide feedback and correctives at each stage of the teaching-learning process' (Bloom, 1969, p. 48). They can help us to 'adapt the teaching to the student needs' (Black and William, 1998, p. 140). They can also help us to 'adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes' (McManus, 2008, p. 3). New Zealand's Ministry of Education (1994) defines formative assessment as 'a range of formal and informal procedures [. . .] undertaken by teachers in the classroom as an integral part of the normal teaching and learning process in order to modify and enhance learning and understanding' (p. 48). Given these definitions, most educators and researchers would agree that the formative assessment should not be limited to single tests, but rather needs to be considered an ongoing process (Popham, 2008; Shepard, 2008). This process consists of a cyclical feedback loop in which (a) the students' current proficiency level is assessed, (b) the assessment-based learning goals are defined, (c) the students' learning progress is monitored by further assessments, and (d) the learning goals and environments are adjusted based on the assessment outcomes (van der Kleij et al., 2015; see also Brookhart, 2003, p. 7).

The conceptual strength of the formative assessment is to make learning visible (see Havnes et al., 2012). It can also aid in using students' strengths and weaknesses to frame appropriate learning goals, monitor their progress toward the goals, and to inform the extent of their success or failure in achieving the goals. In essence, the process concerns three fundamental questions: 'Where am I going?,' 'How am I getting there?,' and 'Where to go next?' (Hattie and Timperley, 2007). The answers can be found in the objective data from the assessments. The process can either directly support learning and self-regulation or be used for diagnostics and data-driven decision making (van der Kleij et al., 2015). It also suits the notions of individualization and differentiated instruction (see Levy, 2008). In fact, the formative assessment can be a prerequisite for individualization and differentiation, as it specifies a student's current standing and her/his extent of progress. The formative assessment is also highly compatible with the current trend toward educational measurements. On the conceptual level, summative and formative assessments share an orientation toward educational outcomes and both can support teaching and learning (Bennett, 2011). On the methodological level, measurement theories that are used include: item-response theory (IRT; see de Ayala, 2009), measurement concepts such as adaptive testing (see Wainer, 2000), and measurement tools such as computer-assisted assessment (see Conole and Warburton, 2005).

There is ample empirical evidence that feedback can substantially benefit learning and self-regulation (e.g., Cawelti and Protheroe, 2001; Campbell and Levin, 2009; Lai et al., 2009; Carlson et al., 2011). Feedback is even considered 'the most powerful single moderator that enhances achievement' (Hattie, 1999). The first studies dating back to the 1950s (e.g., Ammons,

1956), and the more recent meta-analyses, suggest remarkable effect sizes. One of the most comprehensive meta-analyses to date was published by Kluger and DeNisi (1996). They collected 607 effect sizes from 131 studies on the effectiveness of feedback interventions on learning and extracted an average d = 0.41, which corresponds to a small-to-medium effect size (Cohen, 1992).

In the late 1990s, Hattie (1999) published a synthesis of over 500 meta-analyses involving over 400,000 effect sizes from 180,000 studies on various influences on student achievement. The average effect of schooling was d = 0.40 per school year, which can be considered a benchmark against which the effects of feedback can be judged. In sum, 12 previous meta-analyses evaluating 196 studies and almost 7,000 effect sizes were considered. The average effect size was d = 0.79, almost twice the average effect of schooling and large (Cohen, 1992). However, there was considerable variability in the effect sizes, depending on the type of feedback provided. For example, the effect sizes of praise (d = 0.14), punishment (d = 0.20), and reward (d = 0.31) were low, whereas receiving feedback related to a specific task (d = 0.95) and providing cues on how to solve a problem more effectively (d = 1.10) provided the highest effect sizes (see also Hattie and Timperley, 2007).

Empirical evidence concerning effects on self-regulation is less conclusive, although it is widely believed that appropriate feedback should enable the students to monitor the attainments of their learning goals more autonomously (Bernhardt, 2003; Earl and Katz, 2006; Love, 2008; Herman and Winter, 2011). Butler and Winne (1995) suggest that 'research on feedback and research on self-regulated learning should be tightly coupled' (p. 245). Overall, studies show positive effects on motivational, metacognitive, and strategy-use aspects of self-regulation with substantial effect sizes (e.g., d > 1.00 in Dignath et al., 2008), with the feedback type playing a decisive role (e.g., Nicol and Macfarlane-Dick, 2006).

However, not all studies, reviews, and meta-analyses show positive effects of the formative assessment (or feedback, more specifically) on achievement and self-regulation. Rather, the variability in effect sizes is very large, which points to the possibility of substantial moderation by variables that are still poorly understood. Bennett (2011) argues that the studies usually used in meta-analyses might be 'too disparate to be summarized meaningfully' (p. 11). Indeed, 38% of the effects of all studies compiled by Kluger and DeNisi (1996) were *negative*, suggesting higher performance in the control group (see Shute, 2008; Dunn and Mulvenon, 2009; Bennett, 2011).

FORMATIVE ASSESSMENT SYSTEMS: EPISTEMOLOGICAL ASPECTS

As opposed to more traditional approaches to validity and validation (e.g., Cronbach and Meehl, 1955), the current authoritative approach is that of 'validity as an argument' (see **Figure 1**), in which it is not the validity of a test *per se*, but rather the validity of the meaning of test scores and their implications

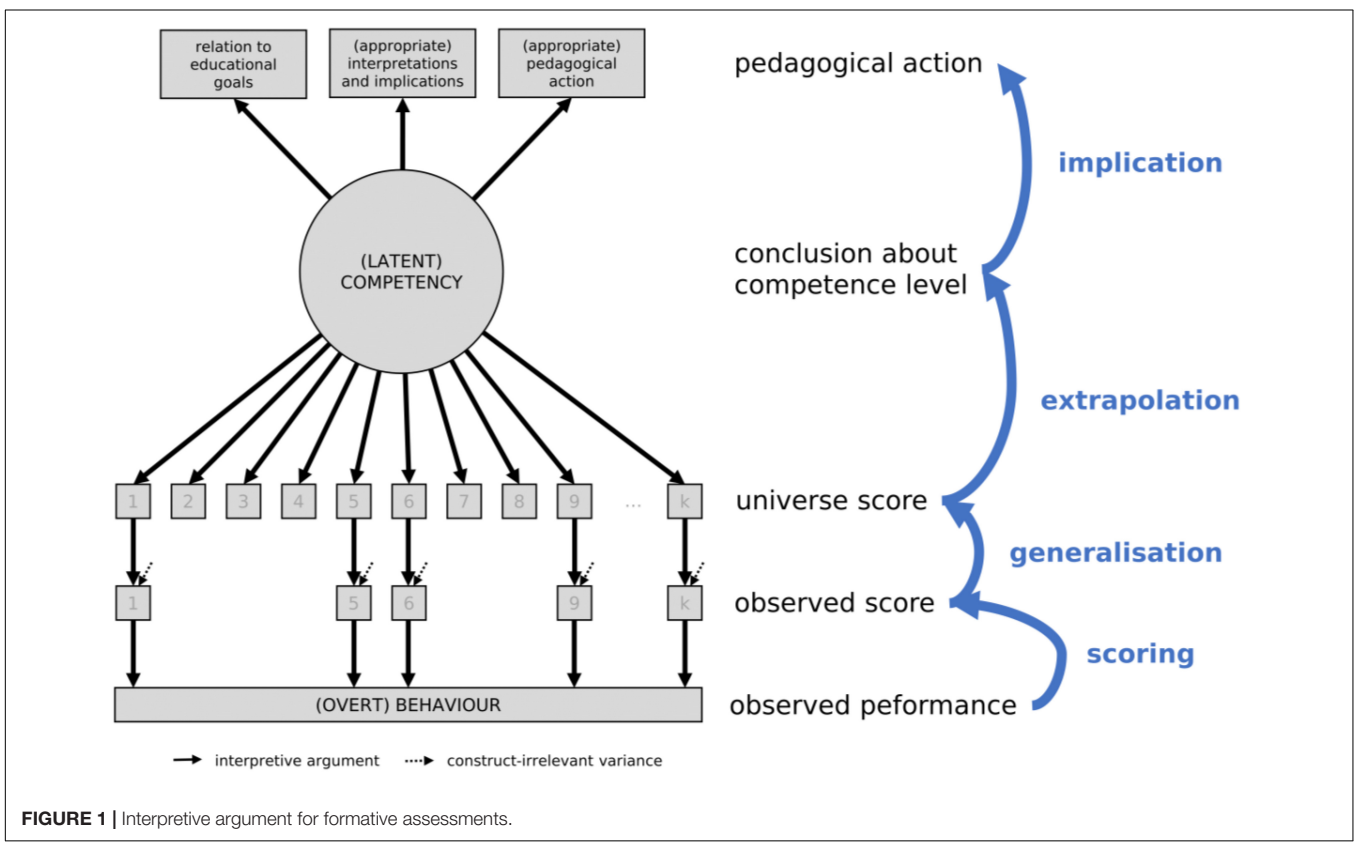

FIGURE 1 | Interpretive argument for formative assessments.

for action that are evaluated (Kane, 2006, 2013; see also Messick, 1989, 1995). Many alternative concepts of validity exist (e.g., Borsboom et al., 2004), and there is an ongoing substantial debate about the relation between validity and truth (e.g., Borsboom et al., 2004; Kane, 2013; Newton and Baird, 2016; for an overview, see Cizek, 2012; Newton and Shaw, 2014). An in-depth discussion of this debate is beyond the scope of this paper; however, we would like to concentrate on the concept of 'validity as an argument,' as a widely accepted notion.

At the core of the concept of validity as an argument is the *interpretive argument*. This can be considered a scientific mini-theory that merits assessment/testing developers' attention. Interpretive argument should be distinguished from the *validity argument*. This latter argument provides an evaluation of the interpretive argument in terms of clarity, consistency, plausibility, and empirical examination. In other words, the interpretive argument is defeasible by failure in the validity argument, and, as with the other scientific theories, such failure can result in the reformulation, restriction or even rejection of the interpretive argument. In the following section, we will develop an interpretive argument for formative assessment by addressing the issues of scoring, generalization, extrapolation, and implication.

Scoring Argument

The interpretive argument for the formative assessment comprises four inferences, namely scoring, generalization, extrapolation, and implication (see **Table 1**). The *scoring inference* is concerned with obtaining valid observed scores from an observed performance. In technical terms, this refers to translating a response, such as a selected multiple choice category or an essay, into a score by means of a scoring key or rating scheme. The scoring rule to do so needs to be free of bias and applied accurately and consistently across all subjects and measurement occasions. This is usually facilitated when standardized tests are used; however, issues may arise when humans are involved in judging performance. In general, the scoring inference for the formative assessment is not much different from those applied to trait interpretations, summative assessments, or placement systems (see Kane, 2006, for more details).

Generalization Inference

The observed scores are based on a sample of observations and a subset of what Kane (2006) labeled the 'universe of generalization.' For example, if the sample of observations contains a set of four items, covering bridging to ten in summation, then the universe of generalization would be all the possible items covering this topic (e.g., '7 + 5 =,' '7 + 6 =,' etc.). Hence, the *generalization inference* is concerned with obtaining a valid universal score from the observed score, an issue that is also discussed in more traditional approaches to validity (e.g., Linn et al., 1991). There are three main issues related to generalization. First, the sample of observations needs to be representative of the universe of generalization and, especially in cases of adaptive or tailored testing, parameter invariance must hold (e.g., Rupp and Zumbo, 2006). Ensuring representativeness is best achieved when the universe of generalization is known and a random sample of items is drawn from this universe. However, in reality, the universe of generalization is only known, if at all, for narrowly circumscribed topics (e.g., bridging to ten) and is not well-defined for more complex ones (e.g., writing

TABLE 1 | Interpretive argument for formative assessments.

Scoring (from observed performance to observed score)	
S1	Scoring rule is appropriate.
S2	Scoring rule is applied accurately and consistently.
S3	Scoring is free of bias.
S4	Data fit the scaling model employed.
Generalization (from observed score to universe score)	
G1	The sample of observations is representative of the universe of generalization.
G2	In case of adaptive or tailored testing, parameter invariance holds.
G3	The sample of observations is large enough to control random error.
Extrapolation (from universe score to conclusion about competence level)	
E1	The universe of generalization is representative of the competency.
E2	There are no construct-irrelevant sources of variability that would seriously bias the interpretation of the competence level.
E3	For extrapolations onto higher aggregate levels (e.g., classes), clear participation rules have been followed.
E4	For extrapolations over time (in terms of learning progress), the learning function must be known.
Implication (from conclusion about competence level to pedagogical action)	
I1	The competence level can be related to an educational goal ('Where am I going?').
I2	The implications associated with the competence level are appropriate, and the semantic interpretation of the
I3	competence level is plausible, legitimate, and accurate ('How am I getting there?').
I4	Whichever pedagogical action is most appropriate depends on the achieved competence level ('Where to go next?').
I5	The decision rules for pedagogical action are appropriate.
	The pedagogical actions taken are effective in improving learning.

Table partly adapted from Kane (2006).

an argumentative essay). In combination with the context-specific nature of learning and thinking (see Greeno, 1989), this makes the selection of items a challenging endeavor. Second, if a measurement model is employed for scaling, which is almost always the case in computer-based assessments, data need to sufficiently fit the model and its assumptions. This is tested routinely in models based on IRT (e.g., Orlando and Thissen, 2000). However, differential models between relevant subgroups (e.g., boys and girls) are not considered extensively. In some cases, this might represent a threat to test fairness and jeopardize the interpretation of inter-individual and group differences. Biased parameter estimates might also arise when unidimensional models are set up but the measured characteristic is not unidimensional (see Ackerman, 1989). This can be the case when the underlying scales are supposed to cover many or even all the school grades. Finally, the number of observations also needs to be large enough to control for random error. This is particularly difficult to achieve in the formative assessment, in which testing time is usually constrained, and only a limited number of items can be presented at any one time.

Extrapolation Inference

The next step in the interpretive argument is the *extrapolation inference* from the universe score to a conclusion about the students' competence levels. For formative assessments, there are three requirements for a valid extrapolation. First, it is necessary that the universe of generalization is representative of the competency or the competency domain to be measured. For example, the universe of all the possible items covering bridging to ten must be representative of the competency to add numbers in the range up to 20. Again, for narrowly described competency domains, this is sometimes self-evident, whereas for more complex domains, this requires more justification. The issue of representativeness in extrapolation has been discussed elsewhere in more detail. For example, Messick (1995) points to the utility of 'task analysis, curriculum analysis, and especially domain theory' (p. 745) for defining the structure and content of the competency or its domain. Labeling it as 'construct domain,' Messick highlights the importance of covering all parts of the construct domain, which can be achieved through ecological sampling, already suggested by Brunswik (1956). This 'content coverage' (Linn et al., 1991) or 'scope' (Frederiksen and Collins, 1989) seems to be particularly relevant in the context of the formative assessment, as gaps in coverage might result in students and teachers underemphasizing those parts of the content that were not considered for assessment. Second, it is equally important that what is captured are only the sources of variability relevant to the targeted competency or its domain, which otherwise would seriously jeopardize the interpretation of the competence level. Construct-irrelevant variability tends to contaminate the task by making it either 'too easy' or 'too difficult' for some students but not for the others. For instance, some items that test the ability of bridging to ten might be color-coded and thus be unduly difficult for color-blind children. Other items might use gender-specific illustrations, thereby eliciting more response from one gender group than from the other. There are many sources of construct-irrelevant variance (see Messick, 1995; Kane, 2006), and they become particularly relevant in the low-stakes testing context of the formative assessment. This is because students tend to reduce test-taking effort in low-stakes assessments, presumably because doing well on the test will bring them limited attainment, intrinsic or utility value for them (Wise and DeMars, 2005). Consistent with the expectancy-value model of achievement motivation (e.g., Wigfield and Eccles, 2000), most research clearly shows that test score validity falls with decreasing test-taking effort, which in turn means that the construct-irrelevant variance and/or error variance more strongly determine the test score. To the best of our knowledge, no extant research has systematically investigated these aspects or has estimated their effects on the validity of formative assessments. We can only speculate that factors such as self-regulation abilities, attention span, conscientiousness at the individual level, classroom climate, availability of computers in the classroom, or teacher support at the system level might be more optimal for some students but not for the others, hence the possibility of construct-irrelevant variance when test-taking effort decreases. Third, teachers or administrative authorities might want to use the formative assessment data to extrapolate a single student's scores of competence to those of groups of students or the entire student population. This can be problematic in the absence of clear participation rules, causing self-selection bias to affect the estimated competence level. At the very least, information is needed about the (non-)participants in formative assessments, and about how these two groups differ in terms of ability and learning progress. To ensure a valid extrapolation from the universe score to conclusions about the competence level, we require broad representativeness, low construct-irrelevant variability, and participation transparency.

Implication Inference

The final step in the interpretive argument is the *implication inference* from the competence level to pedagogical (or administrative) action. Assessment experts consider this step the most important yet the least controllable. It is essential to note that some definitions of the formative assessment *always* encompass a strong functional element. For example, New Zealand's Ministry of Education (1994) defines the formative assessment as 'a range of formal and informal procedures [. . .] undertaken by teachers in the classroom as an integral part of the normal teaching and learning process *in order to modify and enhance learning and understanding*' (p. 48, emphases added). Brown and Cowie (2001) define it as 'the process used by teachers and students to recognize and respond to student learning *in order to enhance that learning* during learning' (p. 510, emphases added); they further argue that 'assessment can be considered formative only if it results in *action by the teacher and students* to enhance student learning' (p. 539, emphases added). Finally, for Black and William (1998), 'assessment becomes 'formative' when the evidence is *actually used to adapt the teaching*' (p. 140, emphases added). Hence, if the purpose of the formative assessment is to enhance learning, then validity is about whether this purpose is achieved or not (see Stobart, 2012). This notion of consequential validity was first proposed by Messick (1989, 1995) and further developed by Kane (2006, 2013), both of whom

focused strongly on the uses (and misuses) of test scores in theorizing about validity and validation.

The implication inference in formative assessments comprises five aspects (see **Table 1**). The first three facets refer to the central functions of the formative assessment, as identified by Hattie and Timperley (2007), whereas the latter two address issues of effectiveness, and whether they instigate the appropriate pedagogical action. Because the purpose of the formative assessment is to 'reduce discrepancies between current understandings/performance and a desired goal' (Hattie and Timperley, 2007, p. 87), an effective formative assessment needs to meet three criteria. The first criterion is 'Where am I going?,' and a student's response to it will define the learning goal. To provide valid accounts of this question, the measured competence level must be related to the learning goal. Both need to be represented on the same dimension and quantified in the same currency. For example, the information that a student 'knows all the letters of the alphabet' would be less relevant for defining the learning goal than 'having a good command of arithmetic in the range up to 20'; however, the information that a student can 'bridge to ten' certainly would. This step might be trivial for the well-defined and specific learning goals, but can present a challenge for the complex and multifaceted learning goals, such as 'writing an argumentative essay' or 'being able to apply trigonometric functions to everyday problems.' In the context of writing a good argumentative essay, for example, one may enquire about the requisite skills and knowledge. The answer would be that one needs to know about text structure, data collection, thesis development, presentation of well-supported (counter-) claims, and presentation of conclusions against the backdrop of logical, rhetorical, and statistical rules and conventions. Assessing and giving feedback about all these aspects is far from being trivial. The second question is 'How am I going?,' and embraces the feedback aspect of the formative assessment. This requires a semantic interpretation of the attained competence level that is plausible, legitimate, and accurate. The implications offered based on this level must be appropriate, too. Due to a lack of training in test theory, it is unlikely that all the students and teachers will arrive at a common interpretation when confronted with a single score in a competency domain. However, even if the students and teachers are formally trained in test interpretation, most decisions made in classrooms and other real-world settings usually tend to be based on holistic qualitative assessments (e.g., Moss, 2003; Stiggins, 2005; Kane, 2006). It is not difficult to imagine that information from isolated formative assessment that is not compatible with the prevailing holistic appraisal will likely be discounted or disregarded at all. This bias poses a most serious threat to the validity of the formative assessment. A similar argument can be made for the third question: 'Where to go next?' However, in this case, other aspects seem more relevant. The ultimate function of the formative assessment is to adjust teaching to the students' competence level. This presupposes that we know which pedagogical action is most appropriate and practicable, given a student's achieved competence level. Gaining this information, however, may not be very easy, and if the differences in students' competence levels are ignored, they may lead to decisions that recommend inappropriate pedagogical actions, seriously damaging the validity of the formative assessment (see Akers et al., 2016). This brings us to the final requirement, which is particularly important for implication inference because it links pedagogical action with learning outcome. This requirement is that pedagogical action informed by data from formative assessment results in significantly better learning outcomes as compared to pedagogical action without these data. This is a very strict validity criterion, especially in settings where instruction quality is high anyway.

METHODOLOGICAL CHALLENGES AND SOLUTIONS

Obtaining information from formative assessment based on computer technology in combination with complex measurement models has some demanding methodological challenges as compared to obtaining information from other sources of information such as ordinary classroom tests or observations. However, when these challenges are met, the epistemological value of such formative assessment and its utility for making truly 'reflective classroom-assessment decisions' (see McMillan, 2003) is much higher. In the following, we want to examine these challenges by focusing on the inferences of scoring, generalization, and extrapolation. We will contrast such assessment with the more traditional ones and point out how they can help increase the validity.

Scoring Inference

Objective, appropriate, accurate, consistent and bias-free scoring is the basis for valid formative assessment. To fulfill these requirements, we need clear, complete, and accurate scoring rules, and we need to ensure that these rules are implemented consistently. Ideally, we also could collect empirical evidence on the quality of the scoring rules and their implementation. To evaluate students' performance in the classroom, teachers usually develop and apply their own, often-intuitive scoring rules (e.g., McMillan, 2003). The objectivity of such scoring largely depends on the teacher. An experienced teacher, for example, is more likely to consider all the appropriate scoring options while developing the scoring rules, compared to a less experienced teacher. Time pressures or preconceptions about students' abilities might also influence the quality of a teacher's use of the scoring rules (e.g., Foster and Ysseldyke, 1976; McKown and Weinstein, 2008). In contrast, computer-based assessment systems offer the advantage of objective scoring through predefined scoring rules; they score the data automatically and independently of the subjects and measurement occasions. The systematic collection of data also allows the empirical validation of the predefined scoring rules via item analyses. This procedure gradually improves scoring quality by identifying wrong or flawed scoring rules (e.g., Linn, 2006). In principle, teachers could also perform such empirical validations of their own scoring rules. However, collecting relevant data and the ability to draw generalizations based on these data may not be very

feasible for teachers, given their limited time and lack of expert knowledge. A computer-based assessment system allows data collected from entire populations of students to be used to validate the scoring.

Generalization Inference

The generalization of an assessment score is especially challenging in the context of the formative assessment. Formative assessments are extremely diverse, as they are used to assess the strengths and weaknesses of each individual student repeatedly in all sorts of educational and instructional settings (e.g., Black and William, 1998; Brookhart, 2003; McMillan, 2003; McManus, 2008). From a methodological perspective, how can we ensure that these diverse assessments result in general and comparable scores with a small margin of random errors? First, a general reference or scale is required to allow us to compare the outcomes of different assessments or assessment versions. Second, item selection needs to be guided to ensure representative sampling from all eligible items. Third, item selection should focus on students' ability levels to minimize the random error of the assessment score.

For traditional classroom assessments, teachers usually use grades as a general metric for comparing the outcomes of different assessments. However, no universal, objective rules exist for generalizing assessment scores to grades. Often, grading is influenced by the performance of the class as a whole in the sense of a norm-referenced score interpretation. Also, teachers are completely free to adjust their grading based on their subjective interpretation of the assessment content and context. For example, they can give higher grades for an average score if they think an assessment is particularly difficult, or that students had too little time to answer all the questions properly. Thus, the comparability of grades from different assessments largely depends on the class context and how teachers interpret students' performance in terms of grades. It also depends on the teacher's ability and experience to assemble representative items for reliable assessments to serve as sufficient information for generalizing a score or an observation (e.g., McMillan, 2003; Smith, 2003). Depending on the target competency, the range of possible assessment items is very broad and difficult to grasp, so it might be very time-consuming for teachers to prepare targeted and reliable assessments for every single student.

Computer-based assessment systems, as noted above, can support teachers in objectifying the generalizability of outcomes from the formative assessment. Computer-based assessment systems particularly allow implementing complex measurement models, such as those based on IRT (e.g., de Ayala, 2009), which can serve as warrants for generalizing the outcomes of different item sets or assessment versions (Kane, 2006). Generally speaking, IRT models imply probabilistic predictions about responses by linking person characteristics and item characteristics by some probability function. The family of Rasch models is a special case of IRT models (see Mellenbergh, 1994) and most often used in the context of educational measurement, so that we will only focus on them in the following. These models state a distinctive, monotonically increasing relation between the probability of answering an item correctly and its difficulty alongside student's ability. One important feature of Rasch models is the underlying assumption of parameter invariance (e.g., Rupp and Zumbo, 2006). Parameter invariance holds that the assessment outcome (i.e., the ability estimate) is independent of (a) the specific items from the range of generalization chosen, (b) the order in which they are presented, and (c) the respondent. Hence, under the (falsifiable) condition that all eligible items refer to the same underlying unidimensional construct, it is possible to provide scores on a common unidimensional scale (e.g., Kolen and Brennan, 2014, p. 191), even though students work on different tailored item sets. These generalized scores are not only comparable among students but also within students across different time points. The transformation from students' observed scores on an item level to a generalized ability score is determined by the underlying model, and is completely standardized across all assessment occasions (Wainer and Mislevy, 2000). Rasch models also serve as a tool for gathering empirical evidence to validate the model assumptions, which are crucial for generalizing the scores of various assessments, including the relation between person characteristic and item characteristic, unidimensionality, and parameter invariance.

Computer-based assessment systems, in tandem with complex measurement models, can also support teachers and students in selecting representative item samples for assessments. Ideally, such systems would include calibrated item banks. These are large pools of independent assessment items with an associated item metadata, such as item difficulty or affiliation to a content domain of the curriculum. Based on this metadata, teachers and students can identify suitable items for creating their own customized assessments, and then decide what they intend to assess and when and how to collect feedback relating to their specific questions (McMillan, 2003; Hattie and Brown, 2008). This autonomy is very important to encourage the parties to accept formative assessments (e.g., Hattie and Brown, 2008). At the same time, test blueprints and item-selection algorithms can help teachers and students select representative items and create reliable assessments. Calibrated item banks can also serve as a basis for administering computer adaptive tests (CAT; Wainer, 2000; van der Linden and Glas, 2010)—an automated form of tailored testing. With CAT, adaptive algorithms use preliminary ability estimates during test taking to select the most suitable items for each individual. These targeted items not only have the advantage of not overly demotivating students by being too easy or too difficult, but they are the most informative with regard to students' ability. The resulting increased measurement efficiency is especially relevant if the target population is heterogeneous and/or testing time is limited. Thus, CAT contributes to the generalizability of assessment results by minimizing the random error (e.g., Lord, 1980; Wainer, 2000; van der Linden and Glas, 2010). In conclusion, we argue that calibrated item banks, based on item response theory, are an ideal tool for addressing reliability and validity. They are particularly useful because they are well adjusted to the context of formative classroom assessments (Brookhart, 2003; McMillan, 2003; Moss, 2003; Smith, 2003), and give teachers sufficient leeway for making decisions that best suit their circumstances. Also, a large item bank is a practical prerequisite that allows setting up formative

assessments as a genuine process, as opposed to being a one-off event or a short-term initiative.

It is vital that data fit the proposed model and its assumptions sufficiently well. This can pose a particular challenge when students' competency levels need to be linked across the grade levels. It is imperative then to look beyond single item fit statistics and focus instead on global fit statistics. To do so, several methods have been suggested, including those specifically developed for item response theory models (see Suárez-Falcón and Glas, 2003) as well as those borrowed from structural equation modeling (see McDonald and Mok, 1995). Models with different dimensionality assumptions should be compared against each other. Principal component analyses should also be applied to the residuals from a one-dimensional model to enable the examination of the degree to which multidimensionality is present (see Chou and Wang, 2010).[1] In practice, it is time-consuming and costly to find adequate items that span abilities across grade levels and still meet the assumption of unidimensionality.

Extrapolation Inference

A score that meets the requirements of scoring and generalization is meaningful only if it can be extrapolated to other competencies. From a methodological perspective, extrapolation requires three techniques. First, it requires supporting and evaluating the representative item selection. Second, it requires detecting and preventing construct-irrelevant variability. Third, it requires collecting information about assessment participation and context. Some traditional classroom assessments might fulfill these requirements while others may not. Teachers normally develop assessments and provide feedback that are closely related to their teaching (Brookhart, 2003). Thus, teaching and assessments focus on the same target competencies. However, teachers do not always have the opportunity to empirically validate whether the assessment is representative of the target competencies or whether it is unaffected by construct-irrelevant sources of variability. This might be a minor problem if the target competency is specific and well-articulated but less so for broader constructs. Regarding the extrapolation of assessment results to higher aggregated levels, teachers are usually in an ideal position to comment on the underlying student sample of an assessment group mean. For example, some students might be excluded from an assessment due to individual learning goals or simply miss the assessment because of illness. Thus, only teachers can place the aggregated values into context and interpret their true meaning. Similarly, teachers are in a favorable position to track and evaluate their students' learning progress longitudinally, whereas it might be difficult for external parties to rely on a snapshot of available data to distinguish 'good' from 'limited' progress.

Within an item-banking system, item-selection algorithms and test blueprints can help teachers to create representative assessments by guiding the item-selection process and reverting to content specifications. Such a system can facilitate tracking previous assessments and visualizing possible gaps in content coverage in all the previous assessments. An underlying unidimensional IRT model, such as the Rasch model, can further enhance the extrapolation from the ability scores to the related competence levels, brought about by the common scales for abilities and difficulties. This relation serves as a basis for criterion-referenced score interpretation (Moser, 2009). In particular, a mastered item content or example item with a high probability can be used to map and describe a specific ability level (Beaton and Allen, 1992; Huynh, 1998). IRT models can also be used to test the construct-irrelevant sources of variability—also known as differential item functioning. This test involves correcting deviations of the probability for solving an item correctly in different groups (e.g., boys and girls), conditional on the specific ability levels in these groups (Camilli and Shepard, 1994), and providing a clear indicator of bias in an item (Lord, 1980). Construct-irrelevant variability can be minimized by targeted assessments or CAT. The administration of the easy items to low-ability students and the more difficult ones to high-ability students might prevent students from getting discouraged or bored by items that do not fit their ability levels (Asseburg and Frey, 2013). Computer-based assessment systems collect and visualize information about the participating student samples, which allow teachers and other stakeholders to use aggregated scores to draw informed conclusions about the competence levels of groups or classes. Such systems have other advantages, too. For example, they enable the longitudinal comparability of assessment results, and provide graphical illustrations of students' learning progress; they also present empirical data about the anticipated learning progress, giving teachers, students, and external parties a broader perspective of students' progress.

PRACTICAL REQUIREMENTS OF FORMATIVE ASSESSMENT SYSTEMS

Due to its nature and scope, the formative assessment requires a huge item bank. The costs of such a bank, however, can only be reasonable if it is delivered to a large number of students. Hence, the objective of making learning visible in day-to-day school life almost inevitably turns into a large-scale project that poses practical challenges. In this section, we will introduce a developing computer-based formative assessment system to serve a population of more than 100,000 students in some German-speaking parts of Switzerland. We will highlight five practical challenges, namely item development, item calibration, item banking, assessment literacy, and ethical considerations.

A Computer-Based Formative Assessment System

We have developed a computer-based formative assessment system[2] to provide students and teachers with an item bank in four school subjects: German (the school's medium of instruction), English and French (the two foreign languages taught), and mathematics. A distinctive feature of this system is its capability to cover topics and competencies from the third grade in the primary school until the third grade in the secondary school, spanning 7 years of compulsory schooling. The item

[1] We are grateful to the reviewer for drawing our attention to this issue.

[2] https://www.mindsteps.ch/

bank is based on a competency-based approach to learning (see Sampson and Fytros, 2008) that emphasizes learning progress and learning outcomes during the learning process. All items used are embedded in the curriculum (see Shepard, 2006, 2008; Shavelson, 2008). Currently, the item bank contains between 4,000 and 12,000 items per school subject; up to 15,000 items per school subject have been planned for the final stage of the project.

Our assessment system has two thematically identical types of item bank: (a) the practice item bank, and (b) the testing item bank. The *practice item bank* is openly available to all the students and teachers for training and teaching purposes. Students can autonomously use this item bank to create and answer an item set from a topic domain they choose or are instructed to choose. This can virtually be done from any place that has an Internet access. Students receive detailed feedback showing which items they answered correctly, and how well they have mastered the topic in question. This item bank is also open to teachers for instruction purposes without any restrictions.

The *testing item bank*, on the other hand, can be used to evaluate students' ability and learning progress and to identify their strengths and weaknesses in a given content domain. Teachers can select items according to the desired competency domains, single competencies, or curricular topics; they can also create tests that can be taken by students on computers at school. There are three 'use cases' for this item bank with three different types of feedback. First, teachers may want to use a general *competency domain*, such as reading comprehension or algebra, to assess their students' ability or learning progress. Second, teachers can test their students on a *single competency*, such as comprehension of simple discontinuous texts or summation in the number range of a million. Finally, teachers can administer tests on *topic-specific knowledge* to assess students' level of mastery. Such topics usually are very narrowly defined and often refer to the content of single instructional units. As opposed to the practice item bank, the testing item bank results are kept confidential in all three use cases, and students are not supposed to receive any help when trying the items. These restrictions are necessary because test results are used to automatically calibrate the item bank in terms of item-difficulty parameters.

Our formative assessment system provides performance feedback at the aggregate level of students and classes. This system can be used to promote a formative approach to instruction to support both students and teachers in setting up learning goals and monitoring their attainments (see Maier, 2015; van der Kleij et al., 2015). It has several features. First, both item banks are available throughout the school year (including break times) and hence allow for continuous monitoring of students' ability levels and their development over time. Second, the system's mathematical model is based on the Rasch model (e.g., Rasch, 1960), the most basic item response theory model, to determine and compare students' ability levels on a metric scale from grade three onward, providing long-term, diagnostic learning trajectories. The Rasch model also facilitates the implementation of adaptive testing algorithms in the assessment system (see Wainer, 2000; van der Linden and Glas, 2010) as well as a fine-tuning calibration of the item difficulty parameters on a running system (see Verschoor and Berger, 2015). Finally, because all the items were developed using the formal competency-based curriculum, our formative assessment system is capable of providing criterion-referenced test scores. Thus, the feedback contains not only abstract test scores, but also tangible examples of the students' competence levels that should help them and their teachers formulate meaningful and appropriate learning goals for each subject.

Valid Content Specifications for Item Development

The core of an item bank for the formative assessment contains thousands, or even tens of thousands, of assessment items. Although teachers usually focus on a specific content area, substantial effort has been expended in developing items to offer students and teachers a wide range of choices. Clear content specifications are crucial for any assessment system to make valid inferences from assessment results (Webb, 2006). However, curricula or content standards, which serve as a theoretical basis for test-content specifications, often lack empirical validation (Fleischer et al., 2013). An assessment system's empirical data contribute to the validation of the theoretical framework and the quality of the assessment items. At the same time, the theory-based content specification allows validating the decisions taken during item calibration, e.g., the selection of an IRT model or a specific linking procedure. The challenge, however, is that neither the theoretical framework nor the empirical data are completely bias-free; both sources are important for verifying each other to establish a valid scale for representing students' genuine abilities.

We used the formal competency-based curriculum as a content framework for item development. The curriculum contains detailed descriptions of students' competence levels, including statements about the development of each level. To put this theoretical framework into practice, we collaborated closely with content experts to develop the items for our item bank. We trained the content experts in test theory and familiarized them with our psychometric and technical guidelines (e.g., item types, number of distractors, styling). These guidelines are an important addition to the content specifications to ensure consistency within the item bank, that the items fulfill the assumptions of the underlying measurement model (e.g., measurement invariance or unidimensionality), and that they meet the system's technical requirements (e.g., available item formats or automated scoring). More than 25,000 items are currently available in our formative assessment system. Considerable effort is needed to validate the match between the theoretical content specification of the items (i.e., their affiliation with specific competence levels in the curriculum) and the empirical, item-response-theory-based item-difficulty estimates. This validation process allows us to detect problematic items, provide feedback to our item developers, and verify our psychometric strategies.

Item Calibration

A general scale is a prerequisite for a flexible item bank. This scale allows representing item parameters independently of a single test or predefined test versions. A vertical scale is required to measure a student's ability longitudinally (i.e., over several school

years), and provide feedback on a long-term learning progress (Tong and Kolen, 2007; Carlson, 2011; Kolen and Brennan, 2014). Unlike a horizontal scale, a vertical one combines item sets of varying average difficulty. Only a vertical scale can provide a panoramic view (7 years in our model) of a student's ability range. A vertical scale is also a precondition for comparing 'students' growth in terms of criterion-referenced magnitude,' 'out of level testing' by means of CAT, setting 'proficiency cut points coherently during standard setting,' and 'evaluating [the alignment of] standards, curriculum and instructions, and assessment [. . .] across grades' (Dadey and Briggs, 2012, p. 8). As far as IRT is concerned, various calibration and linking strategies have been introduced to establish a vertical scale (see Kolen and Brennan, 2014, for a general overview). The challenge here is to identify a calibration design and strategy that corresponds to the size of the available calibration sample, and is compatible with the properties of the measured construct and definitions of growth (i.e., domain vs. grade-to-grade definition of growth) (Kolen and Brennan, 2014).

The calibration of potentially tens of thousands of items in a computer-based item bank is a highly resource-intensive process. To establish vertical scales, we developed a common-item, non-equivalent group design (Kolen and Brennan, 2014). This strategy helped us to calibrate a few hundred anchor items, representative of target competencies and target grades. The calibration design, in more specific terms, consists of a combination of grade-specific and linking items. Grade-specific items are administered to one specific grade cohort only, whereas linking items are shared between two adjacent grade cohorts (Berger et al., 2015). This way, we managed to lay the foundation for establishing a link over different target grades and relating the items to one underlying vertical measurement scale. We then exported the response data to calibrate the anchor items; we did so using the Rasch model (Rasch, 1960) by means of marginal maximum likelihood (MML) estimation procedures. The calibrated items will subsequently serve as anchors for locating additional, uncalibrated items on the scale by means of online calibration (Verschoor and Berger, 2015). For new items with no or very few observations, an Elo update scheme (Elo, 1978) was used to determine the preliminary difficulty estimates of the items. The online-calibration algorithm, in its next move, will automatically switch to a joint maximum likelihood (JML) estimation process (Birnbaum, 1968). Thanks to online calibration, we can start the system after a brief offline calibration phase, which involves extending the item pool and improving the parameter estimates systematically, while students and teachers engage with the system.

Item Bank Development and Maintenance

The development and maintenance of the item bank, i.e., the 'organized collection of items' (Vale, 2006, p. 268), also pose some challenges. Computerized item banking is crucial for inventorying thousands of items, locating relevant items, tracking item usage, and developing an item's state or life cycle. In an item bank, the item content is stored in a respective metadata on the item properties, e.g., a unique item identifier, content classification, scoring key, or the name of the item's author. Additional item properties are based on the empirical data, such as IRT parameters or item exposure. Items can be classified in the item bank by their development state (e.g., new, calibrated, retired) and their relation (i.e., social order) to other items in the item bank (e.g., friend items, which must always appear together or enemy items, which must not be used in the same test; see Vale, 2006). All this information supports item-bank users in item selection and scoring; it is especially relevant when the system itself is responsible for automated item selection and scoring in CAT. However, CAT does not solely rely on an organized collection of items with relevant item properties, such as IRT parameters and content classifications. CAT can provide reliable and efficient ability estimates only if the item bank consists of a sufficient number of items relating to the target competencies and if item overexposure is prevented (Veldkamp and van der Linden, 2010; Thompson and Weiss, 2011). An item-banking system can help psychometricians to use simulation studies to evaluate the fit of the available items ahead of item administration.

In our formative assessment system, we use also the item bank for helping teachers and students to identify the relevant items for constructing their own formative assessments. For this purpose, teachers and students have access to selected item properties within the item bank. In particular, they can filter the contents the item bank in two ways. They can use content categories, namely the curriculum competence levels and related topics, or filter items in relation to the vertical scale, which represents the difficulty of the items on the same scale based on the reported scores. Thus, the outcomes of previous assessments can guide targeted item selection. Additional item properties are automatically used by the system to support teachers and students in constructing sensible assessments. For example, the system informs users about friend items (Vale, 2006), such as listening-comprehension items that are related to the same audio text. The identification of friend items helps teachers and students to create more authentic assessments; this way the students can answer multiple items related to the same support material, rather than switching the context after each item. This is especially relevant in competency domains such as reading and listening comprehension, in which processing the support material during test taking (i.e., reading a text passage or listening to an audio file) can be rather time-consuming.

Technological and Organizational Challenges

Setting up a large-scale computer-based assessment system can inevitably pose several technological and organizational challenges. There are challenges that are purely technological or specific to the design of the human–machine interface. The technology must be capable of perfectly supporting a wide variety of systems, devices, and browsers at school, at home and on the road. Considering the fact that there lacks a central instance for keeping operating systems up to date, in practice, there are a large number of versions and update stages that require supporting. For pragmatic reasons, this limits the prospects of deploying

new versions of the assessment software that would need to be extensively tested on all the various systems. As a compromise between user friendliness and practicality, our assessment system is only fully compatible with the latest two versions of the most popular internet browsers (i.e., Chrome, Firefox, Internet Explorer/Edge, and Safari). The infrastructure must also be capable of supporting several thousands of concurrent users during morning access peaks in the school. This is especially challenging for computer adaptive testing, not least because a continuous real-time communication with the item bank is required to select the appropriate items based on the students' previous responses. To manage the load during peak periods, we implemented multiple instances of the assessment delivery module of our assessment system, which allow us to distribute the load. From a design point of view, the development of an intuitive user interface is crucial, mainly because small deviations from the optimum will immediately result in a surge of customer support requests. Design also needs to take into account the broad age range of users and their scope of digital expertise.

Practical challenges also arise in relation to populating and maintaining the item bank, the large scale of which augments the demands for accuracy and the impact of errors. With thousands of items in each domain, we needed to set up comprehensive, standardized guidelines for designing items across different subjects, content domains, and different school grades or age groups. Quality assurance in a huge item pool is also challenging and labor-intensive: typing errors and errors in the scoring key need to be detected and eliminated, psychometric properties of the items should be constantly monitored, conspicuous items ought to be flagged and double-checked, and content specification needs to be consistently checked to ensure that items are assigned to the most suitable content category within a growing item pool. The maintenance of the item bank also requires a constant investment of time and effort. The item development outside the system needs to be synchronized with the active item pool, and updates of the items need to be carefully integrated into the system. To do so, it is necessary to keep track of the item versions and to decide whether or not updates need to be applied to the item parameters. Subsequently, eliminated items need to be replaced with new ones and matched to the content domains based on the difficulty level.

The quality assurance requires that all data be exported on a regular basis for an offline quality control. This quality control comprises the analysis of item discrimination parameters, a distractor analysis, an investigation of the item fit, and an analysis of differential item functioning between different school grades and types. From a practical point of view, we need to ensure that the data export does not interfere with system performance; that is why it usually takes place outside the usual working hours. We also need to ensure that the export meets all the standards of privacy and data protection. In the future, most of the quality assurance will be implemented automatically within the system to limit the need of data export. This, however, requires even more testing and supervision until the online quality assurance runs flawlessly. We decided to invest this testing and supervision effort and hope that it will pay off in the long run.

A final challenge that deserves a mention, although in passing, concerns designing reporting materials that support a valid interpretation of the results by students of all grades and at all stages of cognitive development. Although there are guidelines and even studies that have investigated design principles for assessment reports, few recommendations exist for age diverse populations. We have needed to adapt our materials several times and are now planning to run randomized controlled trials to investigate which type of report is best understood by whom.

Challenges Concerning Stakeholders' Assessment Literacy

Consequential validity (Messick, 1989, 1995; Kane, 2006, 2013), as the core aspect of the implication inference, strongly requires that all feedback be appropriately understood and interpreted within an inevitable margin of error. In the extant literature, this issue is referred to as 'assessment literacy,' and is defined as the 'understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions' (Popham, 2011, p. 267). Popham emphasizes three important aspects in this definition. First, 'understanding [. . .] concepts and procedures' does not necessarily imply that assessment users are able to develop and run reliable and valid assessments by themselves; equally, they may not know how to calculate ability estimates, standard errors, or reliability coefficients. However, users are expected to recognize the concepts and procedures, and know what they mean to arrive at valid interpretations of them. The focus of the second aspect is on 'fundamental' concepts and procedures, which encompass knowledge that is just about enough and necessary in the respective applied context. Hence, users are not expected to understand the different ways of calculating the different reliability coefficients. However, they should, for instance, understand why a reliability of $\rho = 0.50$ is by far not enough for the interpretation of individual test scores. Popham (2009) has proposed 13 'must-understand topics' for teachers and administrators. One example is the understanding that the function of educational assessment is 'the collection of evidence from which inferences can be made about students' knowledge, skills, and affect' (p. 8). Third, the understanding inherent in the concept of assessment literacy is limited to concepts and procedures that are 'deemed likely to influence educational decisions.' Assessment literacy, as defined above, does not imply that users understand all aspects of assessment but only those that are relevant to everyday decisions. Each of these three points is highly compatible with the concept of consequential validity advanced by Messick (1989, 1995) and Kane (2006, 2013).

There are three more aspects of assessment literacy that have received relatively limited attention. The first aspect is in line with the modern notion of competencies (see Klieme et al., 2008). It refers to the non-cognitive facets of assessment literacy, such as attitudes toward measurement, beliefs about one's own efficacy to make useful decisions based on assessment results, or motivational factors associated with their use. These non-cognitive facets interact with the cognitive ones. A basic understanding of the fundamental assessment concepts and

procedures can cultivate high self-efficacy beliefs and positive attitudes toward educational measurement. In turn, these positive beliefs and attitudes are expected to facilitate the understanding itself. Indeed, there is some evidence that holistic assessment literacy programs that look to assessment literacy as an integral part of professional development are more effective than programs that focus on technical and methodological aspects only (e.g., Koh, 2011). Such programs are probably key to using assessments appropriately. If teachers are extensively supported in conducting, analysing, and interpreting their assessments and learn to relate the assessments to the taught content, chances are good that they will accept formative assessment as a valuable tool in their work, start using it on a regular basis, and develop a sense of self-efficacy when using it.

Second, assessment literacy requires a positive assessment culture in which the process of the formative assessment follows certain requirements, such as the application of intra-individual standards of reference. Black and William (1998) also stress the importance of interaction and dialog in instruction to promote opportunities for students to express their understanding and for teachers to evaluate it. The Assessment Reform Group (1999, p. 7) argues that assessment is more likely to promote learning if it (a) is embedded in a view of teaching and learning of which it is an essential part, (b) involves sharing learning points with students, (c) aims to help students learn and recognize the standards they aim to achieve, (d) involves students in self-assessment, (e) provides feedback that informs students of subsequent action points, (f) is underpinned by confidence that every student can succeed, and (g) if it involves both teachers and students reviewing and reflecting on assessment data. Collectively, these points emphasize a positive and collaborative assessment culture that is a fundamental part of instruction (points a, f, and g), in which students and teachers are not only actively involved but also empowered to draw their own conclusions about their learning processes (points b, c, d, and e).

The third aspect concerns stakeholders' involvement, mainly students and teachers, but also administrators, test developers, and researchers with varying educational backgrounds, interests, and motivations. Teachers need to be assessment-literate to understand the scientific approach to educational measurement and the benefits of the use of formative assessment. Their assessment literacy should at least comprise the key elements of the assessment process, sometimes portrayed as the assessment triangle, comprising 'a model of student cognition and learning in the domain, a set of beliefs about the kinds of observations that will provide evidence of students' competence levels, and an interpretation process for making sense of the evidence' (Pellegrino et al., 2001, p. 44). Although there is evidence that teachers' assessment literacy is linked with notable benefits in students' learning (e.g., Wilson et al., 2001), studies suggest that currently teachers' competence levels in this regard are mediocre at best (Mertler, 2004; DeLuca and Klinger, 2010; Popham, 2011). Similar findings have been reported about teachers' self-described levels of assessment self-efficacy and literacy (e.g., Volante and Fazio, 2007). This is hardly surprising, considering the limited role of assessment literacy in teacher-education programs (e.g., DeLuca and Bellara, 2013). In an extensive review of measurement textbooks, Shepard (2006) found limited guidance 'about how teachers were to make sense of assessment data so as to redesign instruction' (p. 625). Teachers' lack of assessment literacy is likely to pose a serious and hardly controllable threat to validity in formative assessments, despite the existence of several initiatives and interventions to promote teachers' assessment literacy (e.g., Wang et al., 2008; Xu and Brown, 2016).

Students need to be assessment-literate as well to incorporate feedback in their learning processes adequately and get valid answers to Hattie's fundamental questions: where to go, how to get there, and where to go next (Hattie and Timperley, 2007). Equally important are their metacognitive strategies and self-regulation skills, which can be promoted by a competent utilization of formative assessment (Nicol, 2009; Sadler, 2009). Despite the growing interest in and application of testing and formative assessment in schools, there is a paucity of research dealing with this aspect of assessment literacy. However, one can assume that young and/or underachieving students might become overstrained by the demands of complex assessments. Francis (2008), for example, argues that even first-year university students tend to overrate their understanding of the assessment process. Programs that aim to promote assessment literacy in students exist (e.g., Smith et al., 2011), but they are usually targeted at adolescents or young-adult students, and to the best of our knowledge, no program exists for younger children.

Considerations on Ethics and Privacy

The potential benefits of this technology need to be evaluated against the potential ethical concerns that may arise from its usage. The first concern regarding computer-based formative assessments relates to *trust* (e.g., Lee and Nass, 2010). This is particularly crucial when students and teachers make consequential and potentially long-term decisions based on (necessarily) imperfect results. We partially have addressed this issue when discussing the necessity of assessment literacy for understanding and interpreting assessments, but the concern is broader. Computer algorithms might fail and produce flawed outcomes for longer periods of time before being detected. Students and teachers might overestimate the reliability and validity of the results that are neatly presented and appear to be backed scientifically. This may cause disappointments, especially if these expectations are unduly high.

The second ethical concern is the risk of discrimination (see Datta et al., 2015). It is widely recognized that learning algorithms are prone to biases (Caliskan et al., 2017) so that extreme care needs to be put into the selection of algorithms and the interpretation of their results to ensure that these biases are not projected (and possibly exaggerated) by the feedback provided. The nature of this problem is fundamentally different from the *correctness* of results noted above. Here, while results may be considered correct, they may slightly differ for different subjects, hence the discrimination. On the same note, one might also be concerned about the fairness of enhancement (e.g., Savulescu, 2006). If students with greater aptitudes, higher motivation and/or easier physical access to the system benefit more from it than their peers of the reverse profile, formative assessments could widen the existing social discrepancies in education rather than narrowing them. Whether this concern is reasonable or not needs to be scrutinized in carefully designed empirical studies

that track students' learning progress over time, control for any endogeneity bias, and consider the didactic method of teaching. Some didactic setups indeed might widen existing gaps, while others might do the opposite.

The collection of previously unexamined data in educational environments may lead to unintentional leaks about students and/or teachers. These accidental discoveries may range from trivial matters, such as secret friendship between two students (e.g., when log-in times and selection of items are correlated for two students), to more serious affairs, such as bullying or family disruption (e.g., when sharp declines in performance are detected and cannot otherwise be explained). While well documented in the medical research, the manner of dealing with such incidents is yet to be explored in the domain of formative assessments. Finally, the creation of large databases about students' knowledge and beliefs at such a young age raises concerns regarding the potential dual use of these data. While the term 'dual use' has been traditionally used for technology—designed for civilian purposes but with potential military applications—we believe that the recent revelations such as the Cambridge Analytica case illustrates that the capacity for data misuse exceeds the boundaries of this definition. In summary, it is extremely important to carefully consider the manner in which data are collected and disseminated.

In addition to ethical considerations, privacy issues arising from data collection are a serious concern in all kinds of computer-based assessment systems, and even more serious as systems grow both in scale (i.e., the number of students) and scope (i.e., the amount of data, also known as 'big data'). The existing guidelines, however, are surprisingly silent on data protection and privacy. The International Test Commission (2006), for instance, defers to 'local data protection and privacy legislation' (p. 166), whereby most systems incorporate instances of privacy management (e.g., Plichart et al., 2004). We believe that there are two major issues that must be taken into account here. First, when building computer-based assessment systems, a careful consideration of the regulations dealing with the protection of personal data (e.g., GDPR in Europe or COPPA in the United States) is crucial. These legislations address issues that have an effect on how technology has to be designed and deployed. They require, for example, clear statements respecting the nature of the data collected, the purpose for which they have been collected, strict control on individuals who can access the data, the acquisition of consent (parental consent in case of minors), and transparency of data treatment within the system. The intricate educational ecosystem alongside the complexity of algorithms used make some of these tasks extremely difficult.

The design of computer-based assessment systems should always take privacy seriously. Formative assessments, as noted earlier, make learning visible not only to students or teachers but potentially to all parties involved. Also, special caution needs to be exercised when assessment data are being matched with other sources of data (e.g., socioeconomic status or language spoken at home), especially when individual students become identifiable. Indeed, large-scale, computer-based assessment systems must deal with the inherent dilemma between privacy and the right to self-determination over one's own data. However, there is a scientific and administrative desire for rich and abundant data for research and administrative purposes. Thus, care has to be taken that the data collated are strictly necessary in use and exposure. This in some cases may be achieved using advanced privacy-enhancing technologies, such as the processing of encrypted data or anonymization of communication. How to integrate these protection technologies in the workflow of educational tools is a promising subject for future research.

CONCLUSION AND OUTLOOK

In this paper, we discussed the epistemological, methodological, and practical aspects of computer-based tools for formative student assessment, which aims to support learning and data-based decision making. In view of the effects of formative assessment and the benefits of data-based decision making, we are convinced that such tools can offer many advantages, compared with more traditional ways of providing feedback and making educational decisions. From an epistemological perspective, the most compelling advantage lies in the anticipated improvement of validity in computer-based tools, compared with feedback procedures based on teacher intuition and other unsystematic approaches. We have argued that these improvements can extend to all levels of the interpretive argument, ranging from scoring to generalization, extrapolation, and interpretation of results. Obviously, it is difficult to quantify these improvements in advance; however, given the number of aspects involved, one can assume that the scope of improvement will be substantial.

A second advantage of computer-based tools for formative assessment and data-based decision making is their considerable potential for enhancement in terms of availability, versatility, and flexibility at a small cost (in terms of organization and time) for the teachers and students involved. They provide options on the length of assessments, the time of administration, and competencies or topics that are currently relevant. Teachers, for example, can offer them to all their students or only to those whom they consider to be the most in need. Students can choose to run assessments on a regular basis or when they feel that one is necessary. These versatility and flexibility features are a direct function of the size of the item bank; however, once the curriculum has been covered in sufficient breadth and depth, the combinatorial prospects of creating tests can grow considerably.

Computer-based formative assessments have further advantage. They may be used to alleviate social disparities in learning and allow weak students to benefit from an idiosyncratic standard of reference. They can positively influence instruction by improving teachers' curriculum orientation and systematic planning, and contribute to promoting a positive testing culture in schools, in which assessments are not regarded as an external threat, but rather as a beneficial tool.

A flawless, state-of-the-art computer-based tool for the formative assessment needs to keep pace with the current massive technological advancements. Three developments are likely to influence what such systems will look like in the future. The first is the implementation of innovative item formats with interactive

elements that allow assessing students' productive competencies (see Goldin et al., 2017). Such items could contain simulations of conversations with interactive chat bots, writing assignments that are automatically scored with respective algorithms, or geometrical construction tasks with interactive elements. All these would make full use of the computer-based platform and allow assessing both outcomes and the problem-solving processes.

The second potential enhancement resides at the methodological level. By using information on both learning processes and outcomes and reverting to this 'big data,' constantly produced by the system, one could start using such systems as tools for cognitive diagnostics and learning analytics. Cognitive diagnostics instruments enable an in-depth assessment of students' competence levels and automatic presentation of items and tests following suggestions offered based on the collated empirical evidence; these data about the competencies are needed to answer the items and understand how these competencies relate to each other for each individual student. These relations could use cognitive models (e.g., Frischkorn and Schubert, 2018) as a starting point and be further refined by means of automated experiments so that the algorithms could learn by themselves what works best for which students and when. All this is closely related to the concepts and methods put forward in the emerging field of learning analytics (see Siemens, 2013). Here, there is also the idea to discover hidden relations in data but the focus is more on informing and empowering teachers and students about the learning process. A case in point are systems such as the 'Course Signals' at Purdue University (presented in Clow, 2013) that are used to predict success and failure in specific courses based on demographic characteristics, previous academic history, interaction with the system itself and performance on the course to date. This can be done very early during the course and as a consequence, instructors can trigger several interventions meant to prevent failure. Formative feedback systems such as the one introduced above are perfectly suitable as a rich data source for this kind of applications.

Third, given the growing importance of lifelong learning and the popularity of informal learning, it is unlikely that the future of computer-based formative assessments will remain restricted to schools and other educational institutions. This trend is likely to promote personalized learning environments, potentially available to everybody and for a broad range of topics. Combined with innovative and appealing item formats and supported by powerful diagnostic algorithms, we may eventually arrive at truly intelligent tutoring systems that are well-integrated into our daily lives.

AUTHOR CONTRIBUTIONS

UM developed the concept and chaired the practical implementation of the formative assessment system, MINDSTEPS, used here as a sample case. UM and SB were equally involved in developing its methodological foundations. MT drafted the article based on contributions by all authors, particularly UM who wrote on the theoretical background of formative assessments and SB who focused on the methodological and practical issues. All authors have revised the draft and approved the final version to be submitted.

ACKNOWLEDGMENTS

We are grateful to all the teachers and students who supported the development of our system. We are thankful that they agreed to provide user feedback and participate in the pilot studies. We are particularly grateful to Carmela Troncoso from the École Polytechnique Fédérale in Lausanne (EPFL) for her invaluable input on ethical and privacy issues.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Appl. Psych. Meas.* 13, 113–127. doi: 10.1177/014662168901300201

Akers, L., Del Grosso, P., Snell, E., Atkins-Burnett, S., Wasik, B. A., Carta, J., et al. (2016). Tailored teaching: emerging themes from the literature on teachers'use of ongoing child assessment to individualize instruction. *NHSA Dialog* 18, 133–150.

Ammons, R. B. (1956). Effects of knowledge of performance: a survey and tentative theoretical formulation. *J. Gen. Psychol.* 54, 279–299. doi: 10.1080/00221309.1956.9920284

Asseburg, R., and Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychol. Test. Assess. Model.* 55, 92–104.

Assessment Reform Group (1999). *Assessment for Learning: Beyond the Black Box. Cambridge, United Kingdom: University of Cambridge School of Education*. Available at http://www.nuffieldfoundation.org/sites/default/files/files/beyond_blackbox.pdf [accessed November 9, 2017]

Beaton, A. E., and Allen, N. L. (1992). Interpreting scales through scale anchoring. *J. Educ. Behav. Stat.* 17, 191–204. doi: 10.3102/10769986017002191

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Princ. Pol. Pract.* 18, 5–25. doi: 10.1080/0969594X.2010.513678

Berger, S., Moser, U., and Verschoor, A. J. (2015). "Development of an online item bank for adaptive formative assessment," in *Paper presented at the AEA-Europe Conference*, Glasgow, 5–7.

Bernhardt, V. (2003). Using data to improve student achievement. *Educ. Leadersh.* 60, 26–30.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 397–479.

Black, P., and William, D. (1998). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan* 80, 139–148.

Bloom, B. S. (1969). "Some theoretical issues relating to educational evaluation," in *Educational Evaluation: New Roles, New Means (The 63rd Handbook of the National Society for the Study of Education*, Vol. 69, Part 2, ed. R. W. Tyler (Chicago, IL: University of Chicago Press), 26–50.

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

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educ. Meas.* 22, 5–12. doi: 10.1111/j.1745-3992.2003.tb00139.x

Brown, B., and Cowie, B. (2001). The characteristics of formative assessment in science education. *Sci. Educ.* 85, 536–553. doi: 10.1002/sce.1022

Brown, G. T. L. (2013). "asTTle– A National Testing System for Formative Assessment: how the national testing policy ended up helping schools and

teachers," in *A National Developmental and Negotiated Approach to School and Curriculum Evaluation*, eds M. K. Lai and S. Kushner (London: Emerald Group), 39–56.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. Berkeley, CA: University of California Press, doi: 10.3102/00346543065003245

Butler, D. L., and Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical synthesis. *Rev. Educ. Res.* 65, 245–281. doi: 10.3102/00346543065003245

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230

Camilli, G., and Shepard, L. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage Publications.

Campbell, C., and Levin, B. (2009). Using data to support educational improvement. *Educ. Assess. Eval. Acc.* 21, 47–65. doi: 10.1007/s11092-008-9063-x

Carlson, D., Borman, G., and Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educ. Eval. Policy Anal.* 33, 378–398. doi: 10.3102/0162373711412765

Carlson, J. E. (2011). "Statistical models for vertical linking," in *Statistical Models for Test Equating, Scaling, and Linking*, ed. A. A. von Davier (New York, NY: Springer Science+Business Media), 59–70.

Cawelti, G., and Protheroe, N. (2001). *High Student Achievement: How Six School Districts Changed into High-performance Systems*. Arlington, VA: Educational Research Service.

Chou, Y.-T., and Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educ. Psychol. Meas.* 70, 717–731. doi: 10.1177/0013164410379322

Cizek, G. J. (2012). Defining and distinguishing validity: interpretations of score meaning and justification of test use. *Psychol. Methods* 17, 31–43. doi: 10.1037/a0026975

Clow, D. (2013). An overview of learning analytics. *Teach. High Educ.* 18, 683–695. doi: 10.1080/13562517.2013.827653

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

Conole, G., and Warburton, B. (2005). A review of computer-assisted assessment. *Res. Learn. Tech.* 13, 17–31. doi: 10.1080/0968776042000339772

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Dadey, N., and Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Pract. Assess. Res. Eval.* 17, 1–13.

Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated experiments on ad privacy settings: a tale of opacity, choice, and discrimination. *Lect. Notes Comput. Sci.* 2015, 92–112.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: Guilford Press.

DeLuca, C., and Bellara, A. (2013). The current state of assessment education: aligning policy, standards, and teacher education curriculum. *J. Teach. Educ.* 64, 356–372. doi: 10.1177/0022487113488144

DeLuca, C., and Klinger, D. A. (2010). Assessment literacy development: identifying gaps in teacher candidates' learning. *Assess. Educ. Princ. Pol. Pract.* 17, 419–438. doi: 10.1080/0969594X.2010.516643

Dignath, C., Buettner, G., and Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educ. Res. Rev. Neth.* 3, 101–129. doi: 10.1016/j.edurev.2008.02.003

Dunn, K. E., and Mulvenon, S. W. (2009). A critical review of research on formative assessments: the limited scientific evidence of the impact of formative assessment in education. *Pract. Assess. Res. Eval.* 14:7.

Earl, L., and Katz, S. (2006). *Leading in a Data Rich World*. London: Corwin press.

Elo, A. (1978). *The Rating of Chessplayers: Past and Present*. New York, NY: Arco Publishers.

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., and Leutner, D. (2013). Kompetenzmodellierung: struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Z. Erziehwiss* 16, 5–20. doi: 10.1007/s11618-013-0379-z

Foster, G., and Ysseldyke, J. (1976). Expectancy and halo effects as a result of artificially induced teacher bias. *Contemp. Educ. Psychol.* 1, 37–45. doi: 10.1016/0361-476X(76)90005-9

Francis, R. A. (2008). An investigation into the receptivity of undergraduate students to assessment empowerment. *Assess. Eval. High. Educ.* 33, 547–557. doi: 10.1080/02602930701698991

Frederiksen, J. R., and Collins, A. (1989). A systems approach to educational testing. *Educ. Res.* 18, 27–32. doi: 10.3102/0013189X018009027

Frischkorn, G.-T., and Schubert, A.-L. (2018). Cognitive models in intelligence research: advantages and recommendations for their application. *J. Intell.* 6:34. doi: 10.3390/jintelligence6030034

Goldin, I., Narciss, S., Foltz, P., and Bauer, M. (2017). New directions in formative feedback in interactive learning environments. *Int. J. Artif. Intell. Educ.* 27, 385–392. doi: 10.1007/s40593-016-0135-7

Greeno, J. G. (1989). A perspective on thinking. *Am. Psychol.* 44, 134–141. doi: 10.1037/0003-066X.44.2.134

Hattie, J. A. C. (1999). "Influences on student learning," in *Inaugural Lecture Held at the University of Auckland*, New Zealand, 2.

Hattie, J. A. C., and Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *J. Educ. Techn. Syst.* 36, 189–201. doi: 10.2190/ET.36.2.g

Hattie, J. A. C., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Havnes, A., Smith, K., Dysthe, O., and Ludvigsen, K. (2012). Formative assessment and feedback: making learning visible. *Stud. Educ. Eval.* 38, 21–27. doi: 10.1016/j.stueduc.2012.04.001

Herman, J., and Winter, L. (2011). *The Turnaround Toolkit: Managing Rapid, Sustainable School Improvement*. London: Corwin press.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *J. Educ. Behav. Stat.* 23, 35–56. doi: 10.3102/10769986023001035

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *Int. J. Test.* 6, 143–171. doi: 10.1207/s15327574ijt0602_4

Kane, M. T. (2006). "Validation," in *Educational Measurement*, ed. R. L. Brennan (Westport, CT: American Council on Education), 17–64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Measure.* 50, 1–73. doi: 10.1111/jedm.12000

Klieme, E., Hartig, J., and Rauch, D. (2008). "The concept of competence in educational contexts," in *Assessment of Competencies in Educational Contexts*, eds J. Hartig, E. Klieme, and D. Leutner (Cambridge, MA: Hogrefe and Huber Publishers), 3–22.

Kluger, A. N., and DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol. Bull.* 119, 254–284. doi: 10.1037/0033-2909.119.2.254

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teach. Educ.* 22, 255–276. doi: 10.1080/10476210.2011.593164

Kolen, M. J., and Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. New York, NY: Springer. doi: 10.1007/978-1-4939-0317-7

Köller, O. (2014). "Entwicklung und Erträge der jüngeren empirischen Bildungsforschung," in *Das Selbstverständnis der Erziehungswissenschaft: Geschichte und Gegenwart*, eds R. Fatke and J. Oelkers (Weinheim: Beltz), 102–122.

Lai, M. K., McNaughton, S., Amituanai-Toloa, M., Turner, R., and Hsiao, S. (2009). Sustained acceleration of achievement in reading comprehension: the New Zealand experience. *Read. Res. Q.* 44, 30–56. doi: 10.1598/RRQ.44.1.2

Lai, M. K., and Schildkamp, K. (2013). "Data-based Decision Making: an Overview," in *Data-based Decision Making in Education: Challenges and Opportunities*, eds K. Schildkamp, M. K. Lai, and L. Earl (Dordrecht: Springer), 9–21.

Lee, J.-E. R., and Nass, C. I. (2010). "Trust in Computers: the Computers-Are-Social-Actors (CASA) paradigm and trustworthiness perception in human-computer communication," in *Trust and Technology in a Ubiquitous Modern*

Environment: Theoretical and Methodological Perspectives, eds D. Latusek and A. Gerbasi (Hershey, PA: Information Science Reference), 1–15.

Levy, H. M. (2008). Meeting the needs of all students through differentiated instruction: helping every child reach and exceed standards. *Clear. House J. Educ. Strateg. Issues Ideas* 81, 161–164. doi: 10.3200/TCHS.81.4.161-164

Linn, R. L. (2006). "The standards for educational and psychological testing: guidance in test development," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates), 27–38.

Linn, R. L., Baker, E. L., and Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Appl. Psych. Meas.* 24, 15–21. doi: 10.3102/0013189X020008015

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Love, N. (2008). *Using Data to Improve Learning for All: A Collaborative Inquiry Approach*. London: Corwin press.

Maier, U. (2015). *Leistungsdiagnostik in Schule und Unterricht*. Bad Heilbrunn: Julius Klinkhardt.

McDonald, R. P., and Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivar. Behav. Res.* 30, 23–40. doi: 10.1207/s15327906mbr3001_2

McKown, C., and Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *J. Sch. Psychol.* 46, 235–261. doi: 10.1016/j.jsp.2007.05.001

McManus, S. (2008). *Attributes of Effective Formative Assessment*. Washington, DC: Council for Chief State School Officers.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: implications for theory and practice. *Educ. Meas.* 22, 34–43. doi: 10.1111/j.1745-3992.2003.tb00142.x

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychol. Bull.* 115, 300–307. doi: 10.1037/0033-2909.115.2.300

Mertler, C. A. (2004). Secondary teachers' assessment literacy: does classroom experience make a difference? *Am. Second. Educ.* 33, 49–64.

Messick, S. (1989). "Validity," in *Educational Measurement*, ed. R. L. Linn (New York, NY: American Council on Education), 13–103.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Ministry of Education (1994). *Assessment: Policy to Practice*. Wellington: Learning Media.

Moser, U. (2009). "Test," in *Handwörterbuch Erziehungswissenschaft*, eds S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee, and J. Oelkers (Weinheim: Beltz), 866–880.

Moser, U. (2016). "Kompetenzorientiert - adaptiv - digital: adaptives Lernen und Testen für eine zeitgemäße Evaluation des Lernfortschritts im Schulunterricht," in *Digitale Bildungslandschaften*, eds A.-W. Scheer and C. Wachter (Saarbrücken: IMC AG), 327–339.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educ. Measure. Issues Pract.* 22, 13–25. doi: 10.1111/j.1745-3992.2003.tb00140.x

Newton, P. E., and Baird, J.-A. (2016). The great validity debate. *Assess. Educ. Princ. Pol. Pract.* 23, 173–177. doi: 10.1080/0969594X.2016.1172871

Newton, P. E., and Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment*. Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781446288856

Nicol, D. J. (2009). Assessment for learning self-regulation: enhancing achievement in the first year using learning technologies. *Assess. Eval. High. Educ.* 34, 335–352. doi: 10.1080/02602930802255139

Nicol, D. J., and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Stud. High. Educ.* 31, 199–218. doi: 10.1080/03075070600572090

Orlando, M., and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Appl. Psych. Measure.* 24, 50–64. doi: 10.1177/01466216000241003

Pellegrino, J. W., Chudowski, N., and Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Plichart, P., Jadoul, R., Vandenbelle, L., and Latour, T. (2004). "TAO: a collaborative distributed computer-based assessment framework built on semantic web standards," in *Paper presented at the International Conference on Advances in Intelligent Systems (AISTA 2004)*, Luxembourg, 15–18.

Popham, W. J. (2008). *Transformative Assessment*. Alexandria, VA: ASCD.

Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental? *Theor. Pract.* 48, 4–11. doi: 10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: a teacher educator's confession. *Teach. Educat.* 46, 265–273. doi: 10.1080/08878730.2011.605048

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Rupp, A. A., and Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educ. Psychol. Measure.* 66, 63–84. doi: 10.1177/0013164404273942

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assess. Eval. High. Educ.* 34, 159–179. doi: 10.1080/02602930801956059

Sampson, D. G., and Fytros, D. (2008). "Competence models in technology-enhanced competency-based learning," in *International Handbook on Information Technologies for Education and Training*, eds H. H. Adelsberger, J. M. Kinshuk, S. Pawlowski, and D. Sampson (New York, NY: Springer), 1–25.

Savulescu, J. (2006). Justice, fairness, and enhancement. *Ann. N.Y. Acad. Sci.* 1093, 321–338. doi: 10.1196/annals.1382.021

Scheerens, J., Glas, C., and Thomas, S. M. (2003). *Educational Evaluation, Assessment, and Monitoring: A Systemic Approach*. New York, NY: Taylor and Francis.

Schildkamp, K., and Ehren, M. (2013). "From 'Intuition'- to 'Data'-based decision making in Dutch secondary schools," in *Data-based Decision Making in Education: Challenges and Opportunities*, eds K. Schildkamp, M. K. Lai, and L. Earl (Dordrecht: Springer), 49–67.

Schildkamp, K., Lai, M. K., and Earl, L. (2013). *Data-based Decision Making in Education: Challenges and Opportunities*. Dodrecht: Springer. doi: 10.1007/978-94-007-4816-3

Shavelson, R. J. (2008). Guest editor's introduction. *Appl. Measure. Educ.* 21, 293–294. doi: 10.1080/08957340802347613

Shepard, L. A. (2006). "Classroom assessment," in *Educational Measurement*, ed. R. L. Brennan (Westport, CT: American Council on Education), 623–646.

Shepard, L. A. (2008). "Formative assessment: caveat emptor," in *The Future of Assessment: Shaping Teaching and Learning*, ed. C. A. Dwyer (New York, NY: Lawrence Erlbaum Associates), 279–303.

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

Siemens, G. (2013). Learning analytics: the emergence of a discipline. *Am. Behav. Sci.* 51, 1380–1400. doi: 10.1016/j.ijrobp.2018.08.032

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., and McPhail, R. (2011). Assessment literacy and student learning: the case for explicitly developing students "assessment literacy." *Assess. Eval. High. Educ.* 38, 44–60. doi: 10.1080/02602938.2011.598636

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educ. Meas.* 22, 26–33. doi: 10.1111/j.1745-3992.2003.tb00141.x

Stiggins, R. J. (2005). *Student-involved Assessment for Learning*. Upper Saddle River, NJ: Pearson.

Stobart, G. (2012). "Validity in formative assessment," in *Assessment and Learning*, ed. J. Gardner (London: Sage Publications), 233–242.

Suárez-Falcón, J. C., and Glas, C. A. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *Brit. J. Math. Stat. Psychol.* 56, 127–143. doi: 10.1348/000711003321645395

Thompson, N. A., and Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Pract. Assess. Res. Eval.* 16, 1–9.

Tong, Y., and Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Appl. Meas. Educ.* 20, 227–253. doi: 10.1080/08957340701301207

Vale, C. D. (2006). "Computerized item banking," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates), 261–285.

van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., and Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assess. Educ. Princ. Pol. Pract.* 22, 324–343. doi: 10.1080/0969594X.2014.999024

van der Linden, W. J., and Glas, C. A. (2010). *Elements of Adaptive Testing*. New York, NY: Springer. doi: 10.1007/978-0-387-85461-8

Veldkamp, B. P., and van der Linden, W. J. (2010). "Designing item pools for adaptive testing," in *Elements of Adaptive Testing*, eds W. J. van der Linden and C. A. W. Glas (New York, NY: Springer), 231–245.

Verschoor, A. J., and Berger, S. (2015). "Computerized adaptive testing with online JML calibration," in *Paper presented at the IACAT Conference*, Cambridge, MA, 14–16.

Volante, L., and Fazio, X. (2007). Exploring teacher candidates' assessment literacy: implications for teacher education reform and professional development. *Can. J. Educ.* 30, 749–770. doi: 10.2307/20466661

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410605931

Wainer, H., and Mislevy, R. J. (2000). "Item response theory, item calibration, and proficiency estimation," in *Computerized Adaptive Testing: A Primer*, ed. H. Wainer (Mahwah, NJ: Lawrence Erlbaum Associates), 61–100. doi: 10.4324/9781410605931

Wang, T.-H., Wang, K.-H., and Huang, S.-C. (2008). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Comput. Educ.* 51, 448–462. doi: 10.1016/j.compedu.2007.06.010

Webb, N. L. (2006). "Identifying content for student achievement tests," in *Handbook of Test Development*, eds S. M. Downing and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates), 155–180.

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

Wilson, S. M., Floden, R. E., and Ferrini-Mundy, J. (2001). *Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations*. Washington, DC: University of Washington Center for the Study of Teaching and Policy.

Wise, S. L., and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educ. Assess.* 10, 1–17. doi: 10.1207/s15326977ea1001_1

Xu, Y., and Brown, G. T. L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

Conflict of Interest Statement: The formative assessment system used as a case in point in this study was commissioned by the *Bildungsraum Nordwestschweiz*, which funded its development and operation. The authors have disclosed to the article's editor the details of the financial relation between the initiative and the sponsoring institution.

frontiers in Education

REVIEW
published: 25 July 2018
doi: 10.3389/feduc.2018.00061

Why Is Implementation Science Important for Intervention Design and Evaluation Within Educational Settings?

*Taryn Moir**

Educational Psychological Services, North Ayrshire Council, Irvine, United Kingdom

The current challenging economic climate demands, more than ever, value for money in service delivery. Every service is required to maximize positive outcomes in the most cost-effective way. To date, a smorgasbord of interventions have been designed to benefit society. Those worthy of attention have solid foundations in empirical research, offering service providers reassurance that positive outcomes are assured; many of these programmes lie within the field of education and everyday school practice. However, often even these highly supported programmes yield poor results due to poor implementation. Implementation science is the study of the components necessary to promote authentic adoption of evidence-based interventions, thereby increasing their effectiveness. Following a brief definition of key terms and theories, this article will go on to discuss why implementation is not a straightforward process. To do so, this article will draw upon examples of evidence-based but poorly implemented school programmes. Having acknowledged how good implementation positively affects sustainability, we will then look at the growing number of frameworks for practice within this field. One such framework, the Core Components Model, will be used to facilitate discussion about the processes of successful design and evaluation. This article will continue by illustrating how the quality of implementation has directly affected the sustainability of the Incredible Years programmes and the Promoting Alternative Thinking Strategies (PATHS) curriculum. Then, by analyzing implementation science, some of the challenges currently faced within this field will be highlighted and areas for further research discussed. This article will then link to the implications for educational psychologists (EPs) and will conclude that implementation science is crucial to the design and evaluation of interventions, and that the EP is in an ideal position to support sustainable positive change.

Keywords: fidelity, implementation science, readiness to change, education intervention, schools

OPEN ACCESS

Edited by:
Ulrich Dettweiler,
University of Stavanger, Norway

Reviewed by:
Sharinaz Hassan,
Curtin University, Australia
Renae L. Smith-Ray,
Walgreens, United States

***Correspondence:**
Taryn Moir
tarynmoir@north-ayrshire.gcsx.gov.uk

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *05 July 2017*
Accepted: *04 July 2018*
Published: *25 July 2018*

Citation:
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

INTRODUCTION

Implementation science is the study of how evidence-based programmes can be embedded to maximize successful outcomes (Kelly and Perkins, 2012). It is concerned with using a systematic and scientific approach to identify the range of factors which are likely to facilitate administration of an intervention. By studying the success and failure of intervention adoption,

within various disciplines, this scientific approach offers greater understanding of how accredited strategies can be successfully transferred to new contexts. Implementation science, therefore, bridges the gap between theory and effective practice (Fixsen et al., 2009b). Research studies in this field highlight the factors and variables central to successful adoption and sustainability of programmes. Adopting new programmes necessitates change. Implementation science recognizes that people need to be ready for change and that creating optimal conditions for an intervention is crucial to its maintenance. Therefore, implementation science is fundamental to the design of successful interventions. In addition, to understand true effectiveness, both the intervention and its implementation need to be evaluated to fully understand outcomes and impacts (Kelly and Perkins, 2012). Although implementation science has been employed for some time in clinical, health and community settings, its application within the educational domain is still relatively new and there are many areas for further research within this discipline (Lyon et al., 2018).

Definitions Within Implementation Science

An intervention is defined as "a specified set of activities designed to put into practice an activity of known dimensions" (Fixsen et al., 2005). When the intervention has been evaluated as having yielded the expected results, it can be considered effective within targeted populations and settings. For interventions to be effective, it has been persuasively argued (Fixsen et al., 2005) that the programme should be adopted with fidelity, as this ensures sustainability. This means that the programme should have the same content, coverage, frequency and duration as was intended by the designers (Carroll et al., 2007).

Key to intervention design and evaluation are the core components, which are regarded as the essential aspects of the intervention without which the practice or programme will fail to be sustainable or effective (Fixsen et al., 2005).

The Underpinning Theory of Implementation Science

Personal readiness for change depends upon having the capability, opportunity and motivation to change behavior (Michie et al., 2009; Fallon et al., 2018). However, achieving organizational readiness for change is far more complicated. Ideally, individuals within an organization should feel committed and confident in their collective ability to change practices. This is considered to be of critical importance for success in implementing change within an organization (Armenakis et al., 1993; Weiner, 2009). Indeed, it has been suggested that failing to account for such readiness for change can be responsible for a significant proportion of large-scale change efforts being successful or not (Kotter, 1996; Fallon et al., 2018).

Theories of organizational change illustrate the dynamic web of influences within a complex multilevel, multifaceted construct. A three-stage model has been described by Lewin (1951). Stage one attempts to unfreeze fixed mindsets and motivate individuals for change; stage two takes individuals through a transition that enables communication to identify new norms and attitudes; stage three is the embedding of these new ideas into practice. Implementation science describes similar phases in organizing change; these are discussed in the "Frameworks for Practice" section below.

Senge (1990) states: "We tend to focus on snapshots of isolated parts of the system and wonder why our deepest problem never seems to get solved." Implementation science, therefore, acknowledges the impact of systems and coheres with ecological systems theory (Bronfenbrenner, 1979); a constructionist model which illustrates how complex organizational systems need to be aware of wider political, social and cultural influences. Bronfenbrenner illustrates the need for a well-organized and consistent approach. The key internal components of the programme have to be compatible with external influences for full implementation to occur, as seen in **Figure 1** below. Working across these systems with a collaborative focus is necessary for success (Maher et al., 2009).

For any intervention to be successfully embedded, socioeconomic and cultural environments need to be acknowledged, because their variables impact on implementation success. Individuals' readiness for change and group dynamics are enmeshed within their relevant influencing ecological systems. Therefore, when designing and evaluating school-based programmes, it is necessary to clearly understand community cultures and take them into account.

Poor Implementation

There is a tendency for schools to buy new intervention packs marketed as solving all their problems without reference to empirical evidence (Slavin, 2002). In addition, they do not ask many questions about why previously tried programmes have failed. Doing so would, perhaps, be more insightful and cost-effective. Furthermore, while good interventions can be badly implemented, poor interventions can equally be implemented successfully. Therefore, potentially, a theory-based programme may be disbanded while poorly supported interventions may run for years (Kelly and Perkins, 2012).

In essence, having theoretically sound programmes does not, in itself, ensure successful implementation.

One example is a study in Uganda where an empirically supported school-based AIDS education programme was found to be ineffective. Closer examination, using multiple methods, found that this was because it was poorly implemented. Key activities, including role-play, had not been given adequate time. This was due in part to a lack of facilities and in part to a lack of confidence in an intervention concerning such a controversial issue (Kinsman et al., 2001). Here, poor implementation resulted in time, money and resources being wasted.

Furthermore, Barnett found in his review of 36 public programmes that the impacts of empirically based early childhood programmes were affected by the quality of implementation (Barnett, 1995). Also, Greenberg et al. (2005) stated that often, "within-school" initiatives are not implemented with the same quality as the programme designers initially intended, resulting in poor outcomes. Therefore, it is imperative that schools begin to actively embrace implementation considerations when designing and evaluating

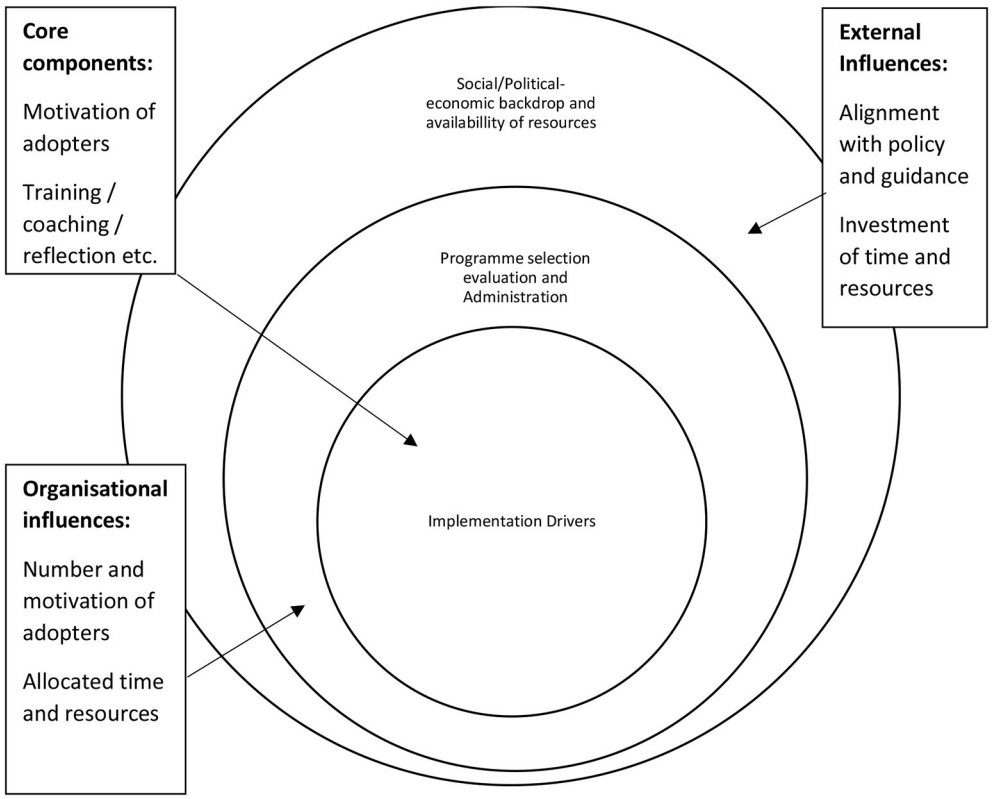

FIGURE 1 | Conceptual model of the ecology of implementation.

initiatives. This will be more cost-effective overall and more efficient in promoting positive change.

FRAMEWORKS FOR PRACTICE

There are many frameworks, from various specific disciplines (Birken et al., 2017). However, Tabak et al. (2012) review of 61 models and Meyers et al. (2012) synthesis of 25 frameworks both indicate that many share commonalities, both in their description of stages of implementation and their core components. One example is CASEL (2012), which offers 10 steps and six sustainability factors. Michie et al. (2011) identified 19 frameworks in their systematic enquiry into characterizing and designing behavior change interventions. From their findings, they developed a "Behavior Change Wheel," which described the key factors of change as being opportunity, capability and motivation. They suggested that the wheel can be used as a framework to identify relevant interventions.

In addition, there is a conceptual framework to measure five indexes of implementation fidelity (Carroll et al., 2007). Measuring fidelity is one way of evaluating implementation, a key process which is just as important as the evaluation of the programme (Fixsen et al., 2005). Initially, three indexes of fidelity were identified, these being exposure, adherence and quality of implementation (Klimes-Dougan et al., 2009). However, Mihalic (2012) later added dimensions of participant responsiveness and programme differentiation. This framework is one of several that are useful when evaluating programme implementation fidelity. More recently, Rojas-Andrade and Bahamondes (2018) analyzed existing data on implementation and found that implementation fidelity of adherence, quality of intervention, exposure to intervention and receptiveness were linked with outcomes 40% of the time, with the latter two indicators having the strongest associations. Measuring fidelity can also be measured via fidelity observations (Pettigrew et al., 2013) perhaps using video (Johnson et al., 2010).

Greenberg et al. (2005) described three phases of implementation—pre-adoption, delivery, and post-adoption—and advised that they should be incorporated into intervention design. Alternatively, the Stages of Implementation Framework (Fixsen et al., 2005) describes six additive stages toward full implementation of programmes. These are:

- current situation exploration
- consideration of change, or installation phase
- preparation for change, or initial implementation phase
- full implementation, where change is being engaged in
- innovation, where after practicing interventions with pure fidelity, subtle adaptations are made to best fit the user
- maintenance of procedures to ensure sustainability

While the selection of implementation frameworks is often driven by previous exposure or convenience rather than theory

(Birken et al., 2017), one framework, the Implementation Components Framework (Fixsen et al., 2009a), is based upon a synthesis of 377 implementation articles. It offers a conceptual model concerned with fundamental aspects necessary for implementation to be successful and identifies the key competency drivers, which are the mechanisms that underpin and therefore sustain implementation:

- staff selection
- pre-service/INSET Training
- consultation and coaching
- staff performance evaluation

Furthermore, organization drivers are described as the mechanisms to sustain systems environments and facilitate implementation:

- decision support data systems
- facilitative administrative support
- systems interventions

This article will continue by looking at each of these drivers as they give great insights into how interventions should be designed and evaluated.

Staff Selection

Getting all staff on board and building a philosophy of joint working is paramount to the success of any new initiative (Maher et al., 2009). In Klimes-Dougan et al. (2009) study of the Early Risers Prevention Programme, she found that staff members' personalities, and not their prior experience, were a predictor of the likelihood that an intervention would be implemented with fidelity. Personality factors include breadth of skill, openness, conscientiousness and levels of commitment in the face of challenges. Staff selection is the first key design consideration in any intervention; however, within the real-world context this can be difficult as it depends upon availability of personnel.

In addition, it is essential to ensure that there are lead players within the organization to guide new interventions. Ideally, there should be a dedicated implementation team. Fixsen et al. (2001) found in their analysis of implementation that designated teams led to an 80% success rate in implementation over a 3-year period, compared to 14% success over a 17-year period for programmes that did not have such teams. It must be noted that this comparison only incorporated two studies as only two could be identified as having the same implementation measures. However, such a significant difference in results still persuasively argues for having dedicated implementation staff. The conclusion can be drawn that without a key stakeholder within the organization who has decision-making authority and the ability to persuade others in the process of implementation, interventions may fall by the wayside or become diluted.

Pre-service/INSET Training

Making a change in organizational practices necessitates training. A threat to effective training can be the difficulty of predicting training needs. Therefore, before any training, best implementation science practice dictates that individuals should complete a pre-INSET questionnaire: a check for readiness. This both offers the facilitator the opportunity to set a benchmark for current knowledge, skill and motivation, and also allows for the negotiation of truly relevant and differentiated sessions (Dunst and Trivette, 2009; Fallon et al., 2018). The process should become a partnership between all involved, as participants' ownership of training increases motivation (Gregson and Sturko, 2007).

In addition, the instructor's characteristics have also been found to be associated with the quality of overall implementation (Spoth et al., 2007); recommendations have been made for having enthusiastic and committed facilitators.

Consultation and Coaching

Modern-day practices require staff to undergo continuous professional development to enhance their competencies. On-the-job coaching not only ensures that these practices will become enmeshed in everyday procedures, but also has the potential to promote a cycle of continuous development. Peer coaching facilitates the development of new school norms and offers the opportunity for sustainable ongoing practice (Joyce and Showers, 2002). Joyce and Showers (2002) found in their meta-analysis of teachers doing training that only 5% put newly learnt strategies into practice. However, coaching and on-the-job training after initial teaching sessions ensured that 95% of teachers used the newly learnt techniques. Coaching, therefore, has a massive impact on the effectiveness of training and should be built into intervention design.

In addition, not only should the coach be proficient, there should also be manuals and materials available to further support new practices (Fixsen et al., 2013). In Dane and Schneider (1998) meta-analysis, only 20% of programmes incorporated both support for staff and training and materials into new interventions. This is, therefore, an area for development.

Staff Performance Evaluation

Once the new methods have been practiced, reflection on the process and discussion with other practitioners will help further embed new ideas. If participants have struggled to put concepts into practice, problem-solving discussions at this stage will prevent the discontinuation of the programme (Kelly, 2012). Feedback from these sessions can be used to further enhance future training sessions; however, as the most successful interventions are those with the greatest fidelity, adaptations should not interfere with programmes' core components.

Decision Support Data Systems

Continual monitoring of implementation helps ensure programme sustainability. Multiple methods should be used to draw together information from a variety of sources, including quality performance indicators, service user feedback and organizational fidelity measures (Fixsen et al., 2005). Durlak (2010) argues that implementation can be measured on a continuum from 0 to 100%. The five indexes of the implementation fidelity model outlined above (Carroll et al., 2007) could potentially be used for this purpose.

Facilitative Administrative Support

Once the practices are becoming embedded, the senior management team (SMT) within the school should ensure that administrative systems, including policies and procedures, are coherent with the new practices. These can then inform and support these new systems.

Systems Interventions

This facet of implementation advises that the organization should observe national policy and other external systems and forces. A changing political climate influences the education system and will therefore directly impact on schools' needs and priorities.

To sum up this section, these core components are fundamental considerations for designing and evaluating interventions. This article will now illustrate how evidence-based programmes' outcomes correlate with implementation quality. Variations in implementation will also highlight associated issues.

OPTIMIZED IMPLEMENTATION?

Research into what works within schools is crucial as it helps authorities and governments to decide on the best ways to help communities. A programme should be empirically based and successfully implemented. Mintra (2012) also states that in addition to programme fidelity, good implementation relies upon building genuine and transparent partnerships. This is illustrated in the example of the implementation of the "Incredible Years" programme.

"Incredible Years" (IY) is anevidence-based programmes aimed at reducing children's aggression and behavioral problems (Webser-Stratton, 2012), yet the success of its implementation has varied. This is attributed to the quality of implementation fidelity. However, given the vast array of countries which have invested in IY, there has been a need to adapt the programme to meet cultural and contextual needs. As Ringwalt et al. (2003) states, adaptation is inevitable and therefore care should be taken to ensure the core components are not undermined. This, therefore, has necessitated the development of guidelines which maximize fidelity but allow flexibility (Reinke et al., 2011). This guidance sets out an eight-point process throughout the implementation phases and has led to optimum implementation across the world, including in Knowsley Central Primary Support Centre in England (CAST, 2012) and the Children and Parents' Service Early Intervention in Manchester (CAPS, 2012).

A similar theme regarding the balance between flexibility and fidelity was found by Jaycox et al. (2006), who looked at three different intervention programmes delivered and evaluated within schools. All were aimed at reducing dating violence in adolescence. Evaluation of each programme illustrated a negotiation between real-world applicability and a tight research design. However, they argued that for optimum implementation, flexibility within the constraints of the design is necessary.

Finally, Promoting Alternative Thinking Strategies (PATHS) (Greenberg and Kusche, 1996) is a "blueprint" programme developed to enhance social and emotional competencies in young children (Mihalic et al., 2001). Although it has a sound evidence base, well-designed implementation is also critical. Kam et al. (2003) evaluated implementation in a study concerning a group of children of low academic achievement living in areas of high deprivation. Their results confirmed the complexity of implementation within the school context and suggested that strong leadership from the school principal and the quality of implementation were predictors of the programme's success in reducing child aggression. Their findings again underline the importance of implementation fidelity with respect to programme dosage, quality of delivery and support and commitment. Furthermore, shockingly, backward trends in pro-social behavior were evident in two out of four establishments where the PATHS programme's implementation lacked sufficient integrity, even when anecdotal evidence suggested effective positive change (Kelly et al., 2012).

These studies highlight the necessity of implementation science considerations within programme design and evaluation. This article will now continue by acknowledging the threats and challenges associated with implementation science.

CHALLENGES/THREATS

Many interventions are implemented without acknowledging the role of implementation science. Leaders need to be aware of the importance of good implementation. This requires training, which is crucial—especially at these early stages of implementation science—to raise awareness of its significance in programme design and evaluation. Raising awareness has far-reaching consequences; therefore, the new language associated with implementation science needs to be taught and embraced (Axford and Morpeth, 2012). Within education (as within other domains), if implementation science is not regarded as important by leaders and the language is not learnt, then dynamic initiatives will fail (Bosworth et al., 2018).

Currently, little time is spent upon implementation (Sullivan et al., 2008); yet effective implementation is likely to take 2–4 years (Fixsen et al., 2009a), and it can take up to 20 years before initiatives are fully embedded into everyday practice (Ogden et al., 2012). However, within our current climate, there is pressure on many organizations, including schools, to make effective changes quickly. In a study of the effectiveness of cooperative learning in secondary schools, Topping et al. (2011) argue that this investment in time may make the cost-effectiveness of intervention questionable. This type of belief, which does not acknowledge the overall cost-effectiveness of these practices, may present barriers to promoting implementation science.

In addition, Carroll et al. (2007) found that the most common reason for deviations from fidelity was time restrictions. Potentially, this could be prevented if recognition of a programme's time commitment is made clear in the initial stages of design. This would ensure realistic goals are set for positive outcomes (Maher et al., 2009). Leaders and teachers need to recognize that it is far more effective to properly invest the necessary time into an initiative, rather than to poorly implement

a series of consecutive ineffective interventions over the same amount of time.

A further challenge is getting the right staff via stringent recruitment procedures: a core component of implementation. These staff, perhaps more highly sought after, may merit raised salaries, partly due to increased duties pertaining to implementation teams or steering groups. Unions and contracts may therefore be barriers, due to increased personnel responsibilities among staff. Furthermore, altering any historic systems within schools can be perceived negatively by either staff or unions. Therefore, funds need to be invested into each programme to cover these associated costs, and unfortunately economic issues are always pressing. Other barriers may include existing policies/procedures and local laws which may not reflect the ethos of implementation science. For example, implementation is a process which can take many years (Fixsen et al., 2009a), yet the cycle of government may lead politicians to be more interested in short- than long-term impact. In such cases it is therefore necessary to disseminate implementation science to policymakers to encourage investment in a longer-term vision of embedded evidence-based interventions.

Furthermore, the reality of many organizations, including schools, is that it is not practically possible to recruit new staff who are open to innovative practices or settings that can facilitate optimum implementation. Therefore, real-world settings need to account for this. For instance, in Scotland there is a national teacher staffing crisis (Hepburn, 2015), whereby rigorous selection of staff is an unobtainable luxury: application pools are small and there are high numbers of unfilled vacancies (Hepburn, 2015).

While there are many challenges, addressing these issues at the beginning of the implementation processes will ensure that interventions are effective, and over the long term, more cost-effective.

Implementation science has been successfully employed in such fields as public health and medicine (Glasgow and Emmons, 2007; Rabin et al., 2010; Scheirer, 2013). However, within education it is a comparatively new science (Lyon et al., 2018), and as such there are many areas for further research at all levels, from global to individual. Global-level areas for research include the development of a greater understanding of the true relationships between core components. This may further inform us whether the components are all-encompassing and whether the core components framework needs to be redefined (Fixsen et al., 2009a). In addition, the model would benefit from further research into each aspect of the framework. Sullivan et al. (2008) argue that this would open "the black box" to give us greater understanding of why this approach works.

Furthermore, while organizations are increasingly trying to ensure that implementation is evaluated, different approaches are being used. Therefore, one goal is to establish a commonality in approaches to the measurement of implementation. A meta-analysis of approaches could then clarify how best to evaluate all its aspects.

In addition, descriptions of interventions and details of their components can be inconsistent, leaving aspects open to interpretation (Michie et al., 2009). As this threatens intervention fidelity, these authors argue for open access to detailed intervention protocols. However, they also acknowledge that intellectual property rights may prevent this from becoming regular practice. Further research is needed to address these issues of consistency. Indeed, lessons can be learnt from other disciplines which have developed research literature to answer similar questions. For example, exploring Re-aim's extended consort diagram, which was developed to translate research into practice by breaking down key factors at each stage of health implementation (Kessler and Glasgow, 2011), could inform implementation within the context of education.

The science of implementation is pertinent in many areas of the service sector, including education, health and social work. Therefore, when researching the conditions which ensure sustainability, findings are transferable between disciplines. This offers huge opportunities for collaborative working across the different domains and creates the potential for rapid advancement of the science of implementation.

IMPLICATIONS FOR EDUCATIONAL PSYCHOLOGISTS

In an ideal world, whenever a theory is supported, its teachings will be transferred into practice to bring about positive change. However, a challenge faced by EPs is ensuring that the interventions schools adopt are effective. EPs have a role in developing clear and widely available information on how to assess interventions by their evidence base and dissemination capacity. There are cases where this has been done in education (Education Endowment Foundation, 2018) and in health (The US National Cancer Institute, 2018). However, support to ensure evidence-based approaches are always used within education continues to be an ongoing goal (Kelly and Perkins, 2012). Recognition that interventions need to be implemented properly gives EPs the opportunity not only to work in line with these principles but also to build capacity within others across an array of settings.

The role of the EP has moved from casework toward more effective systemic ways of working; therefore, the EP is in an exceptional position to:

- Work in collaboration with schools. Jaycox et al. (2006) describe how working in partnership with schools can be effective. They emphasize the importance of becoming familiar with the school staff, its cultures and context through regular contact.
- Jointly discuss options when selecting interventions, ensuring programmes are based on empirical evidence and meet genuine and not perceived needs.
- Ensure staff readiness before implementation.
- Ensure the implementation is designed effectively within the school context.
- Help to measure and assess implementation.
- Undertake research to enhance our understanding of implementation science.
- Develop implementation standards within local authorities.

- Promote effective practice and raise awareness of implementation science.
- Create implementation steering groups which can ensure that implementation is monitored and evaluated throughout the process, and that integrity is maintained (Dane and Schneider, 1998).

In addition, the EP should be sensitive to individuals' workloads by asking school staff only to perform necessary tasks. Throughout the process of implementation, there is a great deal of ongoing monitoring that must take place. Programme implementers should assess success throughout the implementation period and ensure it by adapting the programme to meet the needs of the setting. Therefore, teachers need to understand the importance of implementation monitoring. Players require motivation to fully incorporate these functions into their workload. Furthermore, it may be difficult for an EP to ensure that fidelity is being maintained by teachers, especially when there are competing job pressures. It is therefore of paramount importance that positive working relationships are maintained and that communication is ongoing. The EP should adopt a flexible and sensitive approach in order to yield the best outcomes.

A threat to any intervention is ignoring the whole system of which the school is a part. An example: a teacher wants to implement new class behavior guidelines. For this to be successful, the class rules must be in line with the school and local authority policies and guidelines. Implementation science encourages us all to look at the wider multilevel influences at play. In addition, core implementation components must fit within the organizational components and other social, economic and political influences (Sullivan et al., 2008). If the relationships between these factors are poor, there is less chance of the intervention being implemented with pure fidelity. Here, again, the EP can play a pivotal role in supporting the school's ability to acknowledge all contextual factors.

Within every organization, there are many layers of staff, policies, systems and barriers. Promoting positive change therefore requires a multifaceted approach. If a teacher believes that an intervention is beneficial, they will be more likely to implement it with fidelity (Datnow and Castellano, 2000; Waugh, 2000). Therefore, teachers who have previous experience of an evidence-based intervention which was implemented poorly, thereby yielding disadvantageous outcomes, are unlikely to be motivated to implement the same intervention successfully. Due to these human belief systems, poor implementation could therefore impact on future implementation potential. In such cases the EP may have to sensitively challenge the beliefs that have led to evidence-based programmes being perceived as ineffective. Schools and EPs should work together to design and evaluate initiatives by properly adhering to implementation guidance. Only then is there the best chance of supporting positive change and having maximum impact on the lives of children and families.

Finally, EPs are researchers and have much to offer the study of implementation science. Understanding the fundamentals of this approach and supporting other researchers offer additional opportunities to bring about positive change.

CONCLUSIONS

Implementation science is a universal strategy to ensure that programmes make sustainable positive differences. It acknowledges the systems in place, which interact with each other, and has the potential to significantly improve outcomes for individuals everywhere. Implementation science needs to be incorporated into the design and evaluation of every school programme to ensure effectiveness and sustainability. There are many challenges evident, and players should concentrate on long-term gains rather than short-term fixes to successfully embrace this approach and invest the necessary funding, support and attention. The EP is in an ideal position to support the education system in using these principles and embracing new opportunities of joint working and cross-sector collaboration.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Armenakis, A. A., Harris, S. G., and Mossholder, K. W. (1993). Creating readiness for organisational change. *Hum. Relat.* 46, 681–703. doi: 10.1177/001872679304600601

Axford, N., and Morpeth, L. (2012). "The common language service-development method from strategy development to implementation of evidence based practice," in *Handbook of Implementation Science for Psychology in Education*, eds B. Kelly, and D. Perkins (Cambridge: Cambridge University Press), 443–460.

Barnett, W. (1995). Long-term affects of early childhood interventions on cognitive and school outcomes. *Future Child* 5, 25–50. doi: 10.2307/1602366

Birken, S. A., Powell, B. J., Shea, C. M., Haines, E. R., Kirk, M. A., Leeman, J., et al. (2017). Criteria for selecting implementation science theories and frameworks: results from an international survey. *Implement. Sci.* 12:124. doi: 10.1186/s13012-017-0656-y

Bosworth, K., Garcia, R., Judkins, M., and Saliba, M., (2018). The impact of leadership involvement in enhancing high school climate and reducing bullying: an exploratory study. *J. Sch. Viol.* 17, 354–366. doi: 10.1080/15388220.2017.1376208

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge: Harvard University Press.

CAPS (2012). Available online at: https://incredibleyearsblog.wordpress.com/2015/05/01/children-and-parents-service-caps-early-years-social-and-emotional-wellbeing-awarded-best-practice-congratulations/

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

CASEL (2012). *Collaborative for Academic, Social and Emotional Learning.* Available online at: www.casel.org

CAST (2012). *Incredible Years*. Available online at: http://www.incredibleyears.com/article/cast-study-the-incredible-years-basic-program-in-denmark-in-danish-af-foaeldreprogrammet-basic/

Dane, A., and Schneider, B. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control. *Clin. Psychol. Rev.* 18, 23–45. doi: 10.1016/S0272-7358(97)00043-3

Datnow, A., and Castellano, M. (2000). Teachers' responses to success for all: how beliefs, experiences, and adaptations shape implementation. *Am. Educ. Res. J.* 37, 775–799. doi: 10.3102/00028312037003775

Dunst, C. J., and Trivette, C. M. (2009). Let's be PALS an evidence-based approach to professional development. *Infant Young Child.* 22, 164–176. doi: 10.1097/IYC.0b013e3181abe169

Durlak, J. (2010). The importance of doing well in whatever you do: a commentary on the special section. *Early Child. Res. Q.* 25, 348–357. doi: 10.1016/j.ecresq.2010.03.003

Education Endowment Foundation (2018). Available online at: https://educationendowmentfoundation.org.uk/

Fallon, L. M., Cathcart, S. C., DeFouw, E. R., O'Keeffe, B. V., and Sugai, G. (2018). Promoting teachers' implementation of culturally and contextually relevant class-wide behavior plans. *Psychol. Sch.* 55, 278–294. doi: 10.1002/pits.22107

Fixsen, D. L., Blase, K. A., Timbers, G. D., and Wolf, M. M. (2001). "In search of program implementation: 792 replications of the Teaching-Family Model," in *Offender Rehabilitation in Practice: Implementing and Evaluating Effective Programs*, eds G. A. Bernfeld, D. P. Farrington, and A. W. Leschied (New York, NY: John Wiley & Sons), 149–166.

Fixsen, D. L., Blase, K. A., Naoom, S. F., and Wallace, F. (2009a). Core implementation components. *Res. Soc. Work Pract.* 19:531. doi: 10.1177/1049731509335549

Fixsen, D. L., Blase, K. A., Naoom, S. F., Van Dyke, M., and Wallace, F. (2009b). *Implementation: The Missing Link between Research and Practice. NIRN implementation brief, 1.* Chapel Hill, NC: University of North Carolina at Chapel Hill.

Fixsen, D., Blase, K., Naoom, S., and Duda, M. (2013). *Implementation Drivers: Assessing Best Practices*. Chapel Hill, NC: University of North Carolina at Chapel Hill.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., and Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* Tampa, FL: University of South Florida.

Glasgow, R. E., and Emmons, K. M. (2007). How can we increase translation of research into practice? Types of evidence needed. *Annu. Rev. Public Health* 28, 413–433. doi: 10.1146/annurev.publhealth.28.021406.144145

Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A., and Zins, J. E. (2005). *The Study of Implementation in School-Based Preventive Interventions: Theory, Research and Practice (Unpublished Draft)*. Rockville, MD: U.S. Department of Health and Human Services.

Greenberg, M. T., and Kusche, C. A. (1996). *The PATHS Project: Preventive Intervention for Children.* Final Report to the National Institute of Mental Health, Grant (R01MH42131).

Gregson, J. A., and Sturko, P. A. (2007). Teachers as adult learners: re-conceptualizing professional development. *J. Adult Educ.* 36, 1–18.

Hepburn, H. (2015, July 28) *Teacher Shortages: Act Now Before a 'Full-Blown Crisis' Emerges, Unions Warn.* TES. Available online at: https://www.tes.com/news/teacher-shortages-act-now-full-blown-crisis-emerges-unions-warn

Jaycox, L. H., McCaffrey, D. F., and Ocampo, B. W. (2006). Challenges in the evaluation and implementation of school-based prevention and intervention programs on sensitive topics. *Am. J. Eval.* 27, 320–336. doi: 10.1177/1098214006291010

Johnson, K. W., Ogilvie, K. A., Collins, D. A., Shamblen, S. R., Dirks, L. G., Ringwalt, C. L., et al. (2010). Studying implementation quality of a school-based prevention curriculum in frontier Alaska: application of video-recorded observations and expert panel judgment. *Prevent. Sci.* 11, 275–286. doi: 10.1007/s11121-010-0174-5

Joyce, B., and Showers, B. (2002). "Student Achievement through Staff Development," in *Designing Training and Peer Coaching: Out Needs for learning*, eds B. Joyce and B. Showers (Virginia: National College for School Leadership), 1–5.

Kam, C., Greenberg, M., and Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the paths curriculum. *Prev. Sci.* 4, 55–63. doi: 10.1023/A:1021786811186

Kelly, B. (2012). *Evidence Based in Service Training Methods and Approaches. MSc Educational Psychology Tutorial.* Glasgow: University of Strathclyde.

Kelly, B., Edgerton, C., Robertson, E., and Neil, D. (2012). "The Preschool PATHS curriculum: using implementation science to increase effectiveness," in *SDEP Conference* (Edinburgh: Herriott Watt University).

Kelly, B., and Perkins, D. F. (2012). *Handbook of Implementation Science for Psychology in Education*. Cambridge: Cambridge University Press.

Kessler, R., and Glasgow, R. E. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. *Am. J. Prev. Med.* 40, 637–644. doi: 10.1016/j.amepre.2011.02.023

Kinsman, J., Nakiyingi, J., Kamali, A., Carpenter, L., Quigley, M., Pool, R. et al., (2001). Evaluation of a comprehensive school-based AIDS education programme in rural Masaka, Uganda. *Health Educ. Res.* 16, 85–100. doi: 10.1093/her/16.1.85

Klimes-Dougan, B., August, G., Lee, C.-Y. S., Realmuto, G. M., Bloomquist, M. L., Horowitz, J. L. et al., (2009). Practitioner and site characteristics that relate to fidelity of implementation: the early risers prevention program in a going-to-scale intervention trial. *Prof. Psychol. Res. Pract.* 40, 467–475. doi: 10.1037/a0014623

Kotter, J. P. (1996). *Leading Change*. Boston, MA: Harvard Business.

Lewin, K. (1951). *Field theory in Social Science: Selected Theoretical Papers*. New York, NY: Harper.

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.* 13:5. doi: 10.1186/s13012-017-0705-6

Maher, E. J., Jackson, L. J., Pecora, P. J., Schltz, D. J., Chandra, A., and Barnes-Proby, D. S. (2009). Overcoming challenges to implementing and evaluating evidence -based interventions in child welfare: a matter of necessity. *Child. Youth Serv. Rev.* 31, 555–562. doi: 10.1016/j.childyouth.2008.10.013

Meyers, D. C., Durlak, J. A., and Wandersman, A., (2012). The quality implementation framework: a synthesis of critical steps in the implementation process. *Am. J. Community Psychol.* 50, 462–480. doi: 10.1007/s10464-012-9522-x

Michie, S., Fixsen, D., Grimshaw, J. M., and Eccles, M. (2009). Specifying and reporting in complex behaviour change interventions: the need for a scientific method. *Implement. Sci.* 4:40. doi: 10.1186/1748-5908-4-40

Michie, S., Stralen, M. M., and West, R. (2011). The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement. Sci.* 6:42. doi: 10.1186/1748-5908-6-42

Mihalic, S. (2012). *Blueprints.* Available online at: http://www.colorado.edu/cspv/blueprints/Fidelity.pdf

Mihalic, S., Irwin, K., Elliott, D., Fagan, A., and Hansen, D. (2001). *Blueprints for Violence Prevention*. Juvenile Justice Bulletin.

Mintra, D. (2012). "Increasing student voice in school reform: Building Partnerships, Improving Outcomes," in *Handbook of Implementation Science for Psychology in Education,* eds B. Kelly, and D. Perkins (Cambridge: Cambridge University Press), 21.

Ogden, F., Bjornebekk, G., Kjobli, J., Christiansen, T., Taraldsen, K., and Tollefsen, N. (2012). Measurement of implementation components ten years after a nationwide introduction of empirically supported programs- a pilot study. *Implement. Sci.* 7:49. doi: 10.1186/1748-5908-7-49

Pettigrew, J., Miller-Day, M., Shin, Y., Hecht, M. L., Krieger, J. L., and Graham, J. W. (2013). Describing teacher–student interactions: a qualitative assessment of teacher implementation of the 7th grade keepin'it REAL substance use intervention. *Am. J. Community Psychol.* 51, 43–56. doi: 10.1007/s10464-012-9539-1

Rabin, B. A., Glasgow, R. E., Kerner, J. F., Klump, M. P., and Brownson, R. C., (2010). Dissemination and implementation research on community-based cancer prevention: a systematic review. *Am. J. Prev. Med.* 38, 443–456. doi: 10.1016/j.amepre.2009.12.035

Reinke, W. M., Herman, K. C., and Newcorner, L. L. (2011). The incredible years teacher classroom management training: the methods and principles that support fidelity of training delivery. *Sch. Psych. Rev.* 40, 509–529.

Ringwalt, C. L., Ennett, S., Johnston, T., Rohrbach, L. A., Simons-Rudolf, A., Vincus, A., et al. (2003). Factors associated with fidelity to substance abuse prevention curriculum guides in the nation's middle schools. *Health Educ. Behav.* 30, 375–391. doi: 10.1177/1090198103030003010

Rojas-Andrade, R., and Bahamondes, L. L. (2018). Is implementation fidelity important? a systematic review on school-based mental health programs. *Contemp. Sch. Psychol.* 18, 1–12. doi: 10.1007/s40688-018-0175-0

Scheirer, M. A. (2013). Linking sustainability research to intervention types. *Am. J. Public Health* 103, e73–e80. doi: 10.2105/AJPH.2012.300976

Senge, P. M. (1990). *The Fifth Discipline: The Art and Practice of the Learning Organisation*. New York, NY: Doubleday Currency.

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

Spoth, R., Guyll, M., Lillehoj, C., and Redmond, C. (2007). Prosper study of evidence based intervention implementation quality by community- university partnerships. *Nat. Inst. Health* 35, 981–999. doi: 10.1002/jcop.20207

Sullivan, G., Blevins, D., and Kauth, M. R. (2008). Translating clinical training into practice in complex mental health systems: towards opening the "Black Box" of implementation. *Implement. Sci*. 3:33. doi: 10.1186/1748-5908-3-33

Tabak, R. G., Khoong, E. C., Chambers, D. A., and Brownson, R. C. (2012). Bridging research and practice: models for dissemination and implementation research. *Am. J. Prev. Med*. 43, 337–350. doi: 10.1016/j.amepre.2012.05.024

The US National Cancer Institute (2018). *Research-Tested Intervention Programs (RTIPs)*. Available online at: https://rtips.cancer.gov/rtips/searchResults.doon11/06/18

Topping, K. J., Thurston, A., Tolmie, A., Christie, D., Murray, P., and Karagiannidou, E. (2011). Cooperative learning in science: intervention in the secondary school. *Res. Sci. Technol. Educ*. 29, 91–106. doi: 10.1080/02635143.2010.539972

Waugh, R. F. (2000). Towards a model of teacher receptivity to planned system-wide educational change in a centrally controlled system. *J. Educ Adm*. 38, 350–367. doi: 10.1108/09578230010373615

Webser-Stratton, C. (2012). *Incredible Years*. Available online at: http://www.incredibleyears.com/

Weiner, B. (2009). A theory of organisational readiness for change. *Implement. Sci*. 6, 1–9. doi: 10.1186/1748-5908-4-67

Conflict of Interest Statement: 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.

frontiers in Education

METHODS
published: 25 June 2019
doi: 10.3389/feduc.2019.00053

Linearity vs. Circularity? On Some Common Misconceptions on the Differences in the Research Process in Qualitative and Quantitative Research

*Nina Baur**

Department of Sociology, Technische Universität Berlin, Berlin, Germany

Methodological discussions often oversimplify by distinguishing between "the" quantitative and "the" qualitative paradigm and by arguing that quantitative research processes are organized in a linear, deductive way while qualitative research processes are organized in a circular and inductive way. When comparing two selected quantitative traditions (survey research and big data research) with three qualitative research traditions (qualitative content analysis, grounded theory and social-science hermeneutics), a much more complex picture is revealed: The only differentiation that can be upheld is how "objectivity" and "intersubjectivity" are defined. In contrast, all research traditions agree that partiality is endangering intersubjectivity and objectivity. Countermeasures are self-reflexion and transforming partiality into perspectivity by using social theory. Each research tradition suggests further countermeasures such as falsification, triangulation, parallel coding, theoretical sensitivity or interpretation groups. When looking at the overall organization of the research process, the distinction between qualitative and quantitative research cannot be upheld. Neither is there a continuum between quantitative research, content analysis, grounded theory and social-science hermeneutics. Rather, grounded theory starts inductively and with a general research question at the beginning of analysis which is focused during selective coding. The later research process is organized in a circular way, making strong use of theoretical sampling. All other traditions start research deductively and formulate the research question as precisely as possible at the beginning of the analysis and then organize the overall research process in a linear way. In contrast, data analysis is organized in a circular way. One consequence of this paper is that mixing and combining qualitative and quantitative methods becomes both easier (because the distinction is not as grand as it seems at first sight) and more difficult (because some tricky issues of mixing specific to mixing specific types of methods are usually not addressed in mixed methods discourse).

Keywords: research process, mixed methods, survey research, big data, qualitative content analysis, grounded theory, social-science hermeneutics, objectivity

OPEN ACCESS

Edited by:
Douglas F. Kauffman, Medical University of the Americas–Nevis, United States

Reviewed by:
Jana Uher, University of Greenwich, United Kingdom
Barbara Hanfstingl, Alpen-Adria-Universität Klagenfurt, Austria

***Correspondence:**
Nina Baur
nina.baur@tu-berlin.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *30 June 2018*
Accepted: *22 May 2019*
Published: *25 June 2019*

Citation:
Baur N (2019) Linearity vs. Circularity? On Some Common Misconceptions on the Differences in the Research Process in Qualitative and Quantitative Research. Front. Educ. 4:53. doi: 10.3389/feduc.2019.00053

INTRODUCION

Since the 1920s, two distinct traditions of doing social science research have developed and consolidated (Kelle, 2008, p. 26 ff.; Baur et al., 2017, p. 3; Reichertz, 2019), which are typically depicted as the "qualitative" and the "quantitative" paradigm (Bryman, 1988). Both paradigms have a long tradition of demarcating themselves from each other by ignoring each other at best or criticizing as well as pejoratively devaluating the respective "other" tradition at worst (Baur et al., 2017, 2018, pp. 8–9; Kelle, 2017; Baur and Knoblauch, 2018). Regardless, few authors make the effort of actually defining the difference between the paradigms. Instead, most methodological texts in both research traditions make implicit assumptions about the properties of "qualitative" and "quantitative" research. If one sums up both these (a) implicit assumptions and (b) the few attempts of defining what "qualitative" and "quantitative" research is, the result is a rather crude and oversimplified picture.

"*Qualitative research*" is typically depicted as combination of the following elements (Ametowobla et al., 2017, pp. 737–776; Baur and Blasius, 2019):

- an "interpretative" epistemological stance (e.g., Knoblauch et al., 2018) which is associated e.g., with phenomenology or social constructivism (Knoblauch and Pfadenhauer, 2018) or some branches of pragmatism (Johnson et al., 2017);
- a research process that is circular or spiral (Strübing, 2014);
- single case studies (Baur and Lamnek, 2017a) or small theoretically and purposely drawn samples meaning that relatively few cases are analyzed (Behnke et al., 2010, pp. 194–210);
- for these cases, a lot of data are collected, e.g., by qualitative interviews (Helfferich, 2019), ethnography (Knoblauch and Vollmer, 2019) or so-called "natural" data, i.e., qualitative process-produced data such as visual data (Rose, 2016) or digital data such as web videos (Traue and Schünzel, 2019), websites (Schünzel and Traue, 2019) or blogs (Schmidt, 2019). In all these cases, this means that a lot of information per case is analyzed;
- both the data and the data collection process are open-ended and less structured than in quantitative research;
- data are typically prepared and organized either by hand or by using qualitative data analysis software (such as NVivo, MAXqda or Atlas/ti);
- data analysis procedures themselves are suitable for the more unstructured nature of the data.

In contrast, "*quantitative research*" is seen as a combination of (Ametowobla et al., 2017, pp. 752–754; Baur and Blasius, 2019):

- a "positivist" research stance;
- a linear research process (Baur, 2009a);
- large random samples meaning that many cases are analyzed;
- relatively little information per case, collected via (in comparison to qualitative research) few variables;
- data are collected in a highly structured format, e.g., using surveys (Groves et al., 2009; Blasius and Thiessen, 2012; Baur, 2014) or mass data (Baur, 2009a) which recently have also been called "big data" (Foster et al., 2017; König et al., 2018) and which may comprise e.g., webserver logs and log files (Schmitz and Yanenko, 2019), quantified user-generated information on the internet such as Twitter communication (Mayerl and Faas, 2019) as well as public administrational data (Baur, 2009b; Hartmann and Lengerer, 2019; Salheiser, 2019) and other social bookkeeping data (Baur, 2009a);
- the whole data collection process is highly structured and as standardized as possible;
- data are prepared by building a data base and analyzed using statistical packages (like R, STATA or SPSS) or advanced programming techniques (e.g., Python);
- data are analyzed using diverse statistical (Baur and Lamnek, 2017b) or text mining techniques (Riebling, 2017).

Once these supposed differences are spelled out, it immediately becomes obvious how oversimplified they are because in social science research *practice*, the distinction between the data types is much more fluent. For example, "big data" are usually mixed data, containing both standardized elements (Mayerl and Faas, 2019) such as log files (Schmitz and Yanenko, 2019) and qualitative elements such as texts (Nam, 2019) or videos (Traue and Schünzel, 2019). Accordingly, it is unclear, if text mining is really a "quantitative" method or rather a "qualitative" method. While the fluidity between "qualitative" and "quantitative" research becomes immediately obvious in big data analyses, this issue has also been lingering in "traditional" social science research for decades. For example, many quantitative researchers simultaneously analyse several thousand variables. Survey research has a long tradition of using qualitative methods for pretesting and evaluating survey questions (Langfeldt and Goltz, 2017; Uher, 2018). Almost all questionnaires contain open-ended questions with non-standardized answers which have to be coded afterwards (Züll and Menold, 2019), and if interviewees or interviewers do not agree with the questionnaire, they might add comments on the side—so-called marginalia (Edwards et al., 2017). During data analysis, results of statistical analyses are often "qualified" when interpreting results. While Kuckartz (2017) provides many current examples for qualification of quantitative data, a well-known older example is Pierre Bourdieu's analysis of social space by using correspondence analysis. Likewise, qualitative research has a long tradition of "quantification" of research results (Vogl, 2017), and similarly to text mining, it is unclear, if qualitative content analysis is a "quantitative" method or rather a "qualitative" method.

Despite these obvious overlaps and fluent borders between "qualitative" and "quantitative" research, the oversimplified view of two different "worlds" or "cultures" (Reichertz, 2019) of social science research practice is upheld in methodological discourse. Accordingly, methodological discourse has reacted increasingly by attempting to combine these traditions via *mixed methods research* since the early 1980s (Baur et al., 2017). However, although today many differentiated suggestions exist how to best organize a mixed methods research process (Schoonenboom and Johnson, 2017), mixed methods research in a way consolidates this simple distinction between "qualitative" and "quantitative"

research, as in all attempts of mixing methods, qualitative and quantitative methods still seem distinct methods—which is exactly why it is assumed that they need to be "mixed." Moreover, many qualitative researchers complain that current suggestions for mixing methods ignore important principles of qualitative research and instead enforce the quantitative research logic on qualitative research processes, thus robbing qualitative research of its hugest advantages and transforming it into a lacking version of quantitative research (Baur et al., 2017, for some problems arising when trying to take qualitative research logics seriously in mixed methods research, see Akremi, 2017; Baur and Hering, 2017; Hense, 2017).

In this paper, I will address this criticism by focusing on *social science research design and the organization of the research process*. I will show that the distinction between "qualitative" and "quantitative" research is oversimplified. I will do this by breaking up the debate about "the" qualitative and "the" quantitative research process up in two ways:

Firstly, if one looks closely, *there is not "one way" of doing qualitative or quantitative research*. Instead, in both research traditions, there are sub-schools, which are characterized by the same degree of ignoring themselves or infighting as can be observed between the qualitative and quantitative tradition.

- More specifically, "*quantitative research*" can be at least differentiated into classical survey research (Groves et al., 2009; Blasius and Thiessen, 2012; Baur, 2014) and big data analysis of process-generated mass data ("Massendaten") (Baur, 2009a). Survey data are a good example for research-elicited data, meaning that data are produced by researchers solely for research purposes which is why researchers (at least in theory) can control every step of the research process and therefore also the types of errors that occur. In contrast, process-produced mass data are not produced for research purposes but are a side product of social processes (Baur, 2009a). A classic example for process-produced mass data are public administrational data which are produced by governments, public administrations, companies and other organizations in order to conduct their everyday business (Baur, 2009b; Hartmann and Lengerer, 2019; Salheiser, 2019). For example, governments collect census data for planning purposes; pension funds collect data on their customers in order to assess who later has acquired which types of claims; companies collect data on their customers in order to send them bills etc. Digital data (Foster et al., 2017; König et al., 2018), too, are typically side-products of social processes and therefore count as process-produced data. For example, each time we access the internet, log files are created that protocol our internet activities (Schmitz and Yanenko, 2019), and in many social media, users will leave quantified information—a typical example is Twitter communication (Mayerl and Faas, 2019). Process-produced data can also be analyzed by researchers. In contrast to survey data, they have the advantage of being non-reactive, and for many research questions (e.g., in economic sociology) they are the only data type available (Baur, 2011). However, as they are not produced for research purposes, researchers cannot control the research process or types of errors that may occur during data collection—researchers can only assess how the data are biased before analyzing them (Baur, 2009a). Regardless of researchers using research-elicited or process-produced data, many quantitative researchers aim at replicating results in order to test, if earlier research can uphold scrutiny[1]. Therefore, one can distinguish between primary research (the original study conducted by the first researcher), replication (when a second researcher tries to produce the same results with the same or different data) and meta-analysis (where a researcher compares all results of various studies on a specific topic in order to summarize findings, see Weiß and Wagner, 2019). In contrast, for secondary analysis, researchers re-use an existing data set in order to answer a different research question than the primary researcher asked. As can be seen from this short overview, there are many diverging research traditions within quantitative research, and accordingly, there are many differences and unresolved issues between these traditions. However, for the purpose of this paper, I will subsume them under the term "quantitative research", as I have shown in Baur (2009a) that at least regarding the overall organization of research processes, these various schools of quantitative research largely resemble each other.
- The situation is not as simple for "*qualitative research*": Not only are there more than 50 traditions of qualitative research (Kuckartz, 2010), but these traditions widely diverge in their epistemological assumptions and the way they do research. In order to be able to better discuss these differences and commonalties, in this paper, I will focus on three qualitative research traditions, which have been selected for being as different as possible in the way they organize the research process, namely "*qualitative content analysis*" (Schreier, 2012; Kuckartz, 2014, 2018; Mayring, 2014; see also Ametowobla et al., 2017, pp. 776–786), "*social-science hermeneutics*" ("sozialwissenschaftliche Hermeneutik"), which is sometimes also called "hermeneutical sociology of knowledge" ("hermeneutische Wissenssoziologie") (Reichertz, 2004a; Herbrik, 2018; Kurt and Herbrik, 2019; see also Ametowobla et al., 2017, pp. 786–790) and "*grounded theory*" (Corbin and Strauss, 1990; Strauss and Corbin, 1990; Clarke, 2005; Charmaz, 2006; Strübing, 2014, 2018, 2019). Please note that within these traditions, some authors try to combine and integrate these diverse qualitative approaches. However, in order to be able to explore the commonalities and differences better, I will focus on the more "pure," i.e., original forms of these qualitative paradigms.

Secondly, while it is not possible of speaking of "the" qualitative and "the" quantitative research, *it is neither possible of speaking of "the" research process in the sense that there is only one question to be asked when designing social inquiry*. Instead, when it comes to discussing the differences between qualitative and quantitative research, at least six *issues* have to be discussed:

[1]Note that for many social phenomena, replication is not possible due to the nature of the research object, e.g. for macro-social or fast-changing social phenomena (Kelle, 2017) – see below for more details.

1. How is researchers' perspectivity handled during the research process?
2. How can intersubjectivity be achieved, and what does "objectivity" mean in this context?
3. When and how is the research question focused?
4. Does the research process start deductively or inductively?
5. Is the order of the diverse research phases (sampling, data collection, data preparation, data analysis) organized in a linear or circular way?
6. Is data analysis itself organized in a linear or circular way?

In the following sections, I will discuss for each of these six issues how the four research traditions (quantitative research, qualitative content analysis, grounded theory, social-science hermeneutics) handle them and how they resemble and differ from each other. I will conclude the paper by discussing what this means for the distinction between qualitative and quantitative research as well as mixed methods research.

HANDLING PERSPECTIVITY BY USING SOCIAL THEORY

There are many different types of philosophies of sciences and associated epistemologies, e.g., pragmatism (Johnson et al., 2017), phenomenology (Meidl, 2009, pp. 51–98), critical rationalism (Popper, 1935), critical theory (Adorno, 1962/1969/1993; Habermas, 1981), radical constructivism (von Glasersfeld, 1994), relationism (Kuhn, 1962), postmodernism (Lyotard, 1979/2009), anarchism (Feyerabend, 1975), epistemological historism (Hübner, 2002), fallibism (Lakatos, 1976) or evolutionary epistemology (Riedl, 1985). Moreover, debates within these different schools of thought are often rather refined and organized in sub-schools, as Johnson et al. (2017) illustrate for pragmatism. Regardless, current social science debates simply crudely distinguish between "positivism" and "constructivism". While this is yet another oversimplification which would be worth deconstructing, for the context of this paper it suffices to note that this distinction is rooted in the demarcation between the natural sciences ("Naturwissenschaften") and humanities ("Geisteswissenschaften") in the nineteenth century. It has been the occasion of several debates on the nature of (social) science as well as the methodological and epistemological consequences to be drawn from this definition of (social) science (e.g., Merton, 1942/1973; Smelser, 1976; for an overview see Baur et al., 2018).

In current social science debates, the "quantitative paradigm" is often depicted as being "positivist", while the "qualitative paradigm" is depicted as being "constructivist" or "interpretative" (e.g., Bryman, 1988) which has consequences on how we conceive social science research processes.

One of the issues debated is, whether social reality can be grasped "per se" as a fact. This so-called "positive stance" was taken e.g., by eighteenth and nineteenth century cameralistics and statistics who collected census and other public administrational data in order to improve governing practices and competition between nation states and who strongly believed that their statistical categories were exact images of social reality (Baur et al., 2018). This "positive stance" was also taken e.g., by the representatives of the German School of History who claimed that facts should speak for themselves and focused on a history of events ("Ereignisgeschichte") (Baur, 2005 pp. 25–56).

The criticism of these research practices of *both* the natural sciences (exemplified by early statistics) and the humanities (exemplified by historical research) goes back to the nineteenth century. For example, early German-language sociologists such as Max Weber criticized *both* traditions because they argued that no "facts" exist that speak for themselves, as both the original data producers of sociological or historical data and the researchers using these data see them from a specific perspective and subjectively (re-)interpret them. In other words: Data are highly constructed. If researchers do not reflect this construction process, they unconsciously (re-)produce their own and the data producer's worldview. As in the nineteenth century, both statistical data and historical documents were mostly originally produced by or for the powerful, nineteenth century statistics and humanities unconsciously analyzed society from their own perspective and the perspective of the powerful (Baur, 2008, p. 192). Consequently, early historical science served to politically legitimate historically evolved orders (Wehler, 1980, p. 8, 44, 53–54).

These arguments are reflected in current debates, e.g., by the debates on how social-science methodology in general and statistics in particular are tools of power (e.g., Desrosières, 2002). They are also reflected in postmodern critiques that every research takes place from a specific worldview ("Standortgebundenheit der Forschung"), which is a particular problem for social science research, as researchers are always also part of the social realities they analyze, meaning that their particular subjectivity may distort research. More specifically, as academia today is dominated by white middle-class men from the Global North, social science research is systematically in danger of creating blind spots for other social realities (Connell, 2007; Mignolo, 2011; Shih and Lionnet, 2011)—an issue Merton (1942/1973) had already pointed out.

At the same time, it does not make sense to dissolve social science research in extreme "constructivism", as this will make it impossible to assess the validity of research and to distinguish between solid research and "fake news" or "alternative facts" (Baur and Knoblauch, 2018).

In other words: The distinction between "positivism" and "constructivism" creates a dilemma between either denying the existence of different worldviews or abolishing the standards of good scientific practice. In order to avoid this deadlock, early German sociologists (e.g., Max Weber) and later generations of historians reframed this question: The problem is not, *if* subjectivity influences perception (it does!), but *how* it frames perception (Baur, 2005, 2008; Baur et al., 2018). In other words, one can distinguish between *different types of subjectivity*, which have different effects on the research process. In modern historical sciences, at least three forms of subjectivity are distinguished (Koselleck, 1977):

1. *Partiality ("Parteilichkeit")*: As shown above, subjectivity can distort research because researchers are so entangled in their

own value system that they systematically misinterpret or even peculate data. This kind of subjectivity has to be avoided at all costs.
2. "*Verstehen*": Subjectivity is necessary to understand the meaning of human action (and data in general), so in this sense, it is an important resource for social science research, especially in social-science hermeneutics.
3. *Perspectivity ("Perspektivität")*: Subjectivity is also a prerequisite for grasping reality. The first important steps in social science research are framing a research question as "relevant" and "interesting", addressing this question from a certain theoretical stance and selecting data appropriate for answering that question.

Starting from this distinction, early German sociologists argued that—as one cannot avoid perspectivity—it is important to reflect it and make it explicit. And one does this by making strong use of social theory and methodology when designing and conducting social science research (Baur, 2008, pp. 192–193). The point about this is that social science research still creates blind spots (because reality can never be analyzed as a whole) but as these blind spots are made explicit, they become debatable and can be openly addressed in future research.

If one reframes the question, the debate between "positivism" and "constructivism" implodes, as the comparison of the four research traditions reviewed in this paper illustrates: Quantitative research, qualitative content analysis, grounded theory and social-science hermeneutics all make a strong argument that *social theory* is absolutely necessary for guiding the research process[2]. In order to establish how social theory and empirical research should be linked, one first has to define what "social theory" actually is (Kalthoff, 2008). This is important as theories differ in their level of abstraction and at least three types of theories can be distinguished (Lindemann, 2008; Baur, 2009c; Baur and Ernst, 2011):

1. *Social Theories* ("*Sozialtheorien*"), such as analytical sociology, systems theory, communicative constructivism, actor network theory or figurational sociology, contain general concepts about what society is, which concepts are central to analysis (e.g., actions, interactions, communication), what the nature of reality is, what assumptions have to be made in order to grasp this reality and how—on this basis—theory and data can be linked on a general level.
2. *Middle-range theories* ("*Theorien begrenzter Reichweite*") concentrate on a specific thematic field, a historical period and a geographical region. They model social processes just for this socio-historical context. For example, Esping-Andersen's (1990) model of welfare regimes argues that there have been typical patterns of welfare development in Western European and Northern American societies since about the 1880s. In contrast, in their study "Awareness of Dying", Glaser and Strauss (1975), address topics of medical sociology and claim to have identified typical patterns that are valid for the U.S. in the 1960s and 1970s.
3. *Theories of Society* ("*Gesellschaftstheorien*") try to characterize complete societies by integrating results from various studies to a larger theoretical picture, e.g., "Capitalism", "Functionally Differentiated Society," "Modernity," and "Postmodernity." In other words, theories of society build on middle-range theories and further abstract them. Middle-range theories and theories of society are closely entwined as an analysis of social reality demands "a permanent control of empirical studies by theory and vice versa a permanent review of these theories via empirical results" (Elias, 1987, p. 63). For example, in figurational sociology, the objective is to focus and advance sociological hypotheses and syntheses of isolated findings for the development of a "theory of the increasing social differentiation" (Elias, 1997, p. 369), of planned and unplanned social processes, and of integration and functional differentiation (Baur and Ernst, 2011).

These types of theories are entwined in a very typical way during the research process. Namely, all social science methodologies are constructed in a way that social theory is used to build, test and advance middle-range theories and theories of society (Lindemann, 2008; Baur, 2009c). Therefore, social theory is a prerequisite for social research as it helps researchers decide which data they need and which analysis procedure is appropriate for answering their research question (Baur, 2005, 2008). Social theory also allows researchers to link middle-range theories and theories of society with both methodology and research practice, as not all theories can make use of all research methods and data types (Baur, 2008). For example, rational choice theory needs data on individuals' thoughts and behavior, symbolic interactionism needs data on interactions, i.e., what is going on between individuals.

Due to the importance given to social theory, it is unsurprising that all research traditions stress that the theoretical perspective needs to be disclosed by *explicating the study's social theoretical frame* and defining central terms and terminology at the beginning of the research process (Weil et al., 2008). The dispute between the four methodological traditions discussed in this paper is whether one needs to have a specific *middle-range theory* in mind at the beginning of the research process or not. In quantitative research, specifying one or more middle-range theories in advance is necessary in order to formulate hypotheses to be tested. The opposing point of view is that of grounded theory which explicitly aims at developing new middle-range theories for new research topics and therefore by nature cannot have any middle-range theory in mind at the beginning of the research process. Qualitative content analysis and social-science hermeneutics are someway in between these extreme positions.

All in all, explicating one's social theoretical stance is a major measure against partiality, as assumptions are explicated and thus can be criticized. All research traditions analyzed for this paper also agree on a second measure against partiality: *self-reflection*. In addition, each research tradition has developed *distinct methodologies in order to handle subjectivity and perspectivity*, i.e., in order to avoid partiality crawling back in via the backdoor.

[2]To clarify a common misconception of qualitative research: When qualitative researchers demand that research should be 'open-ended' ('Offenheit'), they do not mean that they are not using theory but that they are using an inductive analytical stance (see below).

In the tradition of critical rationalism (Popper, 1935), *quantitative research* systematically aims at falsification. Ideally, *different middle-range theories and hypotheses compete and are tested against each other*. For example, survey methodology typically tries to test middle-range theories, meaning that at the beginning of the research process not only the general social theory but also middle-range theories must be known and clarified as well as possible. Then, these theories typically are formulated into hypotheses, which then are operationalized and can be falsified during the research process. This idea of testing theories can be seen in two typical ways of doing so:

(a) One can use statistical tests to falsify hypotheses.
(b) The other way of testing theories in quantitative research is using different middle-range theories and see which theory fits the data best.

Qualitative research has repeatedly stated that the point about qualitative research is that very often, no middle range theory exists (which would be needed for testing theories), and the aim of qualitative research is exactly to build these middle range theories. Therefore, in most qualitative research processes, testing theories with data is not a workable solution. Instead, qualitative research has suggested "*triangulation*" of theories, methods, researchers and data (Flick, 1992, 2017; Seale, 1999, pp. 51–72) as an alternative. Note that the idea of triangulating theories in qualitative research is very similar to the idea of testing theories against each other in quantitative research. But in contrast to quantitative research, in qualitative research theories are not necessarily spelled out in advance but instead are built during data analysis. Note also, that the idea that different researchers address the same problem with different data is also equal to the way quantitative research is organized in practice: In the ideal quantitative research process, each single study is just a small pebble in the overall mosaic, to be published e.g., in an academic paper. Then other researchers (e.g., from different institutions) can use other data (e.g., from a different data set) and see how well these fit the theory, i.e., they try to replicate results of the primary study, and the results of various replications can be then summarized in a meta-analysis (Weiß and Wagner, 2019).

In addition, *qualitative content analysis* has a strong tradition of handling researchers' perspectivity, not only by triangulating them but in fact by using different researchers to *code in parallel* e.g. the interview data and then comparing these codings. This procedure works better, the more dissimilar the researchers are concerning disciplinary, theoretical, methodological, political and socio-structural background, as contrasting researchers likely have also very different perspectives on the topic. If two such researchers independently coded the same text passages similarly, hopefully perspectivity can be ruled out. In contrast, if the same passages are coded differently by two persons, then one must interpret and take a closer look on how researchers' subjectivity and perspectivity might have influenced the coding process. All in all, qualitative content analysis makes use of research teams in a way that researchers first work independently and then results are compared. Note that this is similar to the way modern survey research works in practice: Here, questionnaires are typically developed and tested in teams, following the concept of the "Survey Life Cycle" (Groves et al., 2009), and a main means of evaluating survey questions is expert validation by other, external researchers. Similar, during data analysis, it is typical for quantitative researchers to re-analyze data that have already been analyzed by other teams. This is one of the reasons why archiving and documenting data is good-practice in survey methodology.

Social-science hermeneutics, too, have a strong tradition of researchers working together, but this co-operation and reciprocal control is organized in a different way: In contrast to qualitative content analysis and survey research (where researchers first work independently and then results are compared), the order of co-operation and independent research is reversed in social-science hermeneutics: The research team is used at the beginning of data analysis in so-called "*data sessions*" ("Datensitzungen") (Reichertz, 2018). The team focusses on one section of the text and does a so-called "fine-grained analysis" ("Feinanalyse"). During these data sessions, the research teams collectively develops different interpretations or "readings" ("Lesarten") (Kurt and Herbrik, 2019). In fact, these interpretations resemble hypothesis formation in quantitative research, and the following analysis steps also strongly resemble quantitative research, as after the data session, researchers can individually or collectively test the hypotheses (= interpretations). However, in hermeneutics, interpretations are not tested using statistical tests but using "sequential analysis" ("Sequenzanalyse"). During sequential analysis, the text is used as material for testing the hypotheses developed during data sessions: If an interpretation holds true, there should be other hints in the data that point to that interpretation, while other interpretations might be falsified by additional data (Lamnek, 2005, pp. 211–230, 531–546; Kurt and Herbrik, 2019).

Grounded theory handles subjectivity differently in so far that it has developed different procedures for theoretically grounding the research process and for building *theoretical sensitivity* (Strauss and Corbin, 1990). The starting point of discussions on theoretical sensitivity is that researchers—being human—cannot help but entering the field not only with their social theoretical perspective but also with their everyday knowledge and prejudices which may bias both their observations and their interpretations. This may also mislead researchers to gloss over inconsistencies or interesting points in their data too fast. Note that the problem formulated here is very similar to the idea of the "investigator bias" in experimental research. In order to tackle this tendency for misinterpretation, grounded theory states that researchers need to develop theoretical sensitivity, i.e., "to enter the research setting with as few predetermined ideas as possible (…). In this posture, the analyst is able to remain sensitive to the data by being able to record events and detect happenings without first having them filtered through and squared with pre-existing hypotheses and biases" (Glaser, 1978, p. 2–3). In order to develop and uphold this open-mindedness for new ideas, Strauss and Corbin (1990) suggest a number of specific procedures such as systematically asking questions or analyzing words, phrases and sentences. Grounded theory also suggests many ways of systematic comparisons such as "flip flop techniques," "systematic comparison" and so on. Another modus operandi suggested is "raising the red flag." These techniques for increasing theoretical

sensitivity are meant as procedures that can be used if researchers are working on their own and do not have a team of coders who can code in parallel.

Summing up the argument so far, all qualitative and quantitative approaches analyzed suggest a *strong use of social theory*. Social theory transforms partiality into perspectivity by focusing on some aspects of (social) reality which then guide the research process. This necessarily creates blind spots. The difference between unreflected subjectivity and social theory is that by using social theory, blind spots are explicated. Consequently, both their assumptions and consequences can be discussed and criticized. All research traditions also have developed *further methodologies in order to handle perspectivity* in research practice. Two common ideas are *working in teams* and *testing ideas* that have been developed in earlier research. Regardless of research tradition, for these suggestions of using research teams for controlling partiality and handling perspectivity to be effective, it is necessary for the team members to be both knowledgeable about the topic and as different as possible concerning biographical experience (e.g., gender, age, disability, ethnicity, social status etc.) and theoretical stance. Note that none of these countermeasures guarantee impartiality—they just help to better handle it.

ACHIEVING INTERSUBJECTIVITY AND MEANING OF OBJECTIVITY

So far, I have argued that one way to address the distinction between "positivism" and "constructivism" is the issue of how to handle a researcher's subjectivity by transforming it into perspectivity, which in turn makes the specific blind spots created by a researcher's theoretical perspective obvious and opens them for theoretical and methodological reflexion. However, "positivism" and "constructivism" do not only address the relation of social science research to a researcher's personal, subjective perspective but also to the issue of how to achieve intersubjectivity—in other words, how to make research as "objective" as possible in order to be able to distinguish more and less valid research. In this sense, perspectivity is closely linked to the concept of "objectivity."

Now, the debate on "objectivity" is a complicated issue, as the comparison of the four research traditions (quantitative research, qualitative content analysis, grounded theory, and social-science hermeneutics) reveals:

First of all, it is not clear at all *what "objectivity" means* in different research traditions. In *quantitative research*, "objectivity" and "intersubjectivity" are used synonymously and mean that independent researchers studying the same social phenomenon always come to the same results, as long as the social phenomenon remains stable. This concept of objectivity has consequences on the typical way quantitative inquiries are designed: The wish to ensure that researchers can actually independently come to the same result is the main reason why quantitative research tries to standardize everything that can be possibly standardized, as can be exemplified by survey research: sampling (random sampling), the measurement instruments (questionnaires), data collection (interviewer training, interview situation) as well as data analysis (statistics) (Baur et al., 2018). The idea is that by standardization, it does not matter who does the research and results become replicable. Qualitative content analysis tries to copy these procedures by techniques such as parallel coding discussed above. Other examples of aiming at getting as close to objectivity as possible are concepts like intercoder reliability.

However, this aim at achieving objectivity by controlling any effect a researcher's subjectivity might have on the research process does not work in practice at all: Despite all attempts of standardization, quantitative researchers have to make many theoretical and practical decisions during the research process and therefore interpret their data and results (Baur et al., 2018). This is true for all stages of the research process, starting from focussing the research question (Baur, 2008) to designing instruments and data collection (Kelle, 2018), data analysis (Akremi, 2018; Baur, 2018) and generalization using inductive statistics (Ziegler, 2018). In this sense, all quantitative research is "interpretative" as well (Knoblauch et al., 2018)—a fact, that is hidden by terminology: While qualitative research talks about "interpretation", quantitative research talks about "error", but this basically means the same, i.e., that regardless how much researchers might try, social reality cannot be "objectively" grasped by researchers. Instead, there is always a gap between what is represented in the data and what is "truly" happening (Baur and Knoblauch, 2018; Baur et al., 2018).

In order to react to this problem, within the quantitative paradigm, survey research has developed the concept of the "Total Survey Error" (TSE) in the last two decades (Groves et al., 2009). The key argument is that various types of errors might occur during the research process, and these various errors are often related in the sense that—if you reduce one error type—another error increases. For example, in order to minimize measurement error, it is typically recommended to ask many different and detailed questions, a classical example being a psychological test for diagnosis of mental disorders, which usually takes more than 1 hour to answer and is very precise. However, these kinds of questionnaires could not be used in surveys of the general population, as respondents typically are only willing to spend a limited time on answering survey questions. If the survey is too long, they will either outright decline to participate in the survey or drop out during processing the survey—which in turn results in unit or item nonresponse. Therefore, researchers typically ask fewer questions in surveys, which increases the likelihood of measurement error. In this example, there is a trade-off between measurement error (due to short questionnaires) and nonresponse error (due to long questionnaires). The other error types are likewise related. Therefore—while traditionally, these various errors were handled individually—modern survey methodology tries to incorporate them into one concept—the "Total Survey Error." This means that researchers should take into account all errors and try to minimize the error as a whole. However, as errors can only be minimized and never completely deleted, logically, there will always remain a gap between "objective reality" and "measurement." In other words, in research practice, survey

methodology has long abolished the idea that it is possible to "objectively" measure reality. Instead, there might always be a difference between what truly happens and what the data convey (Baur, 2014, pp. 260–262).

The discussion about the Total Survey Error already suggests that trying to achieve "objectivity"—in the sense that everybody who does research may achieve the same result, if the research process is well-organized—only works to a specific point. While quantitative research tries to come as close to this goal as possible, there are many fields of social reality, where the attempt of achieving objectivity via standardization faces huge difficulties. Examples are cross-cultural and comparative research (Baur, 2014) and fields characterized by rapid social change (Kelle, 2017) because here concepts are not stable across contexts.

Based on these observations, most *qualitative research* traditions argue that a concept of objectivity in the quantitative sense does not make sense either for qualitative research or not at all—because it is never possible to achieve, as many social contexts are changing very fast and in fact so fast that it is difficult to build middle range theories that can be tested, as the object of research might have already changed before a researcher can replicate the study. In turn, most qualitative researchers define "objectivity" differently. For example, in social-science hermeneutics, "objectivity" means that researchers should reflect, document and explain how they arrived at their conclusions, i.e., how they collected and interpreted data (Lamnek, 2005, pp. 59–77). Seale (1999, pp. 141–158) calls this "reflexive methodological accounting." The idea is that this makes it possible to criticize and validate research. While this is a more basic concept of "objectivity," this is a concept that all researchers (including quantitative researchers) can agree on.

While I have, firstly, shown that "objectivity" can mean very different things, it is also, secondly, *an oversimplification to claim that all quantitative researchers believe that "objectivity" (in the sense of the quantitative paradigm) can be actually achieved and that all qualitative researchers disbelieve this*. As stated above, many quantitative researchers have long-ago given up the idea that there can be a "true," "objective" measurement of reality. Moreover, there are many qualitative researchers who actually believe that "objectivity" (in the sense of the quantitative paradigm) *can* be achieved, and this believe in the possibility of "true measurement" can be found in all research traditions (Baur et al., 2018), e.g., within qualitative content analysis, Mayring (2014) does believe in "objectivity," Kuckartz (2014) does not and takes a more interpretative stance. Similarly, within the hermeneutical tradition, Oevermann et al. (1979) as well as Wernet (2006) believe in "objectivity," Maiwald (2018) and Herbrik (2018) take a more interpretative stance. All in all, the picture is much more complicated than it seems on first sight—something that is definitively worth exploring in more detail in future research.

FOCUSING THE RESEARCH QUESTION

A consequence of the above discussion is that all social science methodology needs to make strong use of social theory in order to guide a researcher's perspective, in other words: Research questions need to be focused and in fact, research becomes better, if and when it is focused, as it allows researchers to consciously collect and analyze exactly the data they need in order to answer the question. As I have shown in Baur (2009c, pp. 197–206), one can use social theory for focusing, and the research question has to be focused at least concerning four dimensions: (1) action sphere, (2) analysis level, (3) spatiality and (4) temporality with the two sub-dimensions (4a) pattern in time and duration. In additions, researchers have to decide, if there are interactions both within and between these dimensions, e.g., between long-term and short-term developments or between space and time. As a rule of thumb, if one wants to explore one of these dimensions in detail, it is advisable to reduce as much complexity for the other dimensions as possible.

While this need for focusing the research question is something that quantitative research, qualitative content analysis, grounded theory and social-science hermeneutics would agree on, they differ on the question, *when the research question is focused*:

Both *quantitative research*, *qualitative content analysis* and *social-science hermeneutics formulate the research question as precisely as possible at the beginning of the analysis* in order to know what types of data need to be collected and analyzed. Quantitative research needs to know e.g., what kinds of questions to ask in a questionnaire, and in order to be able do this, researchers need to know what they want to know. Similarly, social-science hermeneutics need to know which text passage is especially theoretical relevant and thus worthy of being analyzed in the first data session—usually, only a single or a few sentences are selected, and in order to do this selection, researchers, too, need to know what they want to know. While qualitative content analysis usually collects open-ended data such as interview data or documents, a (for qualitative research) relatively large amount of similarly types of e.g., interviews or documents is collected, and data collection is often divided up between team members—again, in order to be fruitful, this is only possible, if researchers, know what they want to know.

In contrast, *grounded theory* opposes this early focusing, arguing that researchers might miss the most innovative or important points of their research, if they focus too early, especially, if they enter a new research field they know very little about. Instead, grounded theory suggests that *researchers should start with a very general research question at the beginning of the analysis, which is focused during data analysis*. In order to enable a focusing grounded in data, grounded theory has developed an own technique for focusing the research process: selective coding (Corbin and Strauss, 1990).

Among other things, this has effects on the way social research results are written up, e.g., in a paper or book: While in quantitative research, qualitative content analysis and social-science hermeneutics, researchers can more or less decide how their text will be organized after they have focused their question, in grounded theory, the *order of analysis* and the *order of writing* may largely differ—deciding on how the final argument should be structured is an important part of selective coding and can

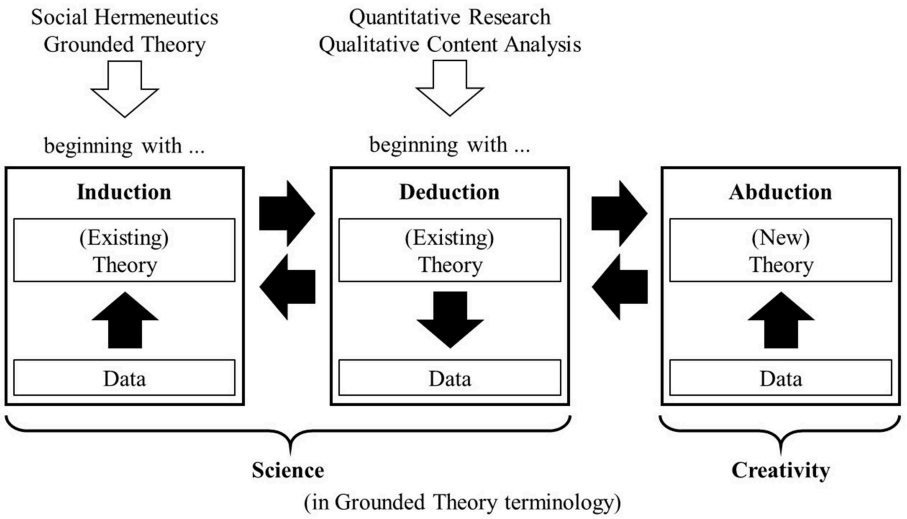

FIGURE 1 | Induction, deduction, and abduction.

only be decided on relatively late during research, i.e., after the research question has been focused.

BEGINNING THE RESEARCH PROCESS: DEDUCTION, INDUCTION AND ABDUCTION

These ideas of how social theory is used for handling objectivity and perspectivity as well as when and how the research question should be focused, strongly influence how the overall research process is designed. Concerning this overall design, the difference between qualitative and quantitative research seems clear:

In current methodological discourse, it is generally assumed that quantitative research is deductive. As depicted in **Figure 1**, "*deduction*" means that researchers start research by deriving hypotheses concerning the research from the selected theory. Researchers then collect and analyze data, in order to test their hypotheses (Hempel and Oppenheim, 1948).

In contrast, it is generally assumed that qualitative research systematically makes use of inductivism. As illustrated in **Figure 1**, "*induction*" starts from the data and then analyses which theory would best fit the data (Strauss and Corbin, 1990).

This simple distinction is another oversimplification in several ways:

First, the idea of induction and deductions has been supplemented by idea of "*abduction*" (Peirce, 1878/1931; Reichertz, 2004b, 2010, 2013), which resembles induction in the sense that both start analysis from data and conclude from data to theory (in contrast to deduction). However, induction only draws on existing theories—if no theory is known that fits or can model the data analysis, induction fails. Researchers can only invent [sic!] a new theory—and this is called abduction (Reichertz, 2004b, 2010, 2013). Grounded theory (Corbin and Strauss, 1990) and social-science hermeneutics (Reichertz, 2004b, 2010, 2013) are the only of the four research traditions which explicitly stress the necessity and importance of abduction, especially as it is the only way of really creating new knowledge. However, in research practice, all researchers in all research traditions need to work abductively at some point, e.g., when they study a completely new social phenomenon (where no prior knowledge can exist).

Secondly, *no actual research process is purely deductive, inductive or abductive*:

- In actual research processes, what usually happens when researchers start their research *deductively*, is that they build a theory, collect and test data—and then research results differ from what was expected. This does not mean that the researcher made a mistake or that research is "bad"—on the contrary: If one assumes that only research questions that can actually yield new results are worthy of being explored, then it is to be expected that results differ from what researchers have deduced from their data. Similarly, if one truly tries to falsify data, it must be possible that the data can actually contradict the theory. The point for the debate about induction and deduction is that researchers usually never end analysis here. Instead, they will take a closer look at the data, re-analyze them and look for other explanations for their results, i.e., they will check, if a different theory than the one considered originally might fit the data better. In the moment they are doing this, they change from the logic of deduction to the logic of induction.
- Likewise, if researchers start data analysis *inductively*, this means that they start interpreting the data and muse, if there is any theory that might fit the data. Once they have identified theories in line with the data, researchers usually go on testing these theories by using further data. An example is the sequence of data sessions and sequence analysis in social-science hermeneutics discussed above. However, in the moment researchers start testing their hypothesis, they have switched from induction to deduction.

– Similarly, if researchers start *abductively*, after abduction, they have a theory that can be tested by deduction.

In further research, researchers will typically switch from induction to deduction and back several time, regardless which paradigm they work with. So in principle, *all social science research makes use of both induction and deduction*. If the existing theories do not fit the data, researchers will additionally make use of abduction.

Rather, the difference between the traditions lies is *how they begin research*:

Quantitative researchers have no choice but to start deductively as they need to know what standardized data they need to collect (e.g., which questions to ask in a survey) and which population the random sample needs to be drawn from—and in order to do so, they need to exactly know what they want to know, i.e., which hypothesis to test. While this is often depicted as an advantage, it is actually a problem when researchers are analyzing unfamiliar fields or if social phenomena are so new that researchers do not know, which theory is appropriate for answering the question.

As suggested by the discussion so far, social-science hermeneutics and grounded theory usually start with induction and then later switch to deduction: Grounded theory specifically aims at building theories for unfamiliar fields, which is exactly one of the reasons why the research question is focused only later in research. In contrast, social-science hermeneutics focus the research question early but only develop hypotheses inductively from the material during data sessions. This illustrates that the logics of deduction and induction are not necessarily linked to the issue when and how the research question is focused.

In contrast, qualitative content analysis also starts deductively. This shows that the simple idea that qualitative research uses induction and quantitative research uses deduction cannot be upheld. On the contrary, there is some qualitative research that starts deductively while other qualitative research might start inductively. Regardless of the logic of beginning, qualitative and quantitative research will swap between the logics in the course of the further research process.

LINEARITY AND CIRCULARITY CONCERNING THE ORDER OF RESEARCH PHASES

The question, if the research process is deductive or inductive is often mingled with the question, if the research process is linear or circular. However, these are not the same things: "Deduction" and "induction" describe ways of linking theory and data. "Linearity" and "circularity" address the issue, how different research phases are ordered, namely, if (a) posing and focusing the research question; (b) sampling; (c) designing instruments; (d) collecting data; (e) preparing data; (f) analyzing data; (g) generalizing results; and (h) archiving data follow one after the other ("linearity"), or if they are iterated ("circularity"). Again, it is generally assumed that quantitative research is organized in a linear way, while qualitative research is organized in a circular way.

Indeed, quantitative research is and always has to be organized in a linear way. This is a direct result of the quantitative concept of objectivity, deduction in combination with the idea of making use of numbers: As stated above, in order to ensure that researchers influence the research process as little as possible but also in order to enable a strong division of labor, research instruments are developed using a prescribed order. Many quantitative techniques do not work, if one deviates from this model. For example, in order to generalize results using inductive statistics, the sample has to be a random sample. A sample is only random, if a population is defined first, then the sample is randomly drawn from this population, and only then data are collected and all units drawn actually participate. If there is unit or item nonresponse, there might be a systematic error (meaning that the sample becomes a nonprobability sample and thus making it impossible to use inductive statistics for generalization). Recruiting additional cases later does not resolve this problem because it contradicts the logic of random sampling (Baur and Florian, 2008; Baur et al., 2018). All in all, this means that both sampling and the development of the instrument must be conducted before data collection. Then, all data must be collected and thereafter analyzed in a bunch. So, the logics of trying to formulate the hypotheses as standardized as possible has the result that quantitative research must be linear in the sense of research phases being organized step-by-step.

Although the need of linearity sometimes is depicted as an advantage, the contrary is true: Often, linearity is a problem because very often, researchers only realize during the actual research that they have made mistakes or false assumptions or forgotten important aspects of the phenomenon under investigation. In circular research processes, these can be easily corrected. However, in linear research processes, this is not possible without setting up a whole new study. That this is not just a general statement but an actual problem that quantitative researchers perceive themselves is reflected in the fact that psychometrics has been using iterative processes of item generation, testing and selection as established practice for several decades. In the last two decades, sociological survey methodologists, too, have tried to derive as much as possible from linearity by developing the concept of the "Survey Life Cycle" (Groves et al., 2009). While in traditional survey methodology, sampling and instrument development were subsequent phases, now at least during instrument development, feedback loops are built in. Panel and trend designs even allow for making slight adjustments both of the instruments and the sample after data analysis in later waves. Regardless, a true circularity is not possible in the logic of quantitative research processes—in principle, the research process concerning the order of building instruments, sampling, data collection and data analysis is linear.

Linearity is also a characteristic of qualitative content analysis, which starts with sampling and collecting data (for example by conducting interviews or sampling texts), then preparing them in a qualitative data analysis software, coding them and afterwards structuring the data. In social-science hermeneutics, too, the overall research process is linear in the sense that usually first data are collected, then transcribed and then analyzed. Similar to survey research, both

qualitative content analysis and social-science hermeneutics might build circular elements into the research process later, e.g., by collecting more data or sampling new cases—still, all in all, all these research processes remain linear in nature, which contrasts common-sense knowledge on qualitative data analysis.

Of the four research traditions analyzed, the only research process truly circular is that of grounded theory. In fact, grounded theory argues most explicitly that linearity is inefficient because a lot of time is wasted on things researchers relatively soon realize they do not need to know and because linearity forces researchers to spend a lot of time before they can actually get started. Thus, grounded theory not only propagates circularity but has also developed suggestions of how to organize this circularity in research practice. In this regard, the key concept is "theoretical sampling," which states that researchers should start analysis as soon as possible with one single case. The first case sampled is ideally the critical case (i.e., a case that should not exist in theory but exists empirically) or the case from which researchers can learn the most given their current understanding. Then data for this case only are collected and immediately analyzed. Depending on what has been learned from the first case, researchers select the second case that likely contrasts the most with the first case. This process is based on the idea that one can learn more from new cases, if they provide as different information as possible. Then data are collected only for the second case and analyzed immediately, then a third contrasting case is selected and so on, until results are "theoretically saturated," i.e., no new ideas or information arises. Theoretical sampling not only allows for developing and adjusting the sampling plan during data analysis but also allows to change the data collection or analysis methods used. For example, researchers could start with qualitative interviews and then later change to ethnography or other kinds of data which will be analyzed. So all in all, given the ways in which the research phases follow each other, it is only grounded theory that differs from the other traditions.

LINEARITY AND CIRCULARITY CONCERNING DATA ANALYSIS

The distinction between linear and circular research processes becomes completely blurred when looking at data analysis. On paper, all qualitative traditions discussed in this paper openly build in circular elements into their data analysis: Researchers using qualitative content analysis conduct different rounds coding the data. Grounded theory, as stated above, as a matter of principle does not only change between different phases of data collection but also differentiates between open, axial and selective coding. Hermeneutics are also circular in the sense that once the different interpretations are developed, the material is tested in different ways.

Quantitative data analysis seems to be completely different, on first sight, as it appears to be linear: If you follow the textbook, quantitative researchers should develop hypotheses at the beginning of the research process, then design their instruments, plan how to analyze them, sample, collect and prepare data. Next, researchers will use statistics to test the hypotheses—and until this step, good quantitative research practice also follows the book.

However, as stated above, what usually happens is that researchers do not achieve the results as expected—and in fact, this is a desirable result, because otherwise research would never produce new insights, and in the sense of quantitative logics, it should be possible to falsify results.

TABLE 1 | Commonalities and differences between research traditions concerning some aspects of the research process.

	Quantitative Research	**Qualitative Content Analysis**	**Grounded Theory**	**Social-Science Hermeneutics**
Handling perspectivity	Perspectivity is a necessary part of the research process and has to be disclosed at the beginning of the research process by explicating the study's theoretical frame and defining central terms and terminology.			
	Aiming at falsification by testing theories and hypothesis	Triangulation (methods, data, theories, researchers)		
		Parallel coding	Theoretical sensitivity	Interpretation groups
Meaning of objectivity	Ideally, different independent researchers should arrive at the same conclusion.	Objectivity in the sense of quantitative researchers is not possible in the social sciences. Instead, researchers should reflect, document and explain how they arrived at their conclusions.		
Focusing the research question	As precisely as possible at the beginning of the analysis	As precisely as possible at the beginning of the analysis	Very general research question at beginning of analysis which is focused during selective coding	As precisely as possible at the beginning of the analysis
Beginning the research process	Deductive	Deductive	Inductive	Inductive
Order of research phases	Linear	Linear	Circular (Theoretical Sampling)	Linear
Data analysis	In theory linear, in practice circular	Circular	Circular	Circular

Still, have you ever read a paper that said "I have done an analysis and did not get the results I wanted or expected … so I am finished now! Sorry!" or "I have falsified my hypotheses and now we do not know anything because what we thought we knew has been falsified"?

That you have very likely never read a paper like this, is because quantitative data analysis is not as strictly deductive-linear as it pretends to be in methods textbooks. But in fact, quantitative data analysis is much more organized in a circular way, similar to the qualitative research traditions. More specifically, when quantitative researchers do not achieve the results they expected, they switch to induction and/or abduction—data analysis now becomes circular in the sense that researchers analyze the dataset in different rounds. After the first round of unexpected results, researchers might e.g., either conduct a more detailed analysis of a specific variable or subgroup which is more interesting, or they might use different statistical procedures to find clues why the results were different than expected. The only difference to qualitative research is that quantitative researchers have to limit their analysis to the data they have—if information is not contained in the data set, they would need to conduct a whole new study. Regardless, the important point for this paper is that—while the overall research process can be either organized in a linear or circular way, during data analysis, all social science research is organized in a circular way in research practice, whether researchers admit this or not.

DISCUSSION

In this paper, I have shown how four research traditions (quantitative research, qualitative content analysis, grounded theory, social-science hermeneutics) handle six issues to be resolved when deciding on a social science research design, namely: How is researchers' perspectivity handled during the research process? How can intersubjectivity be achieved, and what does "objectivity" mean in this context? When and how is the research question is focused? Does the research process start deductively or inductively? Are the diverse research phases (sampling, data collection, data preparation, data analysis) organized in a linear or circular way? Is data analysis organized in a linear or circular way? For each of these issues, I have discussed how the four traditions resemble and differ from each other. **Table 1** sums up the various positions.

When regarding the whole picture depicted in **Table 1**, it is possible to state that the common-sense knowledge that "quantitative research" organizes its research process deductively, tests theories, does objective, positivist research and organizes the research process in a linear way while "qualitative research" organizes its research process inductively, develops theories, has a constructivist stance on research and organizes the research process in a circular way, cannot be upheld for at least three reasons:

1. *Quantitative research is not as objective, deductive and linear as it is often depicted in literature*. It is much more necessary to interpret in *all* phases of quantitative research as quantitative researchers usually admit. During data analysis, quantitative research has always iterated between deduction, induction and abduction, and concerning the overall organization of the research process, quantitative research has recently tried to dissolve linearity as much as possible, as exemplified in the concept of the "Survey Life Cycle."
2. For all these issues, *there are some qualitative traditions that resemble quantitative research more than quantitative research*. As this is not a new revelation, the distinction between "qualitative" and "qualitative" research has often been depicted as continuum, resulting in an order (from strong "quantitativeness" to strong "qualitativeness") from quantitative methods, qualitative content analysis, grounded theory and social-science hermeneutics. The general argument is that qualitative content analysis is "almost" quantitative research, while social-science hermeneutics is one of the "truest" forms of qualitative research.
3. However, *neither can a continuum between qualitative and quantitative research be upheld*, i.e., one can neither claim that qualitative content analysis is *per se* closer to quantitative research than social-science hermeneutics nor is grounded theory positioned in the middle. Rather, depending on the debated issue concerning the research process, social-science hermeneutics might resemble quantitative research much more than qualitative content analysis. For example, while it is true that both quantitative research and qualitative content analysis organize the overall research process more linearly than grounded theory and social-science hermeneutics do, when it comes to handling theory, social-science hermeneutics are "stricter" than the other two qualitative traditions in the sense that they systematically test theories.

To conclude, the oversimplified distinction between "qualitative" and "quantitative" research cannot be upheld. This is both a chance and a challenge for mixed methods research. On the bright sight, mixing and combining qualitative and quantitative methods becomes easier because the distinction is not as grand as it seems at first sight and the boundaries between research traditions are much more blurred. It thus might be easier to focus on practical issues of mixing instead of epistemological debates. On the dark sight, mixing becomes more difficult because some tricky issues of mixing specific types of methods are usually not addressed in current mixed methods discourse. More specifically, mixed methods research so far has strongly focussed on mixing traditions that can be easily mixed due to some similarities in the research process, e.g., quantitative research and qualitative content analysis. When the discussion presented here is taken seriously, it would be much more fruitful to discuss how to combine quantitative research e.g., with grounded theory and social-science hermeneutics because in these traditions the research process is more circular and this circularity is part of their strength. To effectively use the potential of these paradigms, it would be necessary to implement these circular elements in mixed methods research.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Adorno, T. W. (1962/1969/1993). "Zur Logik der Sozialwissenschaften," in *Der Positivismusstreit in der Deutschen Soziologie*, eds T.W. Adorno, R. Dahrendorf, R. Pilot, H. Albert, J. Habermas, and K. R. Popper (München: dtv), 125–143.

Akremi, L. (2017). Mixed-Methods-Sampling als Mittel zur Abgrenzung eines unscharfen und heterogenen Forschungsfeldes. Am Beispiel der Klassifizierung von Zukunftsängsten im dystopischen Spielfilm. *Kölner Z. Soz. Sozialpsychol.* 69 (Suppl. 2), 261–286. doi: 10.1007/s11577-017-0460-3

Akremi, L. (2018). "Interpretativität quantitativer Auswertung. Über multivariate Verfahren zur Erfassung von Sinnstrukturen," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 361–408.

Ametowobla, D., Baur, N., and Norkus, M. (2017). "Analyseverfahren in der empirischen Organisationsforschung," in *Handbuch Empirische Organisationsforschung*, eds S. Liebig, W. Matiaske, and S. Rosenbohm (Wiesbaden: Springer), 749–796. doi: 10.1007/978-3-658-08493-6_33

Baur, N. (2005). *Verlaufsmusteranalyse: Methodologische Konsequenzen der Zeitlichkeit sozialen Handelns*. Wiesbaden: VS-Verlag für Sozialwissenschaften. doi: 10.1007/978-3-322-90815-5

Baur, N. (2008). Taking perspectivity seriously: a suggestion of a conceptual framework for linking theory and methods in longitudinal and comparative research. *Hist. Soc. Res.* 33(4), 191–213. doi: 10.12759/hsr.33.2008.4.191-213

Baur, N. (2009a). Measurement and selection bias in longitudinal data: A framework for re-opening the discussion on data quality and generalizability of social bookkeeping data. *Hist. Soc. Res.* 34(3), 9–50. doi: 10.12759/hsr.34.2009.3.9-50

Baur, N. (ed.). (2009b). *Social Bookkeeping Data: Data Quality and Data Management. Historical Social Research 129*. Mannheim and Köln: GESIS.

Baur, N. (2009c). Problems of linking theory and data in historical sociology and longitudinal research. *Hist. Soc. Res.* 34(1), 7–21. doi: 10.12759/hsr.34.2009.1.7-21

Baur, N. (2011). Mixing process-generated data in market sociology. *Qual. Quant.* 45(6), 1233–1251. doi: 10.1007/s11135-009-9288-x

Baur, N. (2014). Comparing societies and cultures: Challenges of cross-cultural survey research as an approach to spatial analysis. *Hist. Soc. Res.* 39(2), 257–291. doi: 10.12759/hsr.39.2014.2.257-291

Baur, N. (2018). "Kausalität und Interpretativität. Über den Versuch der quantitativen Sozialforschung, zu Erklären, ohne zu Verstehen," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 306–360.

Baur, N., and Blasius, J. (2019). "Methoden der empirischen Sozialforschung. Ein Überblick," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1–30.

Baur, N., and Ernst, S. (2011). Towards a process-oriented methodology. Modern social science research methods and Nobert Elias' figurational sociology. *Sociol. Rev.* 59(777), 117–139. doi: 10.1111/j.1467-954X.2011.01981.x

Baur, N., and Florian, M. (2008). "Stichprobenprobleme bei Online-Umfragen," in *Sozialforschung im Internet. Methodologie und Praxis der Online-Befragung*, eds N. Jackob, H. Schoen, and T. Zerback (Wiesbaden: VS-Verlag), 106–125.

Baur, N., and Hering, L. (2017). Die Kombination von ethnografischer Beobachtung und standardisierter Befragung. Mixed-Methods-Designs jenseits der Kombination von qualitativen Interviews mit quantitativen Surveys. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 387–414. doi: 10.1007/s11577-017-0468-8

Baur, N., Kelle, U., and Kuckartz, U. (2017). Mixed Methods – Stand der Debatte und aktuelle Problemlagen. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 1–37. doi: 10.1007/s11577-017-0450-5

Baur, N., and Knoblauch, H. (2018). Die Interpretativität des Quantitativen, oder: zur Konvergenz von qualitativer und quantitativer empirischer Sozialforschung. *Soziologie* 47(4), 439–461.

Baur, N., Knoblauch, H., Akremi, L., and Traue, B. (2018). "Qualitativ – Quantitativ – Interpretativ: Zum Verhältnis Methodologischer Paradigmen in der Empirischen Sozialforschung," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 246–284.

Baur, N., and Lamnek, S. (2017a): "Einzelfallanalyse," in *Qualitative Medienforschung*, eds L. Mikos and C. Wegener (Konstanz: UVK), 274–284.

Baur, N., and Lamnek, S. (2017b). "Multivariate analysis," in *The Blackwell Encyclopedia of Sociology*, ed G. Ritzer (Oxford: Blackwell Publishing Ltd.). doi: 10.1111/b.9781405124331.2007.x

Behnke, J., Baur, N., and Behnke, N. (2010). *Empirische Methoden der Politikwissenschaft*. Paderborn: Schöningh.

Blasius, J., and Thiessen, V. (2012). *Assessing the Quality of Survey Data*. London: Sage.

Bryman, A. (1988). *Quantity and Quality in Social Research*. London: Routledge and Kegan Paul.

Charmaz, K. (2006). *Constructing Grounded Theory. A Practical Guide through Qualitative Analysis*. London: Sage.

Clarke, A. E. (2005). *Situational Analysis. Grounded Theory After the Postmodern Turn*. Thousand Oaks, CA; London; New Delhi: Sage.

Connell, R. (2007). *Southern Theory*. Cambridge and Malden: Polity.

Corbin, J. M., and Strauss, A. (1990). Grounded theory research: procedures, canons, and evaluative criteria. *Qual. Sociol.* 13(1), 3–21. doi: 10.1007/BF00988593

Desrosières, A. (2002). *The Politics of Large Numbers. A History of Statistical Reasoning*. Harvard: Harvard University Press.

Edwards, R., Goodwin, J., O'Connor, H., and Phoenix, A. (2017). *Working With Paradata, Marginalia and Fieldnotes: The Centrality of By-Products of Social Research*. Cheltenham: Edward Elgar. doi: 10.4337/9781784715250

Elias, N. (1987). *Engagement und Distanzierung. Arbeiten zur Wissenssoziologie I*. Frankfurt a.M.: Suhrkamp.

Elias, N. (1997). Towards a theory of social processes. *Br. J. Sociol.* 48(3), 355–383. doi: 10.2307/591136

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Cambridge; Oxford: Polity and Blackwell.

Feyerabend, P. (1975). *Against Method*. London: New Left Books.

Flick, U. (1992). Triangulation revisited – strategy of or alternative to validation of qualitative data. *J. Theor. Soc. Behav.* 2, 175–197. doi: 10.1111/j.1468-5914.1992.tb00215.x

Flick, U. (2017). Mantras and myths: the disenchantment of mixed-methods research and revisiting triangulation as a perspective. *Qual. Inq.* 23(1), 46–57. doi: 10.1177/1077800416655827

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (eds.). (2017). *Big data and Social Science. A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC Press.

Glaser, B., and Strauss, A. (1975). *Awareness of Dying*. Chicago, IL: Aldine.

Glaser, B. G. (1978). *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*. Mill Valley, CA: The Sociology Press.

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: Wiley.

Habermas, J. (1981). *Theorie des kommunikativen Handelns*, Vol. 2. Frankfurt am Main: Suhrkamp.

Hartmann, P., and Lengerer, A. (2019). "Verwaltungsdaten und Daten der amtlichen Statistik," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1223–1232.

Helfferich, C. (2019). "Leitfaden- und Experteninterviews," in *Handbuch Methoden der empirischen Sozialforschung,* eds N. Baur and J. Blasius (Wiesbaden: Springer), 669–686. doi: 10.1007/978-3-531-18939-0_39

Hempel, C. G., and Oppenheim, P. (1948): Studies in the logic of explanation. *Philos. Sci.* 15, 135–175. doi: 10.1086/286983

Hense, A. (2017). Sequentielles Mixed-Methods-Sampling: Wie quantitative Sekundärdaten qualitative Stichprobenpläne und theoretisches Sampling unterstützen können. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 237–259. doi: 10.1007/s11577-017-0459-9

Herbrik, R. (2018). "Hermeneutische Wissenssoziologie (sozialwissenschaftliche Hermeneutik). Das Beispiel der kommunikativen Konstruktion normativer, sozialer Fiktionen, wie z. B. 'Nachhaltigkeit' oder: Es könnte auch immer anders sein," in *Handbuch Interpretativ forschen,* eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 659–680.

Hübner, K. (2002). *Kritik der wissenschaftlichen Vernunft*. Freiburg: Alber.

Johnson, R. B., de Waal, C., Stefurak, T., and Hildebrand, D. L. (2017). Understanding the philosophical positions of classical and neopragmatists for mixed methods research. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 63–86. doi: 10.1007/s11577-017-0452-3

Kalthoff, H. (2008). "Zur Dialektik von qualitativer Forschung und soziologischer Theoriebildung", *Theoretische Empirie*, eds H. Kalthoff, S. Hirschauer, and G. Lindemann (Frankfurt a.M.: Suhrkamp), 8–34.

Kelle, U. (2008). *Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung*. Wiesbaden: Springer.

Kelle, U. (2017). Die Integration qualitativer und quantitativer Forschung – theoretische Grundlagen von 'Mixed Methods'. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 39–61. doi: 10.1007/s11577-017-0451-4

Kelle, U. (2018). "Datenerhebung in der quantitativen Forschung. Eine interpretative Perspektive auf Fehlerquellen im standardisierten Interview," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 285–305.

Knoblauch, H., Baur, N., Traue, B., and Akremi, L. (2018). "Was heißt, Interpretativ forschen?" in *Handbuch Interpretativ Forschen,* eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 9–35.

Knoblauch, H., and Pfadenhauer, M. (eds.). (2018). *Social Constructivism as Paradigm? The Legacy of the Social Construction of Reality*. London: Routledge.

Knoblauch, H., and Vollmer, T. (2019). "Ethnografie," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 599–618. doi: 10.1007/978-3-658-21308-4_40

König, C., Schröder, J., and Wiegand, E. (eds.). (2018). *Big Data. Chancen, Risiken, Entwicklungstendenzen.* Wiesbaden: Springer. doi: 10.1007/978-3-658-20083-1

Koselleck, R. (1977). "Standortbindung und Zeitlichkeit. Ein Beitrag zur historiographischen Erschließung der geschichtlichen Welt," in *Objektivität und Parteilichkeit in der Geschichtswissenschaft*, eds R. Koselleck, W.J. Mommsen, and J. Rüsen (München: dtv), 17–46. Reprinted 1979 in *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten*, ed R. Koselleck (Frankfurt a.M.: Suhrkamp), 176–207.

Kuckartz, U. (2010). *Einführung in die Computergestützte Analyse Qualitativer Daten*. Wiesbaden: Springer. doi: 10.1007/978-3-531-92126-6

Kuckartz, U. (2014). *Qualitative Text Analysis. A Guide to Methods, Practice and Using Software.* London; Thousand Oaks, CA; New Delhi; Singapore: Sage.

Kuckartz, U. (2017). Datenanalyse in der Mixed-Methods-Forschung. Strategien der Integration von qualitativen und quantitativen Daten und Ergebnissen. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 157–183. doi: 10.1007/s11577-017-0456-z

Kuckartz, U. (2018). "Qualitative Inhaltsanalyse. Am Beispiel einer Studie zu Klimabewusstsein und individuellem Verhalten," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 506–535.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Kurt, R., and Herbrik, R. (2019). "Sozialwissenschaftliche Hermeneutik und hermeneutische Wissenssoziologie," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 545–564. doi: 10.1007/978-3-658-21308-4_37

Lakatos, I. (1976). *Proofs and Refutations*. London: The Logic of Mathematical Discovery.

Lamnek, S. (2005). *Qualitative Sozialforschung*. Weinheim; Basel: Beltz.

Langfeldt, B., and Goltz, E. (2017). Die Funktion qualitativer Vorstudien bei der Entwicklung standardisierter Erhebungsinstrumente. Ein Beispiel aus der Evaluationsforschung in militärischem Kontext. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 313–335. doi: 10.1007/s11577-017-0462-1

Lindemann, G. (2008). "Theoriekonstruktion und empirische Forschung," in *Theoretische Empirie*, eds H. Kalthoff, S. Hirschauer, and G. Lindemann (Frankfurt a.M.: Suhrkamp), 165–187.

Lyotard, J.-F. (1979/2009). *Das Postmoderne Wissen*. Wien: Passagen Verlag.

Maiwald, K.-O. (2018). "Objektive Hermeneutik. Von Keksen, inzestuöser Verführung und dem Problem, die Generationendifferenz zu denken – exemplarische Sequenzanalyse einer Interaktion in einem Fernsehwerbefilm," in *Handbuch Interpretativ Forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 442–478.

Mayerl, J., and Faas, T. (2019). "Quantitative Analyse von Twitter und anderer usergenerierter Kommunikation," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1027–1040. doi: 10.1007/978-3-658-21308-4_73

Mayring, P. (2014). *Qualitative Content Analysis. Theoretical Foundation, Basic Procedures and Software Solution*. Klagenfurt: Beltz.

Meidl, C. N. (2009). *Wissenschaftstheorien für SozialforscherInnen*. Wien; Köln; Weimar: Böhlau.

Merton, R. K. (1942/1973). "The normative structure of science," in *The Sociology of Science: Theoretical and Empirical Investigations*, ed R. K. Merton (Chicago, IL: University of Chicago Press), 267–278.

Mignolo, W. (2011). "I am where I think: Remapping the Order of Knowing," in *The Creolization of Theory*, eds F. Lionnet and S.-M. Shih (Durham; London: Duke University Press), 159–192. doi: 10.1215/9780822393320-007

Nam, S.-H. (2019). "Qualitative Analyse von Chats und anderer usergenerierter Kommunikation," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1041–1052. doi: 10.1007/978-3-658-21308-4_74

Oevermann, U., Allert, T., Konau, E., and Krambeck, J. (1979): "Die Methodologie einer 'objektiven Hermeneutik' und ihre allgemeine forschungslogische Bedeutung in den Sozialwissenschaften," in *Interpretative Verfahren in den Sozial- und Textwissenschaften*, ed H.-G. Soeffner (Stuttgart: Metzler), 352–434.

Peirce, C. S. (1878/1931). "Deduction, induction, and hypothesis," in *Collected Papers of Charles Sanders Peirce,* Vol. 2, eds C. S. Peirce, C. Hartshorne, and P. Weiss, P. (Cambridge: Belknap Press of Harvard University Press), 619–644.

Popper, K. R. (1935): *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft.* Wien: Springer. doi: 10.1007/978-3-7091-4177-9

Reichertz, J. (2004a). Hermeneutic sociology of knowledge. *IHS Newslett.* 12(2), 9–10.

Reichertz, J. (2004b). "Abduction, deduction and induction in qualitative research," in *Companion to Qualitative Research*, eds U. Flick et al. (London: Sage), 159–165.

Reichertz, J. (2010). "Abduction: The logic of discovery of Grounded Theory," in *The Sage Handbook of Grounded Theory*, eds A. Bryant and K. Charmaz (London: Sage), 214–229.

Reichertz, J. (2013). *Die Abduktion in der qualitativen Sozialforschung. Über die Entdeckung des Neuen.* Wiesbaden: Springer VS.

Reichertz, J. (2018). "Interpretieren in Interpretationsgruppen. Versprechungen, Formen, Bedingungen, Probleme," in *Handbuch Interpretativ forschen,* eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 72–107.

Reichertz, J. (2019). "Empirische Sozialforschung und soziologische Theorie," in *Handbuch Methoden der empirischen Sozialforschung,* eds N. Baur and J. Blasius (Wiesbaden: Springer), 31–48. doi: 10.1007/978-3-658-21308-4_2

Riebling, J. (2017). *Methode und Methodologie quantitativer Textanalyse. [dissertation]*. Bamberg: Otto-Friedrich Universität Bamberg

Riedl, R. (1985). *Die Spaltung des Weltbildes. Biologische Grundlagen des Erklärens und Verstehens*. Berlin and Hamburg: Paul Parey.

Rose, G. (2016). *Visual Methodologies*. London; New Delhi; Thousand Oaks, CA: Sage.

Salheiser, A. (2019). "Natürliche Daten: Dokumente," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1119–1134. doi: 10.1007/978-3-658-21308-4_80

Schmidt, J.-H. (2019). "Blogs," in *Handbuch Methoden der empirischen Sozialforschung,* eds N. Baur and J. Blasius (Wiesbaden: Springer), 1015–1026.

Schmitz, A., and Yanenko, O. (2019). "Web Server Logs und Logfiles," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur, and J. Blasius (Wiesbaden: Springer), 991–1000.

Schoonenboom, J., and Johnson, R. B. (2017). How to construct a mixed methods research design. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 107–131. doi: 10.1007/s11577-017-0454-1

Schreier, M. (2012). *Qualitative Content Analyis in Practice*. Los Angeles, CA: London: Sage.

Schünzel, A., and Traue, B. (2019). "Websites," in *Handbuch Methoden der empirischen Sozialforschung,* eds N. Baur and J. Blasius (Wiesbaden: Springer), 1001–1014. doi: 10.1007/978-3-658-21308-4_71

Seale, C. (1999): *The Quality of Qualitative Research*. London; Thousand Oaks; New Delhi: Sage.

Shih, S.-M., and Lionnet, F. (2011). "The creolization of theory," in *The Creolization of Theory,* eds F. Lionnet and S.-M. Shih (Durham; London: Duke University Press), 1–36.

Smelser, N. J. (1976). *Comparative Methods in the Social Sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Strauss, A., and Corbin, J. M. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Thousand Oaks, CA: Sage.

Strübing, J. (2014). *Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung eines pragmatistischen Forschungsstils*. Wiesbaden: Springer VS.

Strübing, J. (2018). "Situationsanalyse. Eine pragmatistische Erweiterung der Grounded Theory unter dem Eindruck der Postmoderne," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 681–707.

Strübing, J. (2019). "Grounded theory," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 525–544.

Traue, B., and Schünzel, A. (2019). "YouTube und andere Webvideos," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1065–1078.

Uher, J. (2018). Data generation methods across the empirical sciences: Differences in the study phenomena's accessibility and the processes of data encoding. *Qual. Quant. Int. J. Methodol.* 53(1), 221–246. doi: 10.1007/s11135-018-0744-3

Vogl, S. (2017). Quantifizierung. Datentransformation von qualitativen Daten in quantitative Daten in Mixed-Methods-Studien. *Kölner Z. Soz. Sozialpsychol.* 69(Suppl. 2), 287–312. doi: 10.1007/s11577-017-0461-2

von Glasersfeld, E. (1994). "Einführung in den radikalen Konstruktivismus," in *Die erfundene Wirklichkeit*, ed P. Watzlawick (München: Piper), 16–38.

Wehler, H.-U. (1980). *Historische Sozialwissenschaft und Geschichtsschreibung*. Göttingen: Vandenhoeck and Ruprecht.

Weil, S., Eberle, T. S., and Flick, U. (2008). Between Reflexivity and Consolidation—Qualitative Research in the Mirror of Handbooks. *Forum Qual. Soc. Res.* 9(3). Available online at: http://nbn-resolving.de/urn:nbn:de:0114-fqs0803280

Weiß, B., and Wagner, M. (2019). "Meta-analyse," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 1511–1520.

Wernet, A. (2006). *Hermeneutik – Kasuistik – Fallverstehen*. Stuttgart: Kohlhammer.

Ziegler, M. (2018). "Interpretativität und schließende Statistik. Das Verhältnis von sozialem Kontext und Generalisierungsstrategien der quantitativen Sozialforschung," in *Handbuch Interpretativ forschen*, eds L. Akremi, N. Baur, H. Knoblauch, and B. Traue (Weinheim; München: Beltz Juventa), 409–441.

Züll, C., and Menold, N. (2019). "Offene Fragen," in *Handbuch Methoden der empirischen Sozialforschung*, eds N. Baur and J. Blasius (Wiesbaden: Springer), 855–862.

Conflict of Interest Statement: 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.

frontiers in Psychology

OPINION
published: 13 August 2019
doi: 10.3389/fpsyg.2019.01866

The Rationality of Science and the Inevitability of Defining Prior Beliefs in Empirical Research

*Ulrich Dettweiler**

Department of Cultural Studies and Languages, Faculty of Arts and Education, University of Stavanger, Stavanger, Norway

Keywords: Bayesian statistics, frequentist statistics, epistemology, prior probability function, rationality of science, philosophy of science

INTRODUCTION

The recent "campaign" in Nature against the concept of "significance testing" (Amrhein et al., 2019), with more than 800 supporting signatories of leading scientists, can be considered as an important milestone and somewhat resounding event in the long on-ongoing struggle and somewhat "silent revolution" (Rodgers, 2010) in statistics over logical, epistemological, and praxeological aspects (Meehl, 1997; Sprenger and Hartmann, 2019), criticizing over-simplified and thoughtless statistical analyses which still can be found in overwhelming many publications to-date. So-called frequentists, the Neyman/Pearson and Fisher schools, and those who apply a hybrid scheme of the two schools (Mayo, 1996) or simple Null Hypothesis Testing (NHST), likelihoodists, and Bayesians alike have debated their approaches over the past decades. This finally led to a discourse facilitated by the *American Statistical Association*, resulting in a special issue of The American Statistician (Vol. 73/2019) titled: "Statistical Inference in the 21st century: A World Beyond $p < 0.05$," with "43 innovative and thought-provoking papers from forward-looking statisticians" (Wasserstein et al., 2019, p. 1). The special issue proposes both new ways to report the importance of research results beyond the arbitrary threshold of a categorical p-value, and some guides of conduct: the researcher should accept uncertainty, be thoughtful, open and modest in their claims (Wasserstein et al., 2019). The future will show if those attempts to statistically better supported science beyond significance testing will be echoed in the publications to come.

A corresponding discourse has been led by the Royal Statistical Society, whereby Andrew Gelman's and Christian Hennig's contribution "Beyond subjective and objective in statistics" has been discussed by more than 50 leading statisticians (Gelman and Hennig, 2017). They suggest to stop using the rather vague terms "objectivity" and "subjectivity," and replace them with "transparency, consensus, impartiality, and correspondence to observable reality" for the former, and "awareness of multiple perspectives and context dependence" for the latter. Together with "stability," these should "make up a collection of virtues" that they consider "helpful in discussions of statistical foundations and practice" (Gelman and Hennig, 2017, p. 967).

Yet, questioning the very concept of "objectivity" might be quite provocative and absurd to most empirical scientists who hold "objectivity" to be a central property of observables, or at least to be the property of scientific method that produces pure, value-free facts. In this light, it is interesting to note that both strategies for overcoming the "statistical crisis in science" (Gelman and Loken, 2014) focus on the researchers' conduct and employ *moral* categories for the *ontological* and *epistemological* problem of *what* we should *believe*.

In this article, I will stress the importance of epistemic beliefs in science for the methods we employ. For this purpose, I will recall an argument that Hilary Putnam proposed more than 35 years ago in his critique of scientific realism. Putnam's philosophy of science had been discussed by statisticians like Meehl and Cronbach at that time (Fiske and Shweder, 1986), but his ideas have since

OPEN ACCESS

Edited by:
Alessandro Giuliani,
Istituto Superiore di Sanità (ISS), Italy

Reviewed by:
Ryota Nomura,
The University of Tokyo, Japan

***Correspondence:**
Ulrich Dettweiler
ulrich.dettweiler@uis.no

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *29 June 2019*
Accepted: *29 July 2019*
Published: *13 August 2019*

Citation:
Dettweiler U (2019) The Rationality of Science and the Inevitability of Defining Prior Beliefs in Empirical Research. Front. Psychol. 10:1866. doi: 10.3389/fpsyg.2019.01866

been overlooked in the above-mentioned discourses. Putnam claims that the concept of rationality, as it is assumed in science, is in fact deeply irrational, if it considers methods to be purely formal, distinct and free from value-judgements. There is also an informal part inherent to rationality in science which depends on the changing beliefs of scientists.

At the core of Putnam's argument lies a fundamental critique of verficationalism with its correspondence theory of truth, which is disguised in the assumption that there are such things as "objective" facts, independent of our "subjective" experiences, thoughts, and language.

THE IMPACT OF SCIENCE ON MODERN CONCEPTIONS OF RATIONALITY

A prominent account of such scientific realism can be found in a later work of John Searle, with whom Putnam fought many philosophical battles (Horowitz, 1996; Cruickshank, 2003).

According to Searle, modern science recurs to "default positions" that are not questioned and "any departure from them requires a conscious effort and a convincing argument." The most central default position implicit in standard empirical research is that we have direct perceptual access to the world through our senses and that the world exists independently of human observation, which is labeled a "correspondence theory of truth" (Searle, 1999).

Yet, the philosophical cost of such an epistemological stance is high: The underlying ontological assumptions in correspondence theories become increasingly counterintuitive and less understandable with the attenuation of their metaphysical ingredients, requiring ability to position the researcher as having an entirely external "god's eye point of view" (Putnam, 1981, p. 49). In other words: Despite the anti-transcendentalist claim of such positivist sciences, the forms of rationality employed derive upon much more substantial metaphysical assumptions than pragmatist methodologies; yet from increased skepticism, the comprehensibility and commonsensical acceptability of science decreases (Dettweiler, 2015).

Despite Putnam has changed his philosophical ideas throughout his life, one constant theme (at least since the 1970ies) is his pragmatist ontological position, which at many points is neither realistic nor ideal. In his claims that, although the world may be *causally* independent of the human mind, the structure of the world (both in terms of individuals and categories) is a function of the human mind and hence is not *ontologically* independent (cf. Brown, 1988). Hereby, Putnam refers to Kant's concept of the dependence of our knowledge of the world on the "categories of thought" and he claims that there is "*a fact of the matter* as to whether the statements people make are warranted or not" (Putnam, 1981, p. 21, cursive by U.D.). This material, realistic reference allows Putnam to talk about warranted truth that is "independent of whether the majority of one's cultural peers would say it is warranted or unwarranted" (ibid). In this respect, Putnam is more than a mere consensus theorist, but not yet a naturalistic realist. He argues instead that "reason can't be naturalized" (Putnam, 1983) and that here and now "truth is independent of justification..., but not independent of *all* justification. To claim a statement is true is to claim it could be justified" (Putnam, 1981, p. 56). Or as Cronbach (1986) reframes Putnam: "Realism is an empirical hypothesis … that can be defended if we observe that a science converges (p. 90).

So, the main challenge to empirical science is the implicit refutation of the claim that the world is accessible independently from the interpretation through our senses and language. It is, according to Putnam, conceptually impossible "to draw a sharp line between the content of science and the method of science," and "the method of science in fact changes constantly as the content of science changes" (Putnam, 1981, p. 191).

"TUNING-FREE" DOES NOT MEAN "VALUE-FREE"

This has, or rather should have, direct implications to the understanding of modern science and the statistical framework it is built on. Putnam argues that any scientific methodology needs to take into account the prior beliefs of scientists and the degree of uncertainty of hypotheses. This means, on the other hand, that we scientists need to make explicit those beliefs that are implicit in the methodologies we apply and quantify in some way uncertainty.

This is often an alien thought to scientists who apply frequentist statistics in their data analyses and reject the "use of subjective uncertainty in the context of scientific inquiry" (Sprenger, 2016, p. 382). It is the very idea of frequentist statistics, that in the long run, the underlying procedure leads to a (probably) correct result irrespective of the researchers' beliefs. Yet the convenience of standard statistical programs with its many default settings should not disguise the many choices implicitly made in the simplest statistical operations. Most researchers hardly question why we fit the data into a Gausian model with a uniform distribution on the infinite range for each of the parameters, and a uniform distribution for the error term as well? With the decision to model the data linearly, according to a normal function within an infinite range of possible values, there are already a number of value-driven presuppositions in the model before we even have started entering the data. The rationale behind the uniform prior probability functions used in standard statistical models is, of course, that it contains as little information as possible, in order to make it a "neutral" procedure. But as Gelman and Hennig (2017) argue, "even using 'no need for tuning' as a criterion for method selection or prioritizing bias, for example, or mean-squared error, is a subjective decision" (p. 971).

There is, as Gelman and Hill (2007) state, nothing wrong with modeling data with uniform distributions on all the parameters. They call those models "reference" models, which provide some important preliminary information in a given data analysis. However, "neutral" does not mean "value-free." We can conceive of many other distribution functions, with more specified parameters, informed by previous research and representing the researchers' prior beliefs, which might better fit the data.

Bayes theorem does provide us with a statistical framework that tells us how data should change our (subjective) degrees of belief in a hypothesis, within a formal model of rational belief provided by the probability calculus. Bayes theorem states that the posterior distribution, i.e., the probability of the parameters given the data, is proportional to the likelihood, which is the probability of observing the data given the parameters (unknowns) multiplied by the prior probability, which represents external knowledge about the parameters.

In fact, Putnam sees subjective Bayesianism as the statistical framework that can assume a formalized language of science in which reliable observations together with some hypotheses can be rationally expressed.

It is from this point that Gelman and Hennig (2017) initiate their proposal to collapse the dichotomy of objectivity and subjectivity altogether. They demonstrate that those prior probability functions are not so much "subjective degrees of belief" but rather "external information" on a specific research question including "restrictions such as smoothness or sparsity that serve to regularize estimates in high dimensional settings, … the choice of the functional form in a regression model, … and … numerical information about particular parameters in a model." This is why Sprenger (2018) argues that the so-called "subjective Bayesianism" should in fact be understood as "objective," thereby defending the language of "objectivity" in science.

GOOD SCIENCE IS A MATTER OF ETHICS, BUT NOT ALONE

I agree with Gelman and Hennig that the dichotomy of "subjective" and "objective" causes a lot of confusion in science, especially when it is applied to classify statistical methodology. It is misleading to (dis)qualify Bayesian statistics as "subjective" when prior probability functions for each parameter in a model are defined with great rigor and transparency. It is also misleading when frequentist researchers use default settings in analyses and claim "objectivity" on their side.

This is, however, not so much a question of ethics. Nor can this tension be solved with introducing rules for the virtuous scientist. It is rather a symptom of a fundamental *epistemological* crisis in modern science. The philosophy of science has been too detached from the empirical sciences and statistics for too long, and those gaps need to be bridged with the education of scientists in epistemology, a claim made by Meehl more than 20 years ago (Meehl, 1997). The enhanced rigor of the scientific enquiry will then follow, since the scientific virtues are inspired by the epistemic beliefs that scientists hold. We simply need to learn again to argue for our epistemological stances, and to define the epistemic claims we make with our statistical analyses, given the data. The epistemologically informed scientist would certainly not be scared to endorse subjectivity as a reliable philosophical concept for empirical science, as Putnam has shown.

Or, as Ian Hacking wittingly summarizes this crisis, all we need to do is think harder, not more objectively (Hacking, 2015).

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The University of Stavanger supported this work with a sabbatical and a grant in the program for Yngre Fremragende Forskere financed by the Norwegian Research Council (project number IN11714).

ACKNOWLEDGMENTS

The brief summary of Putnam's account is nearly verbatim taken from my dissertation (Dettweiler, 2015). Educational Research in the Mirror of Nature. Theoretical, Epistemological, and Empirical Aspects of Mixed-Method Approaches in Outdoor Education (Ph.D. Thesis). Technische Universität München, München). Many thanks to Dr. Mike Rogerson and the reviewer, whose valuable comments very much improved the line of thought in this article.

REFERENCES

Amrhein, V., Greenland, S., and McShane, B. (2019). Retire statistical significance. *Nature* 567, 305–307. doi: 10.1038/d41586-019-00857-9

Brown, C. (1988). Internal realism: transcendental idealism? *Midwest Stud. Philos.* 12, 145–155.

Cronbach, L. J. (1986). "Social inquiry by and for earthlings," in *Metatheory in Social Science. Pluralisms and Subjectivities,* eds D. W. Fiske and R. A. Shweder (Chicago: The University of Chicago Press), 83–107.

Cruickshank, J. (2003). *Realism and Sociology: Anti-Foundationalism, Ontology, and Social Research*, Vol. 5. London, New York, NY: Routledge.

Dettweiler, U. (2015). *Educational Research in the Mirror of Nature. Theoretical, Epistemological, and Empirical Aspects of Mixed-Method Approaches in Outdoor Education.* PhD Thesis, Technische Universität München, München. Retrieved from: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20150602-1241435-1-3

Fiske, D. W., and Shweder, R. A. (eds.). (1986). *Metatheory in Social Science. Pluralisms and Subjectivities*. Chicago: The University of Chicago Press.

Gelman, A., and Hennig, C. (2017). Beyond subjective and objective in statistics. *J. R. Stat. Soc. Ser. A* 180, 967–1033. doi: 10.1111/rssa.12276

Gelman, A., and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York; Cambridge: Cambridge University Press.

Gelman, A., and Loken, E. (2014). The statistical crisis in science. *Am. Sci.* 102:460. doi: 10.1511/2014.111.460

Hacking, I. (2015). "Let's not talk about objectivity," in *Objectivity in Science,* eds F. Padovani, A. Richardson, and J. Y. Tsou (Cham: Springer), 19–33.

Horowitz, A. (1996). Putnam, searle, and externalism. *Philos. Stud.* 81, 27–69.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago; London: University of Chicago Press.

Meehl, P. (1997). "The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical

predictions," in *What If There Were No Significance Tests?* eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah: Erlbaum), 393–425.
Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.
Putnam, H. (1983). *Realism and Reason*. Cambridge: Cambridge University Press.
Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *Am. Psychol.* 65, 1–12. doi: 10.1037/a0018326
Searle, J. R. (1999). *Mind, Language and Society: Doing Philosophy in the Real World*. London: Weidenfeld and Nicolson.
Sprenger, J. (2016). "Bayesianism vs. frequentism in statistical inference," in *The Oxford Handbook of Probability and Philosophy,* eds A. Hájek and C. Hitchcock (Oxford: Oxford University Press), 382–405.
Sprenger, J. (2018). The objectivity of subjective Bayesianism. *Eur. J. Philos. Sci.* 8, 539–558. doi: 10.1007/s13194-018-0200-1
Sprenger, J., and Hartmann, S. (2019). *Bayesian Philosophy of Science. Variations on a Theme by the Reverend Thomas Bayes*. Oxford: Oxford University Press.
Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$". *Am. Stat.* 73, 1–19. doi: 10.1080/00031305.2019.1583913

Conflict of Interest Statement: 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.

frontiers in Psychology

OPINION
published: 26 September 2018
doi: 10.3389/fpsyg.2018.01831

How to Crack Pre-registration: Toward Transparent and Open Science

*Yuki Yamada**

Faculty of Arts and Science, Kyushu University, Fukuoka, Japan

Keywords: QRP, misconduct in research, academic publishing, preregistration, open science

The reproducibility problem that exists in various academic fields has been discussed in recent years, and it has been revealed that scientists discreetly engage in several questionable research practices (QRPs). For example, the practice of *hypothesizing after the results are known* (HARKing) involves the reconstruction of hypotheses and stories after results have been obtained (Kerr, 1998) and thereby promotes the retrospective fabrication of favorable hypotheses (cf. Bem, 2004). P-hacking encompasses various untruthful manipulations for obtaining *p*-values less than 0.05 (Simmons et al., 2011). Such unethical practices dramatically increase the number of false positive findings and thereby encourage the intentional fabrication of evidence as the basis of scientific knowledge and theory, which leads to individual profits for researchers.

OPEN ACCESS

Edited by:
Hannes Schröter, German Institute for Adult Education (LG), Germany

Reviewed by:
Karin Maria Bausenhart, Universität Tübingen, Germany

***Correspondence:**
Yuki Yamada
yamadayuk@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *06 July 2018*
Accepted: *07 September 2018*
Published: *26 September 2018*

Citation:
Yamada Y (2018) How to Crack Pre-registration: Toward Transparent and Open Science. Front. Psychol. 9:1831. doi: 10.3389/fpsyg.2018.01831

BENEFITS OF PRE-REGISTRATION

Pre-registration is a remedy for this problem that involves the submission of research papers for which experimental and analytical methods, including researchers' motivation and hypotheses, have been designed and described completely prior to the collection of actual data (similar to proposal papers). The date on which the research was registered is also recorded. The associated manuscript cannot be modified after research has been registered. In reviewed pre-registration, manuscripts are peer-reviewed prior to registration, and only manuscripts that successfully pass this stage are registered and will be published, regardless of whether the collected data support the registered hypothesis (resulting in publications known as registered reports). It has been repeatedly argued that pre-registration can be a powerful approach for addressing prevalent QRPs (Miguel et al., 2014; Munafò et al., 2017; Nosek et al., 2018). For example, pre-registration can prevent or suppress HARKing, p-hacking, and cherry picking since hypotheses and analytical methods have already been declared before experiments are performed. In cases involving reviewed pre-registration, publication is guaranteed at the registration stage, thereby preventing the occurrence of QRPs. A previous study reported that more than 30% of psychological researchers admitted to the involvement of QRPs (John et al., 2012). Since the object of such researchers who engage in QRPs may be to publish as many research papers as possible, pre-registration eliminates the necessity for such QRPs.

Furthermore, registered reports undergo an additional peer-review stage not present in the conventional publication process. Peer review is conducted both at the time of registration and after results have been reported. The reviewed pre-registration process is relatively laborious for researchers since it requires receiving a decision of acceptance from a journal editor on at least two separate occasions. Therefore, registered reports are considered to be authentic, and research results consistent with postulated hypotheses can achieve greater credibility and approval.

MISUSE OF PRE-REGISTRATION

The preceding paragraphs provide a narrative about QRPs that can be effectively discouraged by pre-registration. However, a detailed examination of the current pre-registration system also reveals problems that this system cannot address. As mentioned, recognition of the value of pre-registration with respect to being able to confer reliability on research findings is becoming increasingly widespread. In terms of reputation management, researchers are motivated to improve their reputation regarding the credibility of their research (and themselves). A subset of researchers may attempt to misuse the pre-registration process to enhance their reputation even if their personality characteristics are not associated with readily engaging in data fabrication or falsification. Alternatively, certain situations may cause normal researchers to misuse this process on a momentary impulse (Schoenherr, 2015; Motyl et al., 2017). Their goals are to enhance the credibility of their research by pre-registering and to show the excellence of their hypothesis by presenting data that support that hypothesis.

There are methods for camouflaging a registered study as successful (van 't Veer and Giner-Sorolla, 2016). One such method is *selective reporting*, which is a type of data fabrication in which data that do not support the hypothesis are not reported (Goodman et al., 2016). Similarly, in the case of *infinite re-experimenting*, malicious researchers repeatedly perform the same experiment multiple times until the desired data to support the hypothesis are obtained and then report these data. Such QRPs cannot be completely prevented unless third parties can manage all of the data from experiments performed by researchers following registration. There is also a method that I call *overissuing*. Researchers who engage in overissuing pre-register a large number of experiments with extremely similar conditions and ultimately report only successful studies. This practice is difficult to discover by reviewers and editors who do not know a researcher's overall registration status; to date, this approach has not been explicitly identified as a QRP.

Another method is an approach that I call *pre-registering after the results are known* (PARKing). Researchers engaging in this practice complete an experiment (possibly with infinite re-experimentation) before pre-registering and write an introduction that conforms to their previously obtained results. Because such researchers apparently get attractive results and misrepresent those results as having been obtained under pre-registration, the research can readily acquire false credibility and impact. Rigorous initial peer-reviews that require revision of protocols may be able to reduce PARKing to some extent, but it is not effective if the malicious researchers involved engage in over issuing or target journals with poor peer-review practices. Furthermore, even if all unprocessed data are shared in a repository, the time stamps of uploaded data files can easily be forged or tampered with in various ways, such as by changing the system date for the operating system that is handling the data file. Therefore, there is currently no method for journals or reviewers to detect PARKing. Because many research resources would be required to implement the unethical methods described above, given the discarding of data that do not fit researchers' hypotheses, such methods can most easily be implemented by laboratories with abundant funds. If the aforementioned QRPs become rampant, their use could not only avoid decreases in false positives (which is a substantial advantage of pre-registration) but also accelerate the Matthew effect of rich people becoming richer (Merton, 1968).

It is easier to fabricate data and falsify results than to engage in *cracking* pre-registration; therefore, why should researchers attempt to crack pre-registration at all? The answer depends on the associated risk. Because data fabrication is a clear case of research misconduct and is subject to punishment, the risk associated with revelation is large. On the other hand, many of the cracking methods introduced here can be performed by simply extending general research practices. For example, suppose that a researcher conducted a paper-based questionnaire survey in the typical manner (without pre-registration) and had obtained significant results that supported his/her hypothesis and written a manuscript about this research. In this case, barriers to PARKing by using the introduction and method sections of the manuscript and subsequently publishing the full article appear to be low. Excel files for data aggregation can be recreated after pre-registration. If such cracking techniques have benefits that outweigh the difficulties and can be used with little risk, researchers who engage in these techniques will readily emerge.

BEYOND PRE-REGISTRATION

Therefore, pre-registration should not be overly trusted: it can easily be cracked. This paper introduces the idea that to prevent such cracking, registered research reports should not be completely believed just because "they were registered"; instead, several replications of the reported research with pre-registration should be performed. In addition, outsourcing experiments to multiple laboratories and agencies that do not share profitable interests with those of the registered researchers can be an effective means of preventing QRPs. If researchers outsourced experiments directly by themselves, some conflicts of interest could arise, so this process should be handled by journals who have received pre-registration protocols. In such cases, a journal would select an outsourcing partner for experimenting based on the protocol with the names of the original researchers being blinded. The candidate for outsourcing could either be selected by the journal in the same manner as reviewer selection or else crowdsourced. In either case, what matters is the precision of the experiment carried out, that this information is preserved for each candidate as a history, and that it can be used in the analysis or on subsequent request.

Preparing funds to implement this is a problem. It is to be hoped that financial support will be provided via various sources of funds based on the idea that such expenditure would help avoid the dissemination of numerous studies that involve the use of cracking approaches. Specifically as part of fraud countermeasure, universities or institutions to which the researchers belong could require them to underwrite this cost. Another way would be for the journal concerned to issue a coupon allowing the researchers entrusted with the experiments

to use it as a resource for their academic activity. Indeed, this system has already been proposed for peer review (Gurwitz, 2017). Hopefully, national/international funding agencies should support and manage outsourcing replication efforts. It would also be effective for fostering a normative standard for high-quality research.

We should change the "positive results = win" mode of thinking that is pervasive throughout the scientific community. An important consideration is transparency. The conventional philosophy of pursuing positive results shrouds research in a fog. How, then, do we bring about such transparent practices? The first step may be to disseminate the pre-registration system to the utmost (i.e., make it mandatory). This will shift the value of pre-registration from an ethical device for distinguishing between ethical and unethical studies to that of research transparency that clearly divides theoretical and empirical work. If all the research is pre-registered, the ethics of that research is not governed by the pre-registration itself. Therefore, at this point, the cracking methods mentioned earlier in this article will lose efficacy.

The second step is research dividing, a successor model to pre-registration. Here I propose a new idea that theoretical and empirical elements no longer need to coexist in one paper. That is, someone can write a paper covering only theoretical elements, while someone else can write a paper focusing solely on empirical material. In the former, the theoretical validity and appropriate hypothesis formation are evaluated; in the latter, appropriate experimentation and analysis are assessed. Detailed discussion will be carried out by those who write a paper on a theoretical issue that advances the previous theory based on those results. Indeed, some idea journals are already in existence (e.g., *Medical Hypotheses* and the *Frontiers* journal's "Hypothesis and Theory"). Such a division of research will promote replication studies as being more natural and easier to conduct. Currently, the hurdle for reviewed pre-registration is too high for many researchers to conduct replication studies. However, for papers focusing solely on empirical material, it would be possible to conduct replications without pre-registration.

If this second step were achieved, the need for QRPs and research misconduct would be reduced. The "positive results reign supreme" attitude in the science community would be discarded because it would not be the yardstick by which researchers would evaluate. As long as the scientific publication system itself is transparent, reliable, and ethical, individual research would not need to be concerned with evaluation of such aspects. The best way to crack pre-registration is to abandon the fixed idea of the structure of scientific articles.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The present study was supported by JSPS KAKENHI Grant Numbers JP15H05709, JP16H03079, JP16H01866, JP17H00875, JP18H04199, and JP18K12015.

ACKNOWLEDGMENTS

The author would like to express great appreciation to the editor and the reviewer for their insightful and constructive comments. The author would also like to thank Davood Gozli and Siqi Zhu for their extremely valuable suggestions on an earlier draft of this article.

REFERENCES

Bem, D. J. (2004). "Writing the empirical journal article," in *The Compleat Academic: A Career Guide 2nd Edn*, eds J. M. Darley, M. P. Zanna, and H. L. Roediger III (Washington, DC: American Psychological Association), 185–219.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Sci. Transl. Med.* 8:341ps12. doi: 10.1126/scitranslmed.aaf5027

Gurwitz, D. (2017). Award bonus points to motivate reviewers. *Nature* 542, 414. doi: 10.1038/542414d

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* 23, 524–532. doi: 10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Person. Soc. Psychol. Rev.* 2, 196–217. doi: 10.1207/s15327957pspr0203_4

Merton, R. K. (1968). The Matthew effect in science: the reward and communication systems of science are considered. *Science* 159, 56–63. doi: 10.1126/science.159.3810.56

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., et al. (2014). Promoting transparency in social science research. *Science* 343, 30–31. doi: 10.1126/science.1245317

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., et al. (2017). The state of social and personality science: rotten to the core, not so bad, getting better, or getting worse? *J. Person. Soc. Psychol.* 113, 34–58. doi: 10.1037/pspa0000084

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., et al. (2017). A manifesto for reproducible science. *Nature Hum. Behav.* 1, 1–9. doi: 10.1038/s41562-016-0021

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proc. Natl. Acad. Sci. U.S.A*, 18:201708274. doi: 10.1073/pnas.1708274114

Schoenherr, J. R. (2015). Social-cognitive barriers to ethical authorship. *Front. Psychol.* 6, 331–335. doi: 10.3389/fpsyg.2015.00877

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

van 't Veer, A. E., and Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *J. Exper. Soc. Psychol.* 67, 2–12. doi: 10.1016/j.jesp.2016.03.004

Conflict of Interest Statement: 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.

frontiers in Psychology

CONCEPTUAL ANALYSIS
published: 04 September 2019
doi: 10.3389/fpsyg.2019.01884

Confounds in "Failed" Replications

*Paola Bressan**

Dipartimento di Psicologia Generale, University of Padova, Padova, Italy

Reproducibility is essential to science, yet a distressingly large number of research findings do not seem to replicate. Here I discuss one underappreciated reason for this state of affairs. I make my case by noting that, due to artifacts, several of the replication failures of the vastly advertised Open Science Collaboration's *Reproducibility Project: Psychology* turned out to be invalid. Although these artifacts would have been obvious on perusal of the data, such perusal was deemed undesirable because of its *post hoc* nature and was left out. However, while data do not lie, unforeseen confounds can render them unable to speak to the question of interest. I look further into one unusual case in which a major artifact could be removed statistically—the nonreplication of the effect of fertility on partnered women's preference for single over attached men. I show that the "failed replication" datasets contain a gross bias in stimulus allocation which is absent in the original dataset; controlling for it replicates the original study's main finding. I conclude that, before being used to make a scientific point, all data should undergo a minimal quality control—a provision, it appears, not always required of those collected for purpose of replication. Because unexpected confounds and biases can be laid bare only after the fact, we must get over our understandable reluctance to engage in anything *post hoc*. The reproach attached to *p*-hacking cannot exempt us from the obligation to (openly) take a good look at our data.

Keywords: replication, confounds, good research practices, Open Science Collaboration, reproducibility project, mate preferences, ovulatory shift

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS), Italy

Reviewed by:
Brian D. Earp,
University of Oxford, United Kingdom
Ulrich Dettweiler,
University of Stavanger, Norway

***Correspondence:**
Paola Bressan
paola.bressan@unipd.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *23 February 2019*
Accepted: *31 July 2019*
Published: *04 September 2019*

Citation:
Bressan P (2019) Confounds in "Failed" Replications. Front. Psychol. 10:1884. doi: 10.3389/fpsyg.2019.01884

> *"Examine [the data] from every angle".*
>
> —Daryl J. Bem (1987, p. 172)

INTRODUCTION

Reproducibility may be crucial in science, but originality presents itself better. Thus, the activity of merely reproducing the work of others has been infrequent (Makel et al., 2012) and regarded with contempt. The spirit of the times has now briskly turned. We are in the midst of a movement that attaches increasing importance to repeating original studies—while loudly questioning the credibility of those findings that do not appear to replicate.

Yet the idea that we should trust a failed replication more than the original study is debatable. A failed replication—unless it has higher statistical power (Maxwell et al., 2015) or does a better job of meeting some implicit auxiliary assumption linking theory to observation (Trafimow and Earp, 2016)—is bound to be just as unreliable as the study it fails to replicate. An effect that truly exists in the world will not always prove "statistically significant" in a faithful replication of the original study; the *p* values produced by repeated simulations of the same experiment bounce around to a rather alarming extent ("the dance of the *p* values": Cumming, 2014; see also Stanley and Spence, 2014; Van Calster et al., 2018). That people would expect *p* values to stay put, naturally, scarcely helps them grasp what nonreplications (do not) entail—reinforcing the feeling of a replication "crisis" (Amrhein et al., 2019).

In this article I illustrate a complementary reason for being skeptical of failed replications: the effect may be there, but remain unseen due to the authors' well-meant unwillingness to treat the new data any differently than the original ones. The wholly understandable aversion to engaging in *post hoc* practices appears to have gone overboard. It is currently feeding the argument that, because "any well-designed study (e.g., an adequately powered study with appropriate measures) provides useful information regardless of the specific findings" (Johnson et al., 2014, p. 320), peer review is only needed before, and not after, data collection. Alas, the property of coming from a well-designed study does not automatically endow data with the distinction of providing useful information. Not only can an "adequately powered study with appropriate measures" produce nonsense, but crucially, there is no knowing ahead of time whether and how it will. We find out if something went *unexpectedly* wrong only by looking at the data (assuming, that is, we are lucky and the data will tell).

THE MILLION ROADS TO THE NULL EFFECT

I shall illustrate my point with actual cases taken from the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), whose results made it into *Science* and proceeded to gather nearly 3,000 citations in 3 years. This project attempted to replicate 100 studies published in 2008 in three respected psychology journals: *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Of the original studies whose results were significant, slightly over 60% failed to replicate—that is, yielded nonsignificant results ($p \geq 0.05$).

Here I showcase a few of the nonreplications that, unbeknownst to the public, turned out to be invalid. I am including only cases in which something amiss was found in the data themselves—rather than, or besides, the methods or analyses—and a response was written up about it. The problems were invariably caught by the original authors; links to each replication report and original authors' response are presented along the original article's reference, in the References list. Most such responses have been added to the corresponding replication record on the Open Science Framework platform. Still, they have not prompted corrections or updates to the replication's status (such as replacing "failed" with "invalid" or "inconclusive") and do not appear to do much else than sit there.

"Failed" Replication of Amodio, Devine, and Harmon-Jones (2008)

In a racial-stereotype inhibition task, people with low levels of prejudice did better when their motivations were only internal (e.g., when being nonprejudiced was personally important to them) rather than external too (e.g., when appearing nonprejudiced served to avoid disapproval from others) (Amodio et al., 2008). This more efficient inhibition of racial stereotypes reflected better stereotype control specifically, as opposed to better cognitive control in general.

The well-validated task used here to measure stereotype control consists in having people classify images of pistols, drills, and suchlike as either guns or tools. Right after seeing very briefly a Black (as opposed to White) face, people are more likely to classify correctly a gun than a tool; the larger one's tendency to do so, the weaker one's stereotype control is surmised to be. This normally observed effect was missing entirely in the replication data, rendering the task invalid as a measuring device. One possible reason is that, although the point was to examine Whites' racial biases toward Blacks, and the original sample included primarily White participants, the majority of participants in the replication sample turned out to be non-White.

"Failed" Replication of Campbell and Robert (2008)

In a practice phase, people repeatedly solved both multiplication problems (such as $7 \times 5 = ?$) and factoring problems (such as $6 = ? \times ?$) (Campbell and Robert, 2008). In the test phase, half of the participants were only asked to multiply and the other half to factor. People who were asked to multiply were faster at solving the problems they had previously practiced as multiplications ($7 \times 5 = ?$) than those they had practiced as factoring ($3 \times 2 = ?$). However, they were faster at problems they had practiced as factoring than at new multiplications altogether. The same result, in reverse, held for those who were asked to factor. Thus, although people did best with problems identical to those practiced earlier (as one would expect), cross-operation transfer was observed too; this was the important result.

In the replication, no evidence of transfer between multiplying and factoring was found. Curiously enough, in the Reproducibility Project database this replication is marked as successful—on the grounds that the significant interaction found in the original study was significant here too. This was a mistake, because (as also pointed out in the replication report itself) the data patterns that produced the interaction were different in the original and replication studies: a practice effect plus a cross-operation transfer in the original, just a practice effect in the replication.

Inspection of the replication data showed that participants failed to become much faster with practice, and during the practice phase continued to make a lot of errors (which, there being no feedback, remained uncorrected and hence did not

promote learning). After the entire set of 20 blocks of practice of the same eight problems, one of the two groups was still making 10 times as many errors as the corresponding group in the original study. Surely, one cannot expect much transfer of something that has not been learned in the first place.

"Failed" Replication of Monin, Sawyer, and Marquez (2008)

Moral rebels are individuals who refuse to comply when complying would compromise their values. According to Monin et al. (2008), people who do comply dislike rebels because their own obedient behavior is implicitly called into question by the rebel's behavior, and this threatens their self-confidence. If so, buttressing people's self-confidence should reduce their need to disparage moral rebels. Here, participants who had just completed a self-affirmation task (i.e., written an essay about a recent experience in which they demonstrated a quality that made them feel good about themselves) disliked moral rebels less than did participants who had completed a control task instead (i.e., listed foods consumed in the last 24 hours).

The crucial manipulation consisted in having participants write a long, mindful essay aimed at increasing their sense of being a good, worthy person. In the original study, this was an 8-minute composition written in the laboratory; the median number of words was 112. In the replication study, which was online, the amount of time participants were required to spend on the essay was not specified; the median number of words turned out to be 29, suggesting that most people had just rushed through the task. On top of that, being Monin et al.'s article about rebels' rejection by their peers, the target person portrayed as either a complier or a rebel ought to have been a peer: in the original study, it was a fellow student of the same age. However, the replication data revealed a median age difference of 15 years between participant and "peer" (consistent with this reduced similarity, the "peer" was liked less overall). I would personally add that, not coming from a peer, the target person's behavior might also have felt less directly relevant to the participant's self-image: less supportive in case of a complier, less threatening in case of a rebel.

"Failed" Replication of Schnall, Benton, and Harvey (2008)

In this study, participants were asked to judge the morality of hypothetical actions (for example, how wrong it would be to put false information on a résumé) (Schnall et al., 2008). People who had previously been exposed to words related to purity and cleanliness (or had washed their hands after watching a disgusting film clip) made more lenient moral judgments than people who had been exposed to neutral words (or had not washed their hands).

Perusal of the replication data disclosed that, across the various moral scenarios, a large percentage of responses was at the top of the scale ("extremely wrong": 41% vs. 28% in the original study). This implies that the lack of effect in the replication study may have resulted purely from lack of variance due to a ceiling effect (Schnall, 2014a,b). (But see Huang, 2014, for discussion of another variable—replication participants' low vs. high response effort—which would appear to be more critical.) The replicators downplayed Schnall's concerns on the grounds that "the distributions themselves provide valuable information for the field about the generalizability of the original findings" (Johnson et al., 2014, p. 320), but this is true only in a loose, uninteresting sense. The specific information they provide is that the original moral scenarios are sensitive to changes in context. They say nothing about the original findings themselves—which, with moral scenarios better suited to the replication sample (i.e., permitting as much variance as in the original study), could replicate just fine.

THE BEST MEN ARE (NOT ALWAYS) ALREADY TAKEN

In each of the cases just reviewed, the replication data were unable to speak to the question of interest and it was too late to do something about it. Amodio, Campbell, Monin, and Schnall had no way of showing that the failed replication would have been successful had the confounds not been there. The causal link between confounding variables and null effects was suspected but not proven.

To make a stronger argument, let us look again at the Reproducibility Project's original studies that failed to replicate. Here I pick yet another such study (Bressan and Stranieri, 2008) that appeared in *Psychological Science* and that I happen to know especially well, being its senior author. The reason why this case deserves closer attention than the others do is that, remarkably and uncommonly, some unconfounding of the confounded replication data turned out to be possible.

The study found that women's preference for faces of men described as single, relative to faces of men described as attached, depended on the ovulatory cycle. Higher-fertility women (those in the middle 2 weeks of their monthly cycle) preferred single men more than did lower-fertility women (those in the first and last week of their cycle). A significant interaction between fertility and women's relationship status indicated that the effect was specific to women who had a partner.

Bressan and Stranieri (2008) pointed out that the effect was consistent with the hypothesis of female dual mating (Pillsworth and Haselton, 2006; see also Thornhill and Gangestad, 2008; Gildersleeve et al., 2014). Over evolutionary history, some women may have benefitted from having their long-term partner raise a child they had conceived with a more attractive man. If the children of these arrangements turned out to be reproductively more successful than the children of women who never strayed (whatever their circumstances), this adaptation would have spread.

Note that the implication of this hypothesis is not that women gain from seeking extrapair partners—only a minority of women in a minority of circumstances would (e.g., Buss and Shackelford, 2008). The implication is, instead, that women have evolved to be able to flexibly implement this strategy *should* they find themselves in these particular circumstances. Indeed—not on moral, but on evolutionary grounds—extrapair mating ought not be pursued liberally. First, sex, especially with a stranger,

invariably involves the risk of infection or injury. Second, female adultery is punished, often harshly, in virtually every society (Buss, 2000). It follows that an adaptation to stray could have evolved only if the hazard brought fruit often enough.

Women, then, might be hardwired to find men more attractive when the odds of conceiving are higher rather than lower. Several lines of evidence suggest that indeed they do: being in the fertile window increases sexual desire for extrapair partners (e.g., Gangestad et al., 2002; Arslan et al., 2018; as can be seen by comparing these two works, evidence of whether this shift extends to in-pair partners is mixed). Single men are more available as extrapair partners than are already attached men. Thus, the effect of fertility on women's preference for single (over attached) men may be an adaptation that increases the benefits of adultery over its costs (Bressan and Stranieri, 2008).

In this article, I am not concerned with "defending" the hypotheses discussed by Bressan and Stranieri (2008); they are just hypotheses. What I care about is whether the main finding is replicable. The Reproducibility Project failed to replicate it across two experiments, one conducted in the laboratory and one online (Frazier and Hasselman, 2016). Although some minor results were replicated, no effect whatsoever of the ovulatory cycle on women's preferences for single men was found. Here I reanalyze these data and show that they contain unforeseen confounds that were absent in the original dataset. Once these confounds are controlled for, the data reveal the same pattern as those in the original study.

METHODOLOGICAL MATTERS

Participants

A total of 769 heterosexual, normally cycling women were included in the analyses (**Figure 1**). Original sample—Italian (Bressan and Stranieri, 2008): Italian ethnicity, median age 21 years, range 18–35. Lab replication sample—American (Frazier and Hasselman, 2016): mixed ethnicities, median age 18 years, range 16–46. Online replication sample—American (Frazier and Hasselman, 2016): mixed ethnicities, median age 21 years, range 18–34.

Participants' eligibility criteria, along with the manner variables were coded, were identical for the original and replication datasets and were the same as in Bressan and Stranieri (2008). In all datasets, each woman's cycle day had been standardized by dividing the number of days since the first day of her last menstrual period by her reported typical cycle length and multiplying the quotient by 28. Based on this index, women had been divided into a higher-conception-risk group (days 8–20) and a lower-conception-risk group (days 1–7 and 21–28). This subdivision (the "average midcycle rule": Lamprecht and Grummer-Strawn, 1996) has the advantage of creating two approximately equal groups. Note that, in Bressan and Stranieri's original dataset, standardized cycle days had been rounded to the nearest integer, so that, for example, a participant on day 7.7 (which rounds to 8) would be in the high-conception-risk group. In both replication datasets, on the opposite, it appears that standardized cycle days had not been rounded, so that a participant on day 7.7 would be in the low-conception-risk group. To render the data comparable, I adopted the least disruptive, most conservative choice, and avoided rounding in both the original and replication datasets. This reclassified as nonfertile four original-study participants (one partnered, three unpartnered) that had been treated as fertile in Bressan and Stranieri (2008). So, all and only women on days 8.0–20.0 were labeled as fertile (high-conception-risk group) in all three datasets.

Women who were taking hormonal contraceptives, were on a standardized cycle day larger than 28 (i.e., experiencing an abnormal ovulatory cycle), reported not being heterosexual, or failed to disclose their relationship status were excluded from all datasets; all other participants were included. In both datasets provided by the Reproducibility Project team (see Data Availability), exclusions had already been made. The lab replication dataset was used as is. The online replication dataset revealed errors in the calculation of women's cycle day; correcting them removed seven participants (see section "Coding errors in the replication datasets" for details). Note, however, that these corrections did not affect the results.

Stimuli

Twelve color photographs of faces of men of various degrees of attractiveness were presented in an album, one per page. Each photo was accompanied by one of four labels: "this person is single," "this person is in love," "this person has a girlfriend," and "this person is married." Four parallel albums were prepared so that each of the 12 faces could be paired, between subjects, with all four labels. Stimuli and albums were the same across the original and replication studies (see Data Availability). Stimuli were presented on paper in the original and lab replication studies, on a computer screen in the online replication study.

Procedure

In the original study, participants were asked to imagine being at a party (with their partner, if they had one) and seeing the man portrayed in the photograph. They read aloud the accompanying label and then rated the man's attractiveness on a scale from 0 (not at all attractive) to 10 (very attractive). The lab replication followed a similar procedure. The online replication's method was adapted to the different interface, and included a memory test for each face/label combination to make sure that the label had been read.

In the original study, after going through the photos, participants answered several questions (some of which were meant to provide information for an unrelated study on female competition) about themselves and their partner, if they had one. The original questionnaire was in Italian; replication participants filled in the exact same questionnaire in an English translation (see Data Availability).

LOOKING AT THE DATA

Inspection of the replication report (Frazier and Hasselman, 2016) and datasets (Data Availability) uncovered reporting errors in

Sample	All	Partnered / Unpartnered	Fertile / Nonfertile	Album A / B / C / D
Original	198	100	49	9
				14
				11
				15
			51	16
				11
				14
				10
		98	44	9
				10
				14
				11
			54	14
				16
				11
				13

Sample	All	Partnered / Unpartnered	Fertile / Nonfertile	Album A / B / C / D
Replication (combined)	571	221	104	19
				28
				24
				33
			117	30
				27
				30
				30
		350	154	46
				34
				46
				28
			196	45
				44
				48
				59

Sample	All	Partnered / Unpartnered	Fertile / Nonfertile	Album A / B / C / D
Replication (Lab)	263	95	45	6
				14
				13
				12
			50	17
				12
				12
				9
		168	75	20
				20
				23
				12
			93	24
				19
				22
				28

Sample	All	Partnered / Unpartnered	Fertile / Nonfertile	Album A / B / C / D
Replication (Online)	308	126	59	13
				14
				11
				21
			67	13
				15
				18
				21
		182	79	26
				14
				23
				16
			103	21
				25
				26
				31

FIGURE 1 | Number of partnered/unpartnered, fertile/nonfertile women who participated in the original (top-left panel), lab replication (bottom-left panel), and online replication (bottom-right panel) studies. The top-right panel presents the combined replication data, which along with the original data were used for my reanalyses. Each participant was shown one photo album out of four possible ones; the distribution of the four albums (A, B, C, D) across participants is indicated in the last column of each panel.

the analyses (one of omission, one of commission), coding errors in the dataset, and sources of random and of systematic noise (confounds). Yet it is important to note that it was the confounds, not the errors, that were responsible for the failure to replicate.

Reporting Errors in the Replication Analyses

Following Bressan and Stranieri (2008), the replication team averaged the attractiveness ratings for the three categories of attached men (married, with a girlfriend, and in love) for each participant. The preference for single men was computed as the mean rating given to single men minus the mean rating given to attached men. These measures had already been calculated for both replication datasets, and in my reanalyses I used them exactly as they appear in the Reproducibility Project's files (Data Availability).

The replication authors reported (Frazier and Hasselman, 2016) that, unlike in Bressan and Stranieri's original study, the interaction between man's availability (single, attached),

participant's conception risk (low, high), and participant's partnership status (partnered, unpartnered) was not significant ($F < 1$ in both the lab and online replications; repeated-measures ANOVAs). I reran their analyses on exactly the same data and in exactly the same way.

The lab replication analysis came out identical. In the online replication analysis report I found one error (surely a typo) and one remarkable omission. Along with the critical triple interaction ($p = 0.746$), the authors reported the following effects: partnership status ($p = 0.008$), conception risk ($p = 0.548$), and man's availability ($p = 0.091$; the corresponding F was misreported as 16.90 whereas it should have been 2.88). This list of significant and nonsignificant effects failed, however, to include the nearly significant interaction between conception risk and man's availability: $F(1, 314) = 3.71$, $p = 0.055$. As shown by separate ANOVAs, this interaction was due to the fact that fertile women liked single men better than attached men, $F(1, 139) = 6.62$, $p = 0.011$, whereas nonfertile women did not, $F < 1$. This effect is in the same direction as that found in the original study (where it was further qualified by the interaction with partnership status).

Coding Errors in the Replication Dataset

Inspection of the online replication data file (Data Availability) uncovered a systematic error in the calculation of women's cycle day. Day 1 (referring to participants on their first day of menstruation) had been miscoded as Day 0 and so on, so that all cycle-day values were off by 1. Correcting these data led to the reassignment of seven low-fertility women to the high-fertility group and of eight high-fertility women to the low-fertility group, and to the loss from the database of six women whose standardized cycle day of 28 turned out to be 29 (meaning that they were experiencing an abnormal ovulatory cycle). One further inclusion error was found: one participant had been assigned a negative cycle day, because the first day of her last menstrual period had been set in the future. Note, however, that neither the correction of the cycle-day values nor the removal of these seven participants had any bearing on the results.

Sources of Random Noise in the Replication Dataset

Inspection of the lab replication data file (Data Availability) revealed that: (1) 16 participants "arrived late/early, did not follow instructions, had previous knowledge of the study, etc"; (2) 39 participants "forgot to read labels, misread labels, gave ratings before reading labels, questioned labels, asked explicitly whether label should affect her rating"; and (3) 41 participants were "not paying attention, went through very fast, phone usage." None of these participants (89 overall, because a few fell in more than one category) had been excluded from the analyses run by the replication team. These sources of noise in the data (absent in the original study) were indeed hard to remove, because the study's statistical power would decrease substantially by dropping these participants en masse[1]. Given the arbitrariness of any decisions about which cases to exclude and which to include, I discarded none of them from my reanalyses either.

Some of the participants in the replication studies had given abnormally low ratings to their current partner's personality; these women may have been more interested in replacing him altogether than in having him raise their child. However, given that no outlier exclusions based on partner traits had been made by Bressan and Stranieri (2008), I did not make any in my present analyses of the replication data either. Note that the conception-risk effect reported below does become stronger if these outliers are removed; but not being the focus of the current paper, here this point is neither spelled out nor discussed further.

Sources of Systematic Noise in the Replication Dataset

Confounds: Album

Before rerunning the analyses on the corrected replication data, I checked for any relevant differences between the original and replication samples. I began by examining the distribution of the four albums across participants. Different participants saw different albums (with one-fourth of each sample's participants sharing a specific assortment of face/label combinations). However, because the 12 men whose pictures were used as stimuli had been deliberately chosen so as to cover different degrees of attractiveness (see Bressan and Stranieri, 2008), the three specific men labeled as "single" in each of the four albums were not equally attractive across albums. Hence, the choice of counterbalancing the face/label combinations represented an inevitable source of noise. Album had indeed a significant main effect on the preference for single men in all three datasets. Combining the datasets revealed a large overall preference for singles in two albums (A and C; both p's < 0.0001, one-sample t), a large overall preference for attached men in another (B; $p < 0.0001$), and no significant preferences in the remaining album (D; $p = 0.173$). (The pattern of these preferences across albums was the same for partnered and unpartnered women.)

In the original study, album did not interact with any of the other variables and contributed random noise only. In the replication study, however—presumably due to some quirk in the recruitment of participants—albums were not uniformly distributed across the various categories of relationship status and conception risk. Crucially (see top-right panel in **Figure 1**), the two albums with the most attractive single men turned out to have been overwhelmingly presented to fertile women

[1]Relative to the original study, the numerosity of the group of participants who should show the effect—partnered women—was only marginally higher in the online study and actually *lower* in the lab study (see **Figure 1**), which falls short of qualifying either study as a high-powered replication (e.g., Simonsohn, 2015). Any power calculations the replication team performed prior to data collection (based on the effect size obtained in the original study, a method that leads to inadequate statistical power *per se*: Maxwell et al., 2015; Simonsohn, 2015) were made pointless by the introduction of random and systematic noise that was absent in the original study—not just regular noise, alas, but participants who "forgot to read labels" and biased assignments of stimuli. To ensure sufficient power, the data of the two studies were therefore analyzed together, with type of study (lab replication, online replication) as a factor.

who were unpartnered (unpartnered: 92; partnered: 43; $X^2 = 8.42$, $p = 0.004$; the corresponding figure for the original study is $X^2 = 0.47$, $p = 0.493$), while the two albums with the least attractive singles were presented to equivalent numbers of unpartnered (62) and partnered (61) fertile women. Put differently, the least attractive single men had been shown more often to nonfertile (103) than to fertile (62) unpartnered women, whereas the most attractive singles had been shown to equivalent numbers of nonfertile (93) and fertile (92) unpartnered women.

The original study found that fertile *partnered* women preferred singles. The figures above show that, in the replication study, fertile *unpartnered* women got—by some turn of chance—to rate the best singles. This confound had the remarkable consequence of creating a spurious "fertility effect" for unpartnered women, in the same direction as the original fertility effect for partnered women. Therefore, it made it impossible to detect the original study's interaction between fertility and relationship status.

It appears indisputable, at this point, that any analysis of the *replication* data that addresses the effects of fertility and relationship status on preference for singles while neglecting the confound of album assignment is bound to deliver noise as an answer. Therefore, I kept track of the effect of album in all analyses (including, as a robustness check, the reanalyses of the original data).

Potential Confounds: Self-Confidence With Men

As mentioned earlier, the replication team did find (though it failed to report) a nearly significant interaction between women's conception risk and man's availability. Yet, unlike in the original study, these two factors did not participate in a triple interaction with women's partnership status. Once one simply controls for the bias in album assignments, as we will see, the interaction with partnership status becomes $p = 0.170$ (**Figure 2**), raising the question of whether it may have reached significance if only the replication study had been less messy. However, let us assume that the lack of interaction in the replication is to be taken at face value. The first point that comes to mind is then whether partnership status might have affected the American and Italian samples' women in different ways. In an exploratory rather than confirmatory spirit, I investigated this issue by using the participants' responses to the questionnaire (see Data Availability). Partnered and unpartnered women saw two different versions of the questionnaire; any shared questions about the partner referred to the current partner in the former case and to a hypothetical partner in the latter. I considered the only question that had been presented identically, and with the same meaning, to both partnered and unpartnered women. This was: "In relationships with the opposite sex, how self-confident are you?" Responses were given on a 1–5 scale (1 = not at all, 2 = a little, 3 = moderately, 4 = a lot, 5 = very much).

The distribution of these responses differed significantly between the original and replication datasets (Mann-Whitney U test, $p = 0.001$). Over 30% of women in either replication sample reported being more than moderately self-confident with men (responses 4 and 5: "a lot" and "very much"), as opposed to less than 20% of women in the original sample

		Fertility		Fertility x Partnered		Fertility x Self-confidence		Fertility x Partnered x Self-confidence	
Sample	**Control variables**	**Original**	**Replication**	**Original**	**Replication**	**Original**	**Replication**	**Original**	**Replication**
All women	None	**0.012**	**0.003**	**0.007**	**0.170**				
		(198)	(571)	(198)	(571)				
	Self-confidence	**0.011**	**0.124**	**0.002**	**0.078**	**0.356**	**0.743**	**0.281**	**0.015**
		(196)	(570)	(196)	(570)	(196)	(570)	(196)	(570)
Unpartnered women	None	**0.894**	**0.218**						
		(98)	(350)						
	Self-confidence	**0.661**	**0.876**			**0.128**	**0.139**		
		(96)	(349)			(96)	(349)		
Partnered women	None	**0.001**	**0.002**						
		(100)	(221)						
	Self-confidence	**0.000**	**0.013**			**0.918**	**0.036**		
		(100)	(221)			(100)	(221)		
Partnered non-self-confident	None	**0.036**	**0.771**						
		(29)	(113)						
Partnered self-confident	None	**0.001**	**0.002**						
		(71)	(108)						

FIGURE 2 | Visualization of the effect of fertility on women's preference for single relative to attached men. The figure presents the results of univariate ANOVAs; fertility is reported as a main effect and in its interaction with relationship status and/or self-confidence with men, whenever either factor appears in the analysis. The column "Control variables" indicates whether the ANOVA contains factors other than type of study, album, relationship status, and fertility. For each analysis and effect, the cell indicates the *p* value (rounded to the first three digits; *N* within brackets), separately for the original and the replication data. Significant results ($p < 0.05$) are shown in white on a dark background. Significant main effects that are qualified by an interaction are shown on a lighter gray background. Nonsignificant effects are shown in black on a white background.

(only responses 4; nobody chose the value 5). Critically, in the replication data the distribution of responses was consistently different for partnered and unpartnered women (Mann-Whitney U test, $p < 0.0005$ in both replication samples), unlike in the original data (Mann-Whitney U test, $p = 0.587$). Basically, in American women self-confidence with men was strongly associated with relationship status, being lower for unpartnered than for partnered women. This was not the case in Italian women, which might reflect a general cultural difference or merely a sample difference.

This divergence between the original and replication studies was especially disturbing because in the replication data, unlike in the original data, participants' self-confidence with men interacted not only with relationship status, as mentioned above, but also with conception risk and album. Strikingly, for example, among extremely self-confident partnered women (response 5) the albums with the most attractive single men had been shown nearly exclusively to those who were nonfertile (nonfertile: 14; fertile: 1), while the least attractive singles were presented to equivalent numbers of nonfertile and fertile women (nonfertile: 10; fertile: 9).

Because a woman's self-confidence with men is likely to increase the extent to which she perceives a man to be available to her, this set of asymmetries created a serious potential confound that was absent in the original data. For this reason, data were analyzed both with and without considering participants' self-confidence with men—median-split into "low" and "high"[2]—as a factor in the ANOVA. As a robustness check, this was done for both the original and replication data.

Note that this confound oddly complements and compounds those identified previously. In sum, albums were poorly allocated across (1) partnered and unpartnered fertile women; (2) fertile and nonfertile unpartnered women; (3) fertile and nonfertile self-confident partnered women. All misallocations tended to spuriously increase the ratings given to single men by unpartnered fertile women and/or decrease the ratings given to single men by the most self-confident partnered fertile women. Thus, each confound biased the data in the same direction—opposite to the original result.

"FAILED REPLICATION" DATA REANALYSIS

Results

What is at issue here is not how much confidence we should place in the original finding, but whether the Reproducibility Project did indeed, as claimed, fail to replicate it. Hence, I will not be evaluating the magnitude of the effect, the strength of the evidence for it, or the likelihood that the hypothesis is "true"—these matters are all beside the point. Instead, I will be running the very same analyses, only correcting for confounds, and adopting the very same rules and statistical standards as the Reproducibility Project did—whether or not these are the wisest. And because the criterion used to judge success or failure in the Reproducibility Project replications was the presence or absence of statistical significance, this is the criterion I will use, too.

The main analysis reported in Bressan and Stranieri's original study was a repeated-measures ANOVA on attractiveness ratings with a within-subjects factor of man's availability (single, attached). For simplicity, it is replaced here with a univariate ANOVA on preferences for single men relative to attached men; the two analyses (repeated-measures on a two-level within-subject variable and univariate on the difference between such levels) are conceptually identical and give identical results.

The fixed factors were album (A, B, C, D), participant's partnership status (partnered, unpartnered), and participant's conception risk (low, high). The same univariate ANOVA was run on both the original data and the combined replication data (see text footnote 1). In the latter, type of study (lab replication, online replication) was also added as a factor. Interactions were explored by stratifying the data (by partnership status, as in Bressan and Stranieri (2008), whenever this variable participated in the interaction) and repeating the ANOVA within each subgoup. All ANOVAs were run with and without the potential confounder of self-confidence with men (below the median, above the median) as an additional fixed factor.

For reasons of transparency and completeness of information, main and interaction effects that were significant in one sample and not in the other were explored in both, and all results are reported[3]. **Figure 2** presents the p values of all effects, separately for the original and replication studies.

Conception risk was significant as a main effect in both the original[4] and replication data (**Figure 2**, row 1: compare cells 1 and 2). Overall, higher-fertility women preferred single over attached men more than did lower-fertility women. In the original sample, the main effect of fertility was qualified by a significant interaction with partnership status, whether or not self-confidence with men was added to the analysis (row 1, cell 3; row 3, cell 3). In the replication sample, the main effect of fertility was qualified by a significant interaction with partnership status *and* self-confidence with men (row 3, cell 8).

To unpack these interactions, data were stratified by partnership status and fed into separate ANOVAs. In partnered women, fertility was always significant whether or not self-confidence

[2]Owing to the different distribution of responses, median splits—i.e., those that created two groups as close as possible in numerosity—turned out to be different in the original and replication samples. Both replication samples: responses 1-2-3 vs. 4-5. Original sample: responses 1-2 vs. 3-4 (nobody chose the value 5). Note that using subdivisions other than the median split would result in extremely unequal cell sizes. Overall, for example, the middle point of the scale (response 3) was chosen by nearly half of the participants; responses 1 and 5 were chosen by only 3% of, respectively, partnered and unpartnered women.

[3]The only results not reported in **Figure 1** are the main and interaction effects of type of study and album, which have no bearing on the ovulatory shift hypothesis; they can be found in the analysis outputs (Data Availability).

[4]The significant main effect of fertility on preference for single men corresponds to the significant interaction between conception risk and man's availability in Bressan and Stranieri's (2008) repeated-measures ANOVA. Following ANOVA reporting conventions, this effect was not originally reported because it was further qualified by the interaction with participant's partnership status.

with men was taken into account (rows 9 and 11: compare cells 1 and 2); in unpartnered women, it never was (rows 5 and 7: compare cells 1 and 2). This was true in both the original and replication studies. In the latter, the significant effect of fertility in partnered women was further qualified by a significant interaction with self-confidence with men (row 11, cell 6). Exploring this interaction showed that the effect was entirely driven by self-confident women (row 15, cell 2; see **Figure 3**).

Discussion

Failures to replicate can certainly suggest that the original findings emerged by chance, but we should contemplate that eventuality only after we have made an honest effort to understand whether discrepancies may have arisen from *other* causes, be they trivial or interesting. In this case, the cause was trivial: a significantly biased allocation of the face/label combinations used as stimuli. If this confound is controlled for in the analyses, the main result of the original study *is* replicated. In the original Italian sample, being fertile raised partnered women's attraction to single, relative to attached, men (p = 0.001). In the (albeit much noisier) American replication sample, it did too (p = 0.002).

In both the original and replication studies, the effect was significant for partnered women and not for unpartnered women. However, in the original study the *difference* between partnered and unpartnered women was significant as well (p = 0.007), whereas in the replication study it was not (p = 0.170). An obvious reason could be that the replication data were simply too noisy for the interaction to emerge. A conceptually more interesting possible reason concerns a variable that was irrelevant in the original sample but relevant and confounded in the replication sample. The role of this variable (self-confidence with men) was unpredicted, hence this finding should be interpreted as exploratory—a potential factor to track in future research. In the replication but not in the original sample, being partnered strongly covaried with feeling self-confident with men, and feeling self-confident with men was confounded with both face/label allocation and conception risk. Controlling for self-confidence with men replicated the original significant difference between partnered and unpartnered women. The notion that self-confidence with men could play a role is far from counterintuitive: lack of self-confidence may decrease a woman's perceived chances of success in pursuing an extrapair man, or increase her fear that pursuing an extrapair man could endanger her relationship with her current partner. Still, the effect of fertility on partnered women's preferences for singles was also significant overall, fully replicating the original finding even if the unanticipated differences in self-confidence between participants are *not* taken into account.

It is important to note that, in principle, the failure to consider the peculiar distribution of women's self-confidence in the American sample might have hampered replicability entirely, and we would be none the wiser. And yet, the original paper could not possibly have alerted future replicators about the importance of this variable: self-confidence with men was strongly associated with having a partner (and again, confounded with album) in the replication sample, but was unrelated to it (and unconfounded) in the original sample.

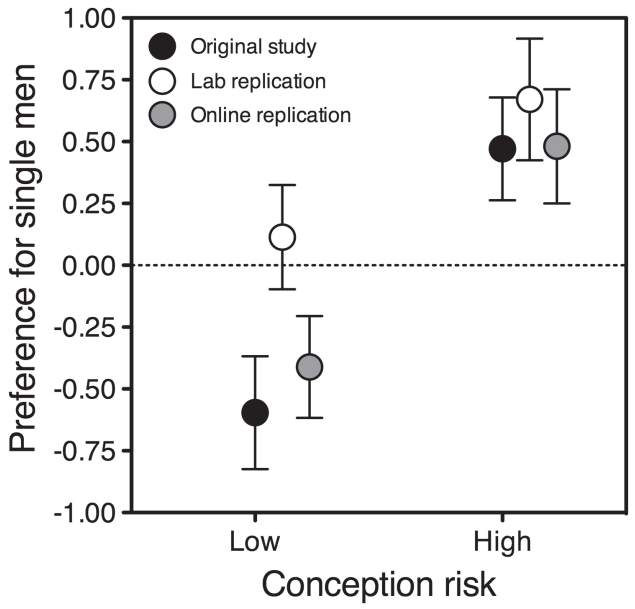

FIGURE 3 | Preference for single over attached men as a function of conception risk, in partnered women who reported above-the-median self-confidence with men. The ovulatory shift is represented by the difference between each symbol on the left and the corresponding symbol on the right. Symbols depict estimated marginal means adjusted for the effect of album; error bars indicate the standard error of the mean. Black symbols: original study, N = 71; gray symbols: online replication, N = 62; white symbols: lab replication, N = 46.

FREQUENTLY ASKED QUESTIONS

In the spirit of open debate I report, with his consent and nearly verbatim, some critical comments made by Ruben Arslan in a signed review of a previous version of this paper. Other readers might easily entertain similar doubts; they sound reasonable but are, I will argue, misplaced. I respond to them here.

> *1. The replication data produce the same result as the original data only when two post-hoc-plausible decisions are made. This is a perfect illustration of the problems that led to the reproducibility crisis in psychology. Adjusting for the imbalance in conditions is reasonable, but that alone does not turn the effect significant.*

The original study's major finding was the effect of cycle on partnered women's preference for single over attached men. Adjusting for the imbalance in conditions *is* enough to replicate it. Thus, a more appropriate conclusion is that the significant effect for partnered women was replicated, and (although an effect emerged only for partnered and not for unpartnered women)

the significant *difference* between partnered and unpartnered women was not.

One may stop there and learn nothing; or wonder why, and perhaps learn something (see also Stroebe and Strack, 2014; Van Bavel et al., 2016; Penders and Janssens, 2018). Here the failure of the $p = 0.170$ interaction to attain significance may very well have been due to the low power of a messy study, or even simply to basic sampling error and random measurement error (Stanley and Spence, 2014), but suppose for the sake of argument that there is a "real" difference between the original and the replication results. To move on, we must look at the data. Self-confidence with men was distributed differently in the original and replication samples, and (only) in the latter it was confounded with partnership status, conception risk, and face/label combinations. I looked exclusively at self-confidence with men because it happened to be the only question that was presented identically to all women. Yet if there had been 10 such theoretically meaningful questions, and one investigated them all to identify those potentially responsible for the difference between the outcomes of the original and replication studies, that would be perfectly rational: what I would ask is that this is done in the open and that the new "findings" are explicitly treated as exploratory. With their help, we may work out better hypotheses, to be tested in future—possibly, preregistered—studies.

> *2. The author's reaction to a nonreplication of her work is to double down on her initial interpretation and reanalyze the data following the Bem advice that has become known as a recipe for overfitting.*

My reaction to a nonreplication of my work is not to prove that my interpretation was correct or my findings "true" (I am in no position to know whether they are), but to understand why the original and replication studies produced different results. Until the day when this attempt to understand is expected from who has failed to replicate—and replication studies and datasets are examined for obvious confounds as scrupulously as original studies and datasets should—the burden is going to fall, alas, on the original authors.

Of Bem's (1987) otherwise unfortunate recommendations, one should not be dismissed with the rest and it is the only one I have followed here: look at the data. The nonreplication has prompted me, before all else, to reanalyze *my* old data to check to which extent the results I obtained depended on the analytic choices I made (as per Steegen et al., 2016). Even though the multiverse of possible choices in such a complex study is inevitably too large for present-day comfort, the original results have turned out to be remarkably robust—and this includes plausible alternative classifications of participants into high- and low-fertility groups. In fact, the results came out stronger using the stricter window (days 10–15) defined as "peak fertility" in Gildersleeve et al.'s (2014) meta-analysis.

Incidentally, I am not inclined to take this finding as additional evidence for the ovulatory shift hypothesis. Psychology studies typically rely upon relatively small samples. If indeed the fertile window falls entirely between days 10 and 17 in only 30% of women (Wilcox et al., 2000), its average position may be expected to vary even widely from one small sample to the next. Thus, finding the strongest effect for days 10–15 is not necessarily more persuasive than finding it for days 7–14 or 8–20.

> *3. Without realizing it, the author is doing what she criticizes herself: wander through the garden of forking paths.*

The "garden of forking paths" (Gelman and Loken, 2013) refers to the idea that the route toward statistical significance appears predetermined but is in fact the result of a hidden chain of choices that, albeit defensible and made in good faith, are arbitrary. Alternative data can lead to equally reasonable alternative analyses and equally reasonable ways to support the research hypothesis; "significant" patterns are thus perpetually revealed in what is actually noise. Very true. But because making reasonable choices cannot be avoided, the only moral is that we should not be so sure of our findings.

In this paper I have not tried out different data-cleaning, data-coding, and/or data-analytic alternatives in the attempt to produce the original results from the replication dataset. And as far as I am aware, I have not made any "reasonable choices" that had not been made in the original study either: I merely checked no obvious confounds had been introduced. I found at least a major one, concerning stimulus allocation, in the replication sample (but not in the original sample). Controlling for it in the analysis revealed a significant cycle shift in preference for single men among partnered women; this shift was in the same direction as reported by Bressan and Stranieri (2008) and replicated their main result.

For exploration purposes, I also showed that controlling for another likely confound (self-confidence with men) replicates their secondary result too. It should be clear that labeling this as a confound rests on the assumption that a woman's self-confidence with men affects her probability to become involved in an extrapair liason: a reasonable assumption in a world of alternative reasonable assumptions—one path in the garden of forking paths. Even worse, one taken in the context of the replication study, a dismally noisy and confounded dataset.

Of course, *both* the original result and its replication could just be side effects of phenomena unrelated to the hypothesis; or—far from impossible, considering how imprecise all these measures, most notably fertility ones (Wilcox et al., 2000), are bound to be—they might be plain noise themselves. The original findings of Amodio et al. (2008), Campbell and Robert (2008), Monin et al. (2008), and Schnall et al. (2008)—and even the findings "successfully" replicated by the Reproducibility Project, for that matter—might all turn out to be false positives. Yet this is irrelevant to the point I wish to make: let us check whether our data contain obvious confounds before doing anything with them. And if openly controlling for a *demonstrated* confound (not simply a *possible* or *plausible* confound) is now to be considered as a discretionary, arbitrary choice in data analysis, well, we should think again.

CODA: LET US HUNT FOR ARTIFACTS

Undisclosed flexibility in data coding and analysis may be responsible for the better part of the replication "crisis" in psychological (and nonpsychological: Begley and Ellis, 2012; Camerer et al., 2016) research (Simmons et al., 2011). Bem (1987) famously encouraged the beginning social scientist to examine the data from every angle; and then, to "cut and polish" the dataset and "craft the best setting for it" as though it were a jewel. Advice of that description tends now to be less popular than it once was. We have become aware that decisions as minor as whether or not to remove outliers, or include a certain factor in the analysis, are capable of swaying a study's results to the extent of reversing statistical significance (Steegen et al., 2016). Typically, such choices are not portrayed as arbitrary and any alternatives remain hidden. No discretionary paths were taken here. All coding, processing, and analytic choices were identical to those in Bressan and Stranieri (2008); the replication datasets were analyzed as they were provided by the Reproducibility Project. Only transparently verifiable errors and confounds were, respectively, corrected and controlled for. In the interest of cross-validation, each new analysis run on the replication data was repeated identically on the original data. All outcomes are reported.

The potential role played by methodological or statistical problems in purported failures to replicate has been voiced before (see Zwaan et al., 2018), although it appears that the original authors' viewpoints struggle to be heard, confined as they often are to blogs and comment sections. The particular case I have dissected here stands out from the rest in that the major confound could not only be identified but also controlled for—revealing results that were similar to those reported in the original study. More often (as in all the other cases I have illustrated), methodological confounds are impossible to control after the data have been collected, but that is exactly the stage when they tend to be found. Hardly everything that could possibly go wrong with a study can be predicted ahead of time, even when the study has been meticulously laid out and has received all necessary blessings. And although data are expected to be scrutinized by both authors and peer reviewers before being added to the published record, it appears that no after-the-fact quality control is required of data collected for purpose of replication (see also Schnall, 2014a,b). After-the-fact *any*thing is bunched together with questionable research practices. Johnson et al. (2014) dismissed Schnall's (2014a) exposure of a ceiling effect in their data as "hunting for artifacts." But hunting for artifacts is precisely what we should all do before taking our data seriously. If our data *are* the result of artifacts, they carry no evidentiary value; we should dispose of them (well, store them away) and start afresh.

One is left to wonder how many replications "fail" (and also, of course, how many original studies "succeed") solely because one has not bothered to look carefully at the data. No help will be forthcoming from preregistrations and similar declarations of intent—because whether a replication has failed owing to an unpredictable stimulus misallocation, or an accidental recruitment quirk, or an unexpected sample difference, can be established only *post hoc*. None of us gets a kick out of establishing things *post hoc*. Still, if we are really curious about the truth—as opposed to just craving to prove a point—it might be best to have a good look at the data; yes, to examine them (in full public view) from every angle.

DATA AVAILABILITY

All data, materials, analysis scripts and outputs, along with the Supplementary Material file, are publicly available via the Open Science Framework and can be accessed at https://osf.io/amrwu.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I am grateful to Becca Frazier for redoing our original study and to Ruben Arslan for allowing me to report his objections. Thanks, as always, to Peter Kramer for criticism and encouragement.

REFERENCES

Amodio, D. M., Devine, P. G., and Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: the role of conflict monitoring and neural signals for control. *J. Pers. Soc. Psychol.* 94, 60–74. doi: 10.1037/0022-3514.94.1.60 (Replication report by Johnson, Hayes, and Graham (2014): https://osf.io/ysxmf. Original authors' commentary on the replication (2015): https://osf.io/a5hdb).

Amrhein, V., Trafimow, D., and Greenland, S. (2019). Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *Am. Stat.* 73, 262–270. doi: 10.1080/00031305.2018.1543137

Arslan, R. C., Schilling, K. M., Gerlach, T. M., and Penke, L. (2018). Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *J. Pers. Soc. Psychol.* doi: 10.1037/pspp0000208

Begley, C. G., and Ellis, L. M. (2012). Drug development: raise standards for preclinical cancer research. *Nature* 483, 531–533. doi: 10.1038/483531a

Bem, D. J. (1987). "Writing the empirical journal article" in *The compleat academic: A practical guide for the beginning social scientist*. eds. M. P. Zanna and J. M. Darley (New York: Random House), 171–201.

Bressan, P., and Stranieri, D. (2008). The best men are (not always) already taken: female preference for single versus attached males depends on conception risk. *Psychol. Sci.* 19, 145–151. doi: 10.1111/j.1467-9280.2008.02060.x

Buss, D. M. (2000). *The dangerous passion: Why jealousy is as necessary as love and sex*. New York: Free Press.

Buss, D. M., and Shackelford, T. K. (2008). Attractive women want it all: good genes, economic investment, parenting proclivities, and emotional commitment. *Evol. Psychol.* 6, 134–146. doi: 10.1177/147470490800600116

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* 351, 1433–1436. doi: 10.1126/science.aaf0918

Campbell, J. I. D., and Robert, N. D. (2008). Bidirectional associations in multiplication memory: conditions of negative and positive transfer. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 546–555. doi: 10.1037/0278-7393.34.3.546 (Replication report by Riker and Saide (2015): https://osf.io/bux7k. Original author's commentary on the replication (2015): https://osf.io/mvudt).

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Frazier, R. S., and Hasselman, F. (2016). *Replication of Study 3 by Bressan & Stranieri (2008, Psychological Science)*. Available at: https://osf.io/7vriw (Accessed May 18, 2018).

Gangestad, S. W., Thornhill, R., and Garver, C. E. (2002). Changes in women's sexual interests and their partners' mate-retention tactics across the menstrual cycle: evidence for shifting conflicts of interest. *Proc. Biol. Sci.* 269, 975–982. doi: 10.1098/rspb.2001.1952

Gelman, A., and Loken, E. (2013). *The Garden of Forking Paths: Why Multiple Comparisons Can be a Problem, Even When There is no "Fishing Expedition" or "p-Hacking" and the Research Hypothesis was Posited Ahead of Time*. Available at: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf (Accessed July 1, 2018).

Gildersleeve, K., Haselton, M. G., and Fales, M. R. (2014). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychol. Bull.* 140, 1205–1259. doi: 10.1037/a0035438

Huang, J. L. (2014). Does cleanliness influence moral judgments? Response effort moderates the effect of cleanliness priming on moral judgments. *Front. Psychol.* 5:1276. doi: 10.3389/fpsyg.2014.01276

Johnson, D. J., Cheung, F., and Donnellan, M. B. (2014). Hunting for artifacts: the perils of dismissing inconsistent replication results. *Soc. Psychol.* 45, 318–320. doi: 10.1027/a000001

Lamprecht, V. M., and Grummer-Strawn, L. (1996). Development of new formulas to identify the fertile time of the menstrual cycle. *Contraception* 54, 339–343. doi: 10.1016/S0010-7824(96)00202-8

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012). Replications in psychology research: how often do they really occur? *Perspect. Psychol. Sci.* 7, 537–542. doi: 10.1177/1745691612460688

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

Monin, B., Sawyer, P. J., and Marquez, M. J. (2008). The rejection of moral rebels: resenting those who do the right thing. *J. Pers. Soc. Psychol.* 95, 76–93. doi: 10.1037/0022-3514.95.1.76 (Replication report by Holubar (2015): https://osf.io/a4fmg. Original author's commentary on the replication (2015): https://osf.io/3s2zd).

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aaa5612

Penders, B., and Janssens, A. C. J. W. (2018). Finding wealth in waste: irreplicability re-examined. *BioEssays* 1800173. doi: 10.1002/bies.201800173

Pillsworth, E. G., and Haselton, M. G. (2006). Women's sexual strategies: the evolution of long-term bonds and extrapair sex. *Annu. Rev. Sex Res.* 17, 59–100. doi: 10.1080/10532528.2006.10559837

Schnall, S. (2014a). Clean data: statistical artifacts wash out replication efforts. *Soc. Psychol.* 45, 315–317. doi: 10.1027/1864-9335/a000204

Schnall, S. (2014b). *Social Media and the Crowd-Sourcing of Social Psychology*. Available at: www.psychol.cam.ac.uk/cece/blog (Accessed November 7, 2018).

Schnall, S., Benton, J., and Harvey, S. (2008). With a clean conscience: cleanliness reduces the severity of moral judgments. *Psychol. Sci.* 19, 1219–1222. doi: 10.1111/j.1467-9280.2008.02227.x (Replication report by Johnson, Cheung, and Donnellan (2014): https://osf.io/maq28. Original author's commentary on the replication: see Schnall, 2014a,b).

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366. doi: 10.1177/0956797611417632

Simonsohn, U. (2015). Small telescopes: detectability and the evaluation of replication results. *Psychol. Sci.* 26, 559–569. doi: 10.1177/0956797614567341

Stanley, D. J., and Spence, J. R. (2014). Expectations for replications: are yours realistic? *Perspect. Psychol. Sci.* 9, 305–318. doi: 10.1177/1745691614528518

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* 11, 702–712. doi: 10.1177/1745691616658637

Stroebe, W., and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9, 59–71. doi: 10.1177/1745691613514450

Thornhill, R., and Gangestad, S. W. (2008). *The evolutionary biology of human female sexuality*. New York: Oxford University Press.

Trafimow, D., and Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: comment on Klein. *Theory Psychol.* 26, 540–548. doi: 10.1177/0959354316637136

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., and Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proc. Natl. Acad. Sci. U. S. A.* 113, 6454–6459. doi: 10.1073/pnas.1521897113

Van Calster, B., Steyerberg, E. W., Collins, G. S., and Smits, T. (2018). Consequences of relying on statistical significance: some illustrations. *Eur. J. Clin. Investig.* 48:e12912. doi: 10.1111/eci.12912

Wilcox, A. J., Dunson, D., and Baird, D. D. (2000). The timing of the "fertile window" in the menstrual cycle: day specific estimates from a prospective study. *BMJ* 321, 1259–1262. doi: 10.1136/bmj.321.7271.1259

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behav. Brain Sci.* 41:e120. doi: 10.1017/S0140525X17001972

Conflict of Interest Statement: 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.

frontiers in Psychology

METHODS
published: 21 December 2018
doi: 10.3389/fpsyg.2018.02599

Quantitative Data From Rating Scales: An Epistemological and Methodological Enquiry

*Jana Uher**

London School of Economics and Political Science, London, United Kingdom

Rating scales are popular methods for generating quantitative data directly by persons rather than automated technologies. But scholars increasingly challenge their foundations. This article contributes epistemological and methodological analyses of the processes involved in person-generated quantification. They are crucial for measurement because data analyses can reveal information about study phenomena only if relevant properties were encoded systematically in the data. The Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm) is applied to explore psychological and social-science concepts of measurement and quantification, including representational measurement theory, psychometric theories and their precursors in psychophysics. These are compared to theories from metrology specifying object-dependence of measurement processes and subject-independence of outcomes as key criteria, which allow tracing data to the instances measured and the ways they were quantified. Separate histories notwithstanding, the article's basic premise is that general principles of scientific measurement and quantification should apply to all sciences. It elaborates principles by which these metrological criteria can be implemented also in psychology and social sciences, while considering their research objects' peculiarities. Application of these principles is illustrated by quantifications of individual-specific behaviors ('personality'). The demands rating methods impose on data-generating persons are deconstructed and compared with the demands involved in other quantitative methods (e.g., ethological observations). These analyses highlight problematic requirements for raters. Rating methods sufficiently specify neither the empirical study phenomena nor the symbolic systems used as data nor rules of assignment between them. Instead, pronounced individual differences in raters' interpretation and use of items and scales indicate considerable subjectivity in data generation. Together with recoding scale categories into numbers, this introduces a twofold break in the traceability of rating data, compromising interpretability of findings. These insights question common reliability and validity concepts for ratings and provide novel explanations for replicability problems. Specifically, rating methods standardize only data formats but not the actual data generation. Measurement requires data generation processes to be adapted to the study phenomena's properties and the

OPEN ACCESS

Edited by:
Ulrich Dettweiler,
University of Stavanger, Norway

Reviewed by:
Martin Junge,
University of Greifswald, Germany
Barbara Hanfstingl,
Alpen-Adria-Universität Klagenfurt, Austria
Jennifer Hofmann,
University of Zürich, Switzerland

***Correspondence:**
Jana Uher
mail@janauher.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *14 May 2018*
Accepted: *03 December 2018*
Published: *21 December 2018*

Citation:
Uher J (2018) Quantitative Data From Rating Scales: An Epistemological and Methodological Enquiry. Front. Psychol. 9:2599. doi: 10.3389/fpsyg.2018.02599

measurement-executing persons' abilities and interpretations, rather than to numerical outcome formats facilitating statistical analyses. Researchers must finally investigate how people actually generate ratings to specify the representational systems underlying rating data.

Keywords: qualitative-quantitative integration, observational methods, assessment methods, transdisciplinary approach, quantitative methods in the social sciences, measurement, quantification, data

INTRODUCTION

Quantifications are central to many fields of research and applied settings because numerical data allow to analyze information using the power of mathematics (Chalmers, 2013; Porter, 1995; Trierweiler and Stricker, 1998). In psychology and social sciences (e.g., education, sociology, political science), quantitative data are often generated with rating methods in which persons indicate their judgments of predefined statements on multi-stage scales (e.g., standardized assessments, surveys or questionnaires). Rating scales are also used in many applied sectors (e.g., government, business, management, industry, public media) to help answer key questions, make decisions and develop strategies, such as for national policies, health programs, personnel selection and marketing (Menon and Yorkston, 2000; Abran et al., 2012; Hammersley, 2013). Accurate quantifications are thus critically important.

Increasing Criticism of Rating Scales

The strong reliance on rating methods is, however, increasingly criticized (Baumeister et al., 2007; Fahrenberg et al., 2007; Grzyb, 2016; Doliński, 2018). Scholars from various disciplines scrutinize their underlying epistemologies and measurement theories (Wagoner and Valsiner, 2005; Trendler, 2009; Vautier et al., 2012; Hammersley, 2013; Bringmann and Eronen, 2015; Buntins et al., 2016; Tafreshi et al., 2016; Bruschi, 2017; Humphry, 2017; Valsiner, 2017; Guyon et al., 2018). These developments are still largely unnoticed by mainstream psychologists who currently focus on the replication crisis, which they aim to solve by scrutinizing the epistemological foundations of significance testing, confidence interval estimations and Bayesian approaches (Nosek et al., 2015; Open Science Collaboration, 2015; Wagenmakers et al., 2016; Zwaan et al., 2017)—thus, by improving issues of data *analysis*.

But processes of data *generation* are largely understudied. Discussions are limited to debates about so-called 'qualitative' versus 'quantitative' methods, a common polarization suggesting some methods could be quantitative but not qualitative, and vice versa. Previous debates revolve around irreconcilable differences in underlying epistemologies (e.g., constructivist versus naïve-realist). To balance their respective advantages and disadvantages, both methods are combined in mixed-method designs (Creswell, 2003). But the methodological foundations of the operational procedures by which 'quantitative' and 'qualitative' data are generated are hardly discussed. Specifically, so-called 'quantitative' data are commonly generated by lay people who may be largely unaware of the positivist epistemology underlying the scales they are ticking. But even if they knew, what would this tell them about how to generate data? Likewise, laypeople are commonly unfamiliar with measurement theories. So how can they, by intuitively judging and ticking scales, produce data that meet the axioms of quantity and measurement? What considerations and decisions must raters actually make to justify interpretation of rating outcomes as 'quantitative' data? And in what ways do scientists' axioms and theories of measurement inform raters' decisions? Current debates are surprisingly silent about these fundamental issues.

Problematic findings with rating scales increasingly emerge. On widely used Big Five personality scales, differences between student and general public samples varied substantially and randomly across 59 countries, showing that, contrary to common assumptions, student findings cannot be generalized (Hanel and Vione, 2016). The empirical interrelations among ratings items used to assess the same personality factor (e.g., 'outgoing' and 'not reserved' for Extraversion) varied unsystematically across 25 countries, averaging around zero (Ludeke and Larsen, 2017). These findings seriously question what information these ratings actually capture.

For wide applications, rating scales are worded in everyday language, thus capitalizing on raters' and scientists' everyday knowledge. But everyday knowledge is often incoherent, contradictive and context-dependent (Laucken, 1974; Hammersley, 2013). What specific knowledge do raters actually apply? Could it be that 'outgoing' has not the same meaning for students and the general public and not the same for people from different countries? How do raters choose the scale categories to indicate their judgements? What does "agree" actually mean to different people and in what ways is this related to their intuitive judgements and scientists' axioms of quantity? Rating data have been used intensely for almost a century now (Thurstone, 1928; Likert, 1932); but still little is known about the processes by which raters actually generate these data.

Aims of This Article

This article contributes to current debates an enquiry of the epistemological and methodological foundations of rating scales, which psychologists and social scientists widely use to generate quantitative data *directly by persons* rather than using technologies (see concepts of 'persons as data generation systems', 'human-based measurement', 'measurement with persons[1]',

[1]The distinction between 'data generated *with* persons' versus 'data generated *on* persons' (frequently made in metrology) is irrelevant for the present analyses that focus on the processes by which persons generate data, no matter whether these data are about persons, non-human animals or objects.

'humans as measurement instrument'; Berglund, 2012; Pendrill, 2014). The focus is on intuitive judgements on multi-stage rating scales (e.g., Likert-style), not considering comparative judgment methods (Thurstone, 1927) or questionnaires involving right and wrong answers (e.g., intelligence tests). The article explores *processes of data generation*—before any methods of data *analysis* can be applied. These processes are crucial for measurement and quantification because data can reveal information about study phenomena *only if* relevant properties have been encoded systematically in the data. No method of analysis, however, sophisticated, can substitute these essential steps.

A transdisciplinary perspective is adopted to elaborate epistemological, metatheoretical and methodological foundations of theories and methods of data generation, measurement and quantification from psychology and social sciences but also from biology, physics and especially *metrology*, the science of measurement (BIPM, 2006). Metrology was key to the successes of the physical sciences (e.g., physics, chemistry, astronomy) but did not form the basis for measurement theories in psychology and social sciences (Michell, 1999; Mari, 2013). This notwithstanding, the article's basic premise is that *general principles of scientific measurement and quantification should apply to all sciences* (see also McGrane, 2015; Mari et al., 2017). This is no utopic ideal. It is a necessity arising from the complexity of today's real-world problems that require application of inter-, multi- and transdisciplinary approaches. Big Data gain momentum. But statistical results can be interpreted with regard to real-world phenomena *only if* the data fulfill elementary criteria of measurement and quantification that can be understood and used in the same way across sciences—without ignoring peculiarities of their objects of research.

Psychologists and social scientists encounter particular challenges because their study phenomena are intangible, highly adaptive and complex, and less rigorously rule-bound than those explored in other fields (but see Hossenfelder, 2018). Therefore, measurement technologies from physical sciences and engineering cannot be applied. Moreover, as all persons are individuals and members of social communities, scientists exploring these phenomena cannot be independent of their objects of research. This entails particular risks of (unintentionally) introducing all kinds of ego-centric and ethno-centric biases (Uher et al., 2013b; Uher, 2015c).

To elaborate principles by which basic criteria of measurement and quantification can be met in all sciences while considering fundamental differences in their objects of research, this article applies the *Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals* (*TPS-Paradigm*; Uher, 2015a,b,c,d,e, 2016a,b, 2018b,c). It is well suited for this purpose because it provides unitary frameworks in which concepts from psychology, life sciences, social sciences, physical sciences and metrology that are relevant for research on individuals have been systematically integrated. It also puts into focus the individuals who are doing research and generating data, thus opening up a meta-perspective on research processes.

First, these frameworks and relevant concepts are briefly introduced and used to explore epistemological foundations of measurement and quantification considering concepts from psychology, social sciences and metrology. Then, principles by which metrological criteria can also be met in person-generated quantifications are outlined, highlighting challenges and limitations. Application of these principles is illustrated by the example of investigations of individual-specific behaviors ('personality'). The demands that rating methods impose on data-generating persons are systematically deconstructed and compared with the demands involved in other quantitative methods (e.g., ethological observations). Closing, the article highlights problematic assumptions underlying rating methods as well as implications for their utility to improve replicability and transparency in psychology and social sciences.

TRANSDISCIPLINARY PHILOSOPHY-OF-SCIENCE PARADIGM FOR RESEARCH ON INDIVIDUALS (TPS-PARADIGM)

The TPS-Paradigm comprises a system of interrelated philosophical, metatheoretical and methodological frameworks (*paradigm*) in which concepts, approaches and methods from various disciplines (*transdisciplinary*) for exploring phenomena in or in relation to *individuals* were systematically integrated, further developed and complemented by novel ones. Its purpose is to make explicit the presuppositions, metatheories and methodologies underlying scientific systems (*philosophy-of-science*) to help researchers critically reflect on; discuss and refine their theories, models and practices; and derive ideas for novel developments (for a schematic overview, see **Figure 1**; for introductions Uher, 2015a,c, 2018c; for more information and empirical applications[2]).

Philosophical Framework

The philosophical framework specifies presuppositions made about individuals' nature and properties and the fundamental notions by which knowledge about them can be gained. Three presuppositions are important.

Complexity Theories

Complexity theories, developed amongst others in philosophy (Hartmann, 1964), thermodynamics (Prigogine and Stengers, 1984), physics of life (Capra, 1997), theoretical biology (von Bertalanffy, 1937), medicine (Rothschuh, 1963), and psychology (Wundt, 1863; Koffka, 1935; Vygotsky and Luria, 1994) allow to conceive individuals as living organisms organized at different levels forming nested systems, from molecules and cells over individuals up to societies. At each level, they function as integrated wholes in which dynamic non-linear processes occur from which new properties emerge not completely predictable from their constituents (*principle of emergence*). These new properties can feed back to the constituents from which they

[2] researchonindividuals.org

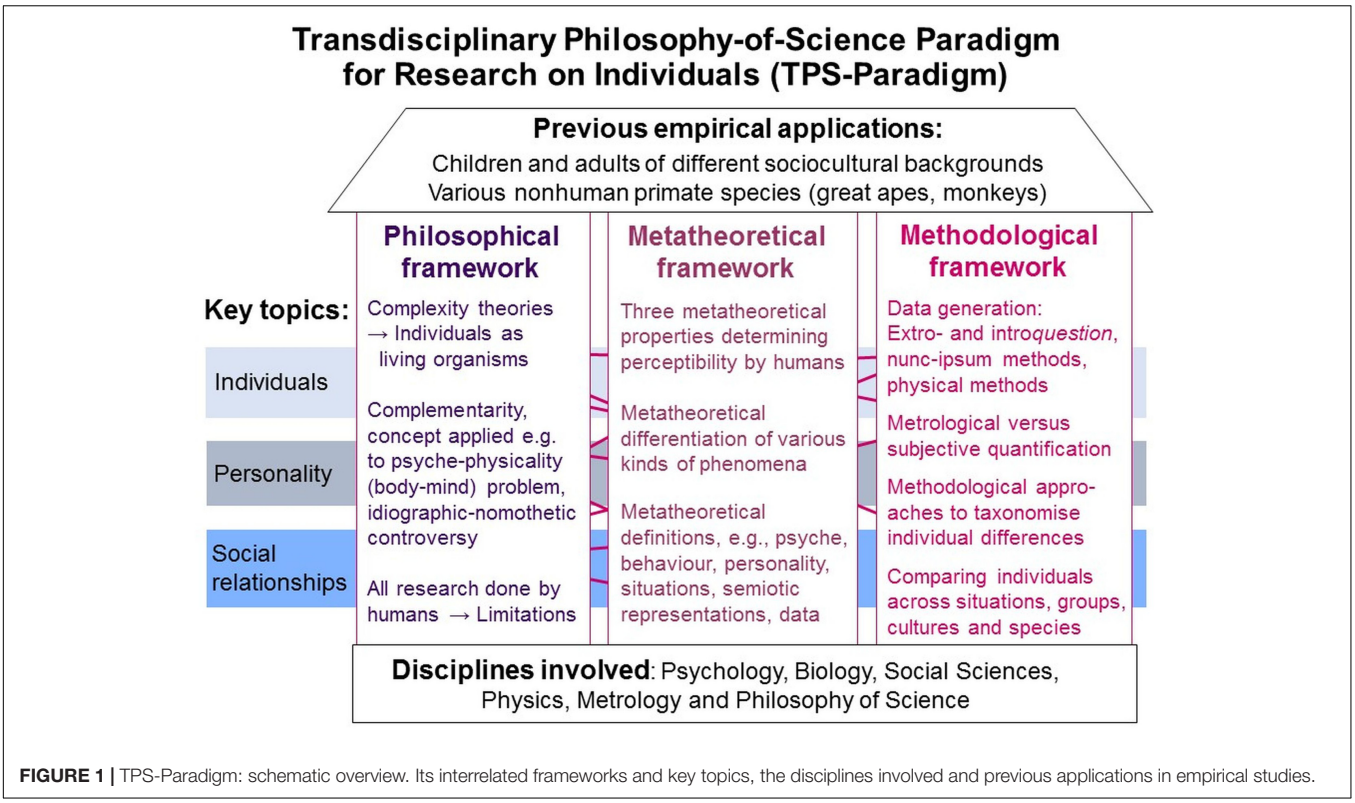

FIGURE 1 | TPS-Paradigm: schematic overview. Its interrelated frameworks and key topics, the disciplines involved and previous applications in empirical studies.

emerge, causing complex patterns of upward and downward causation. With increasing levels of organization, ever more complex systems and phenomena emerge that are less rule-bound, highly adaptive and historically unique (Morin, 2008). This applies especially to psychological and social-science objects of research.

Complementarity

This concept highlights that particular objects of research can be exhaustively understood only by describing two mutually exclusive properties that are irreducible and maximally incompatible with one another, thus requiring different frames of reference, truth criteria and investigative methods, and that may therefore be regarded as *complementary* to one another (Fahrenberg, 1979, 2013; Hoche, 2008; Walach, 2013). This concept was applied to the wave-particle dilemma in research on the nature of light (Heisenberg, 1927; Bohr, 1937) and to the body-mind problem (Brody and Oppenheim, 1969; Fahrenberg, 1979, 2013; Walach and Römer, 2011). In this problem, called *psyche-physicality problem* in the TPS-Paradigm given its particular terminology (see below; Uher, 2015c), complementarity takes a metaphysically neutral stance without making assumptions of either ontological dualism or monism while emphasizing the necessity for *methodical dualism* to account for observations of two categorically different realities that require different frames of reference, approaches and methods (Walach, 2013). In the TPS-Paradigm, complementarity is also applied to resolve the nomothetic-idiographic controversy in 'personality' research (Uher, 2015d).

Human-Made Science

The third presupposition concerns explicit recognition that *all science is created by humans*, hence on the basis of humans' perceptual (Wundt, 1907) and conceptual abilities (interpretations; Peirce, 1958, CP 2.308). This does not imply ideas of radical constructivism (von Glasersfeld, 1991), positing that concepts had no representational connection with a reality existing outside of individuals' minds and that knowledge could be developed without reference to an ontological reality in which humans have evolved over millions of years (Uher, 2015a). But it also clearly rejects naïve realist assumptions that individuals' senses could enable direct and objective perceptions of the external reality 'as it really is'. Instead, it highlights that we can gain access to this reality only through our human perceptual and cognitive abilities, which inevitably limits our possibilities to explore and understand this reality. This epistemological position comes close to those of critical realism (Bhaskar and Danermark, 2006) and pragmatism-realism (Guyon et al., 2018). They emphasize the reality of the objects of research and their knowability but also that our knowledge about this reality is created on the basis of our practical engagement with and collective appraisal of that reality. Knowledge is therefore theory-laden, socially embedded and historically contingent.

As science inherently involves an anthropocentric perspective, a *phenomenon* is defined in the TPS-Paradigm as anything that humans can perceive or (technically) make perceivable and/or that humans can conceive (Uher, 2015c). This notion differs from various philosophical definitions (e.g., Kant's, 1781/1998).

Metatheoretical Framework

Three Metatheoretical Properties Determining Perceptibility by Humans

The TPS-Paradigm's metatheoretical framework builds on *three metatheoretical properties* conceivable in different forms for phenomena studied in research on individuals. These particular properties are considered because they determine a phenomenon's perceptibility, which has important methodological implications (see below). Given the focus on research on individuals, these properties are conceived in dimensions of everyday experiences (e.g., scaled to human bodies, international time standards), ignoring micro- or macro-dimensions explored in some fields (e.g., atomic and outer-space dimensions).

These properties are (1) a phenomenon's location in relation to the studied individual's body (e.g., internal, external), (2) its temporal extension (e.g., transient, temporally extended)—both dimensional properties—and (3) its spatial extension conceived as physical versus "non-physical". Physicality here refers to concepts of classical physics, because they match everyday experiences, unlike quantum physical ones. Physical denotes corporeal/bodily/material phenomena (matter) as well as immaterial physical phenomena (e.g., heat, movements), which are not corporeal in themselves but become manifest in material phenomena with which they are systematically connected. All physical phenomena are spatially extended. But spatial properties cannot be conceived for "non-physical" phenomena, which are not simply contrasted against the physical (as indicated by the quotation marks) and therefore conceived as complementary. This distinction resembles Descartes' res extensa and res cogitans (Hirschberger, 1980) but implies only a methodical rather than an ontological dualism (Uher, 2015c, 2016a). These properties are labeled metatheoretical because they reflect a level of abstraction not commonly considered, and only time and space constitute ontological categories.

Different Kinds of Phenomena Studied in Research on Individuals

The three properties are used to metatheoretically differentiate various *kinds of phenomena*, which differ in their particular constellations in these properties' forms. For example, morphological phenomena (living organism's structures and their constituting parts) are internal/external, temporally extended and material physical. Physiological phenomena (morphology's chemical and physical functioning) are primarily internal, mostly transient and immaterial physical (**Figure 2A**). These *conceptual* differentiations, as they are accessibility-based, have important methodical implications for data generation shown below.

Basic kinds of phenomena: inseparable from the individual's bodily entity

Four kinds of phenomena are conceived as *basic* because they are inseparable from the intact individual's body: morphology, physiology, behavior and psyche (see **Figure 2A**; for details, Uher, 2015a). For the present analyses, the conceptual distinction between psyche and behavior is important.

Behaviors are defined as the "external changes or activities of living organisms that are functionally mediated by other external phenomena in the present moment" (Uher, 2016b, p. 490). Thus, behaviors are external, transient and (mostly immaterial) physical phenomena (e.g., movements, vocalizations). The *psyche* is defined as "the entirety of the phenomena of the immediate experiential reality both conscious and non-conscious of living organisms" (Uher, 2016a, p. 303; with immediacy indicating absence of phenomena mediating their perception; see Wundt, 1894). The psyche's phenomena are essential for all sciences because they are the means by which any science is made. A science exploring the psyche must therefore distinguish between its objects of research and its tools for investigating them. Therefore, the psyche's phenomena in themselves are termed *psychical*, whereas *psychological* denotes the body of knowledge (Greek -λογßα, -logia) about psychical phenomena[3].

Psychical phenomena (e.g., cognitions, emotions, and motivations) are conceived as located entirely internal and perceivable only by each individual itself and nobody else[4] (Locke, 1999). Differences in temporal extension distinguish *experiencings* (Erleben), which are transient and bound to the here-and-now (e.g., thoughts, emotions), from *memorized psychical resultants* or commonly *experiences* (Erfahrung), which are, although accessible only through experiencings, temporally more extended in themselves (e.g., sensory and psychical representations, knowledge, abilities; with memorisation here broadly referring to any retention process). Unlike immaterial physical phenomena (e.g., heat, x-radiation), the psyche's immaterial properties show neither spatial properties in themselves nor systematic relations to the spatial properties of physical phenomena to which they are bound (e.g., brain matter and physiology) and are therefore conceived as "non-physical", reflecting complementary psyche-physicality relations (see Fahrenberg, 2013).

Internality, imperceptibility by others and lack of spatial properties differentiate psyche from possible externalizations in behaviors and language from which psychical phenomena can only be inferred indirectly. This has important implications for language-based methods like ratings as shown below.

Composite kinds of phenomena comprising both phenomena inseparable from the individual's body and phenomena independent of it

In the TPS-Paradigm, three further kinds of phenomena are conceptually distinguished: *semiotic representations* (e.g., written and spoken language)—phenomena essential for rating methods—as well as *artificial outer-appearance modifications* (e.g., clothes) and *contexts* (e.g., situations) not considered here (see Uher, 2015a,c). They are conceived as *composites* because they comprise phenomena of different kind (as distinguished by

[3]This distinction is made in many languages (e.g., Dutch, French, German, Italian, and Russian) but not commonly in the English.

[4]In the TPS-Paradigm, assumptions of extended mind are rejected because psychical phenomena in themselves are differentiated from their possible expression in behaviors and language, which form inherent parts of extended mind concepts (Clark and Chalmers, 1998; Logan, 2007).

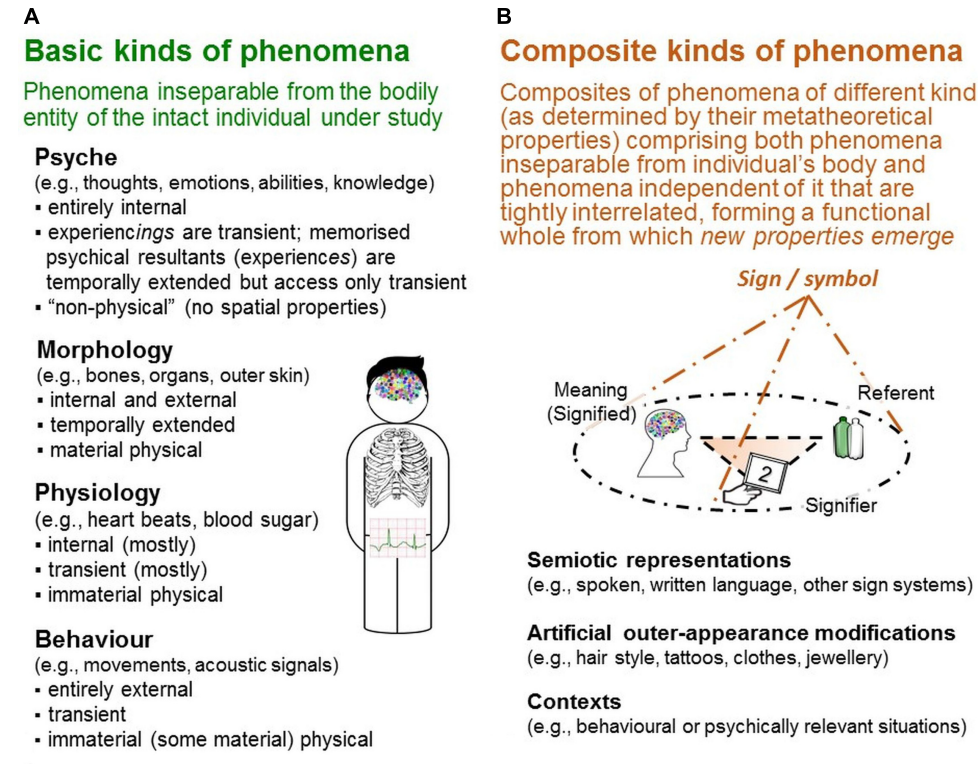

FIGURE 2 | Kinds of phenomena. In the TPS-Paradigm, various kinds of phenomena are conceptually differentiated by the particular constellation of forms regarding the three metatheoretical properties determinating their perceptibility. Two types are distinguished: **(A)** basic kinds of phenomena are characterized by their inseparability from individuals' bodies, and **(B)** composite kinds of phenomena by their complexity and heterogeneity of the phenomena involved, some of which are independent of individuals' bodies.

the three metatheoretical properties) that are tightly interrelated with one another, forming a functional whole from which new properties emerge (**Figure 2B**). These new properties can be explored only by studying the composite's constituents in their functional interdependence. Importantly, these composites are conceptual and not demarcated by physical boundaries (unlike, e.g., biological cells). Instead, their constituents are located apart from one another, which considerably complicates their exploration as semiotic representations illustrate.

Semiotic representations (e.g., written language) are composites in which (a) particular *psychical constituents* (the *signified*; e.g., meanings, mental representations) are tightly interrelated with (b) particular *physical constituents external* to individuals' bodies (the *signifier*; e.g., ink on paper, vocalizations) and (c) particular *referents* to which both refer and which may be located external or internal to individuals' bodies. These three constituents form a functional composite from which new properties emerge—those of *signs* (sign includes the notion of *symbol* in the TPS-Paradigm). For example, a semiotic representation may comprise (a) ideas and meanings of bottles, (b) visual (graphic) patterns shaped like "BOTTLE" or "Flasche" (German for bottle), or acoustic (phonetic) patterns like ['bɒt.əl] or ['flaʃə] and (c) some bottles to which both refer (**Figure 3**). Visual and acoustic patterns are external physical and can thus be perceived by others and used to decode the meanings and referents someone may have encoded in them. Importantly, meanings are not inherent to the physical signifiers in themselves but only *assigned* to them. Meaning construction occurs in people's minds; it is internal and psychical ("non-physical"). The term semiotic *representation* highlights that individuals' psychical representations are the essential component that interconnects a sign's signifier with its referent. These three constituents are all located apart and not demarcated as an entity and therefore cannot be straightforwardly recognized as a composite. Socially shared assignments turn such composites into *signs*—but only for persons making such attributions (the signs' psychical constituent). Such assignments are arbitrary and therefore vary (e.g., different alphabets; Vygotsky, 1962; Westen, 1996; Uher, 2015a). For these reasons, semiotic representations are complex and metatheoretically heterogeneous, involving external and internal, physical and "non-physical", temporally extended and transient phenomena. This considerably complicates explorations, such as their function as data.

Excurse: data – semiotic representations used to encode information about the study phenomena

Data are signs (symbols) that scientists use to represent information about the study phenomena in physically

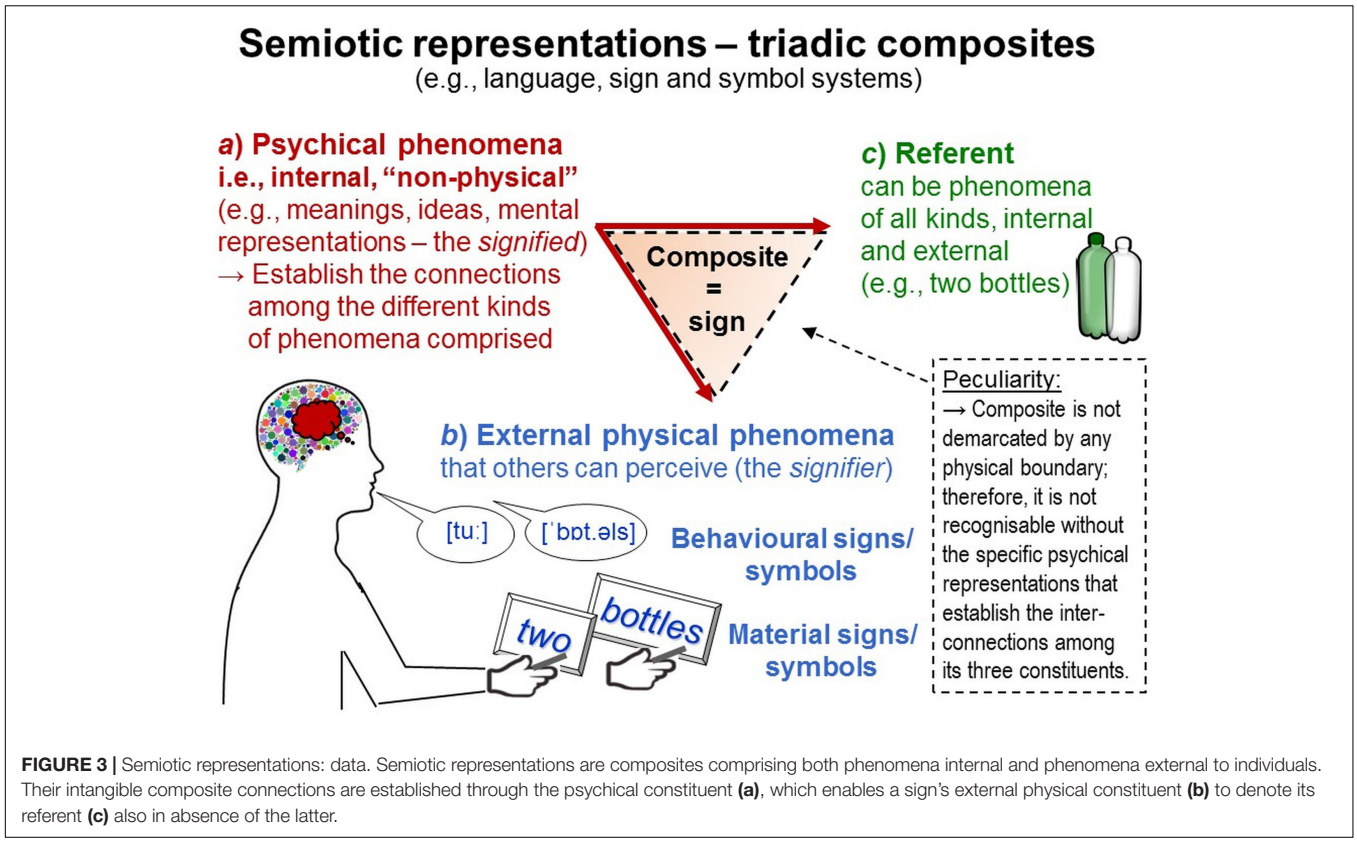

FIGURE 3 | Semiotic representations: data. Semiotic representations are composites comprising both phenomena internal and phenomena external to individuals. Their intangible composite connections are established through the psychical constituent **(a)**, which enables a sign's external physical constituent **(b)** to denote its referent **(c)** also in absence of the latter.

persistent and easily perceivable ways. Thus, data are semiotic representations—composites comprising particular *physical constituents* (e.g., visible patterns like "two" or "2") to which particular persons (e.g., scientists, observers) assign particular *meanings* (e.g., mathematical properties) and refer both to particular *referents*—the properties and phenomena under study (e.g., numbers of bottles; **Figure 3**).

Important types of data are *numerals* (apart from textual data). Numerals comprise physical constituents (e.g., visible patterns shaped like 1, 5, 10, and 50) to which individuals often—but not always—attribute the meaning of *numbers*. As such attributions are arbitrary, the meaning of numbers can also be attributed to other physical constituents (e.g., visible patterns shaped like I, V, X, L). Vice versa, different meanings can be assigned to the same signifiers that then constitute different signs (e.g., Roman numerals also represent alphabet characters). Consequently, *not all numerals represent numbers*. Whether or not numerals represent numbers depends on the meanings attributed by their creators—an important point for data generation.

Data, as they are signs (symbols), can be stored, manipulated, decomposed and recomposed, that is, *analyzed in lieu of the actual phenomena under study* (the referents) and in ways not applicable to these latter. But inferences about the study phenomena can be made *only if* the data represent relevant properties of these phenomena in appropriate ways. This is a further important point for data generation taken up again below.

Methodological Framework

Data Generation Methods Are Determined by the Study Phenomena's Modes of Perceptibility: Basic Principles and Method Classes

The three properties, because they describe modes of perceptibility under everyday conditions, also specify the ways to make phenomena accessible under research conditions. Therefore, these metatheoretical properties are used in the TPS-Paradigm to derive methodological principles and define basic method classes that cut across common classifications, which specify properties of data once these are generated (e.g., 'qualitative', 'quantitative'; Uher, 2018a).

External phenomena (e.g., behaviors) are publicly accessible and can be studied without any mechanism standing between observer and observed using *observational methods*. *Internal physical* phenomena (e.g., brain), by contrast, are imperceptible under everyday conditions but can be made perceptible under research conditions using *invasive or technical methods* (e.g., surgery, X-ray).

Temporally extended phenomena (e.g., body morphology) do not change quickly, which facilitates perception and enables repeated perception of the same entity. *Transient phenomena* (e.g., behaviors, nerve potentials), by contrast, can be perceived and recorded only in the brief moments when they occur, thus real-time using so-called *nunc-ipsum*[5] *methods* (e.g., observations, EEG).

[5]Derived from the Latin *nunc ipsum* for at this very instant.

Physical phenomena, both material and immaterial (e.g., morphology, heat), are spatially extended. Therefore, they can be captured with *physical methods*, which rely on the spatial extensions of materials that are systematically related to and more easily perceivable than the study phenomena (Uher, 2018a), such as mercury in glass tubes for measuring temperature (see Chang, 2004). *Psychical* phenomena, given their non-spatial ("non-physical") properties, are inaccessible by any physical method and cannot be made perceivable by others. This unique property is used in the TPS-Paradigm to distinguish methods enabling access to psychical phenomena from those that cannot.

Introquestive[6] *methods* are all procedures for studying phenomena that can be perceived *only from within the individual itself and by nobody else in principle under all possible conditions*. This applies to psychical phenomena, which can be explored by others *only indirectly* through individuals' externalizations (e.g., behaviors, language). Accordingly, all methods of self-report and inner self-observation are introquestive. *Extroquestive*[7] *methods*, by contrast, are all procedures for studying phenomena that *are or can (technically) be made perceptible by multiple individuals* (**Figure 4**). This applies to all physical phenomena (including internal and immaterial ones, e.g., inner organs, heat) because they can be made perceptible using invasive or technical methods (e.g., surgery, EEG). Joint perception of the same entity by multiple individuals (e.g., observers) is essential for data quality assurance (e.g., establishing intersubjectivity; see below; Uher, 2016a, 2018a).

Previous concepts of intro*spection* versus extro*spection* are distinguished from one another with regard to the studied individual by denoting its "inward perspective" versus "outward perspective", respectively (Schwitzgebel, 2016). These two perspectives are, however, not perceived as separate channels of information. Instead, they are always merged in the multifaceted unity emerging from the composite of all perceptions available at any moment (Wundt, 1894). Therefore, introspection and extrospection cannot be differentiated as methods. By contrast, extro*question* and intro*question* are defined and differentiated on the basis of (a) the particular study phenomena (e.g., sounds, thoughts), considering that other internal and external phenomena can be simultaneously perceived, and of (b) the particular persons who perceive the study phenomena and generate from their perceptions data about these phenomena (Uher, 2016a).

These concepts highlight that psychophysical investigations of relations between sensory perceptions and physical stimuli (Fechner, 1860; Titchener, 1905)—commonly interpreted as introspective—are actually extroquestive methods. The physical stimuli (e.g., lights, sounds) are external to participants' bodies and therefore perceivable also by the experimenters. Only because physical stimuli are extroquestively accessible can they be experimentally varied and compared with individuals' subjective judgements. Thus, contrary to widespread assumptions, psychophysical findings about sensory perceptions cannot be generalized to perceptions of phenomena that are accessible only introquestively. Involvement of perceptions does not qualify investigations as introquestive because perceptions are always involved in any investigation (e.g., natural-science observation; Uher, 2016a, 2018a).

This perceptibility-based classification of data generation methods highlights that a phenomenon's modes of accessibility determine *unequivocally* the class of methods required for its investigation. Each kind of phenomenon can be captured only with particular method classes and no method class allows for exploring all kinds of phenomena (see complementarity; Uher, 2018a). These are further important points for data generation taken up again below.

[6]Derived from the Latin *intro* for in, within; and *quaerere* for to seek, enquire.
[7]Derived from the Latin *extro* for beyond, outside.

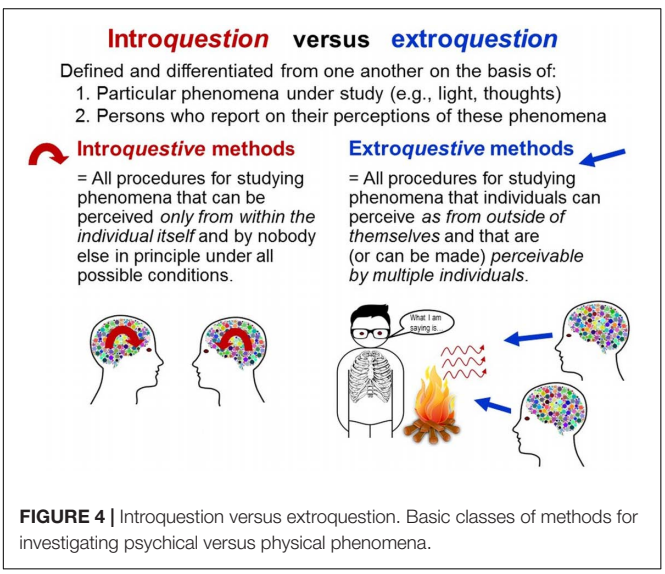

FIGURE 4 | Introquestion versus extroquestion. Basic classes of methods for investigating psychical versus physical phenomena.

MEASUREMENT AND QUANTIFICATION ACROSS THE SCIENCES

The TPS-Paradigm's frameworks and the concepts outlined above are now applied to explore concepts of measurement and quantification, highlighting commonalities and differences among sciences.

Measurement Versus Quantification

In psychology, quantification and measurement are often considered synonyms; but they are not the same. *Quantification* generally denotes the assignment of numbers, whereas *measurement* denotes a purposeful multi-step process, comprising operative structures for making such assignments in reliable and valid ways together with explanations of how this is achieved (Maul et al., 2018). Hence, not every quantification is an outcome of measurement (Abran et al., 2012).

Concepts of Quantity and Early Measurement Theories

What Is a Quantity?

A *quantity* is a divisible property of entities of the same kind—thus, of the same quality. Two types are distinguished, multitudes and magnitudes (Hartmann, 1964).

Multitudes are discontinuous and discrete quantities that are divisible into indivisibles and discontinuous parts, which are countable—numerable—and therefore expressible as a number (e.g., persons, eyeblinks). Thus, multitudes are quantities by their ontological nature (Hartmann, 1964). *Magnitudes*, by contrast, are continuous and unified quantities that are divisible into divisibles and continuous parts. Magnitudes can be directly compared and rank-ordered in terms of 'more', 'less', or 'equal' (e.g., body length). By comparing a target property's magnitude with the magnitudes of designated references of the same kind of property (e.g., length of units on rulers), which constitute multitudes and are thus countable, their ratios can be expressed as a measurement unit (e.g., meter) and a number (JCGM200:2012, 2012).

Early Measurement Theories and the Fundamental Problem of Psychological and Social-Science Measurement

From an epistemological analysis of counting and measuring (von Helmholtz, 1887), Hölder (1901) axiomatized equality/inequality, ordering and additivity relations among physical magnitudes, thereby laying the foundations for their measurement (Michell, 1997; Finkelstein, 2003). Quantities for which additive operations can be empirically constructed and quantities that can be derived from them led to further measurement concepts. In *fundamental (direct) measurement*, quantities are obtained directly (e.g., length). In *derived measurement,* the target quantity is obtained *indirectly* from relations between other directly measurable quantities (e.g., volume from length; Campbell, 1920). In *associative measurement*, the target quantity is obtained *indirectly* through measurement of another quantity with which it is systematically connected (e.g., temperature through length of mercury in glass tubes; Ellis, 1966; Chang, 2004).

Psychophysicists, pioneers of early psychology, studied equality and ordering relationships of sensory perceptions of physical stimuli (e.g., just-noticeable-differences and comparative judgements of light stimuli; Titchener, 1905), which is possible only because they constitute extroquestive explorations. But the properties of psychical phenomena in themselves, especially non-sensory ones (e.g., thoughts, emotions, and motivations), cannot be empirically added (concatenated) or derived from additive quantities. The possibility of their measurement was therefore rejected by the British Association's for the Advancement of Science committee for quantitative methods (Ferguson et al., 1940; see also Kant, 1786/2016; Trendler, 2018). This led psychologists and social scientists to focus on relational models, operational theories and utility concepts (Michell, 1999; Finkelstein, 2003).

Representational Theory of Measurement

Representational theory of measurement, developed in the social sciences, formalizes (non-contradictory) axiomatic conditions by which empirical relational structures can be mapped onto symbolic relational structures, especially numerical ones (Krantz et al., 1971; Suppes, 2002). For measurement, these many-to-one mappings (homo- or isomorphisms) must be performed such that the study phenomena's properties and their interrelations are appropriately represented by the properties and interrelations of the signs used as data (*representation theorem*). Permissible transformations specify how the numerical representations can be further transformed without breaking the mapping between the empirical relations under study and the numerical ones generated (*uniqueness theorem*; **Figure 5**; Vessonen, 2017).

In physical sciences and engineering, representational theory plays no role, however, despite its applicability (Finkelstein, 2003). This may be because it formalizes initial stages of measurement and important conditions of measurability but does not stipulate any particular measurement procedures (Mari et al., 2017). Another problem concerns establishing measurability (i.e., evidence of ordered additive structures of the same quality) because not just any mapping of numbers onto empirical relational structures constitutes measurement. But the appropriateness of particular numerical representations

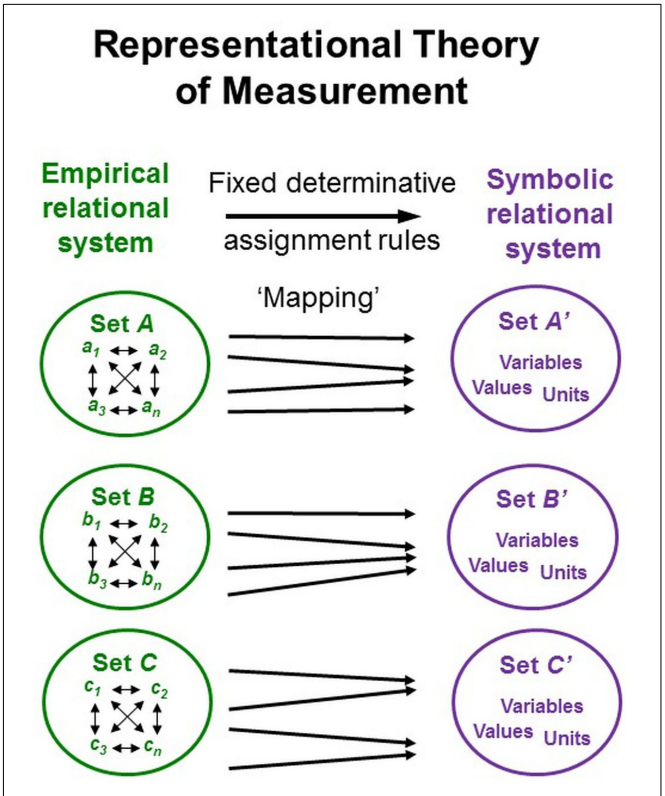

FIGURE 5 | Representational theory of measurement. Key elements of representational systems frequently used in psychological and social-science concepts of measurement.

is often only assumed rather than established, thereby reducing the interpretability of the generated symbolic representation *regarding the empirical phenomena under study* (Blanton and Jaccard, 2006; Vessonen, 2017).

Psychometric Theories of Measurement

Psychometric theories are concerned with statistical modeling approaches, building on various *positivist* epistemologies that focus on empirical evidence and predictive ability (*instrumentalist* focus) rather than on finding true explanations of reality. Therefore, some psychometricians apply *operationalist* epistemologies and determine study phenomena by the methods used for their exploration (Bridgman, 1927), such as by defining intelligence as "what an IQ-test measures" (Boring, 1923; van der Maas et al., 2014). This, however, reduces measurement to any number-yielding operation (Dingle, 1950). It also ignores that measurement results constitute information that can be understood also outside the specific context in which they were generated (Mari et al., 2017). The ability to represent information also in absence of their referents is a key feature of semiotic representations like data (Uher, 2015a, 2016b).

Psychometricians applying classical test theory or probabilistic latent trait theory (e.g., item response theory, Rasch modeling) sometimes build on *naïve realist* epistemologies by assuming ratios of invariant quantities exist in the world and independently of the methods used (Mari et al., 2017). Hence, they assume that ideal methods (e.g., purposefully designed rating scales) allow to empirically implement an identity function, turning pre-existing 'real' scores into estimated (manifest) scores—although with errors or only certain probabilities, which, however, can be defined with reference to the assumed 'true' scores or 'latent trait' scores, respectively. But this ignores that interactions between study properties and methods always influence the results obtained (Heisenberg, 1927; Bohr, 1937; Mari et al., 2017). In human-generated measurement, these interactions are intricate because they are mediated by the data-generating persons who perceive and interpret—thus interact with—both the study properties (whether located internally or externally) and the methods used (e.g., rating scales, observation schemes). Metrologists' concepts of 'humans as measuring instruments' (Pendrill, 2014) and psychometrician's concepts of rating scales as 'measuring instruments' do not reflect these triadic relationships.

Metrological Concepts of Measurement and Scientific Quantification

To justify that quantifications can be attributed to the objects of research, metrologists define measurement as a purposive process comprising operative structures that establish evidence for its object-dependence ("objectivity") and the subject-independence of its results ("intersubjectivity"; **Figure 6** Frigerio et al., 2010; Mari et al., 2012, 2017). Importantly, in metrology, "*object*ivity" refers to the *object* of research and denotes that measurement processes depend on the objects and properties under study (therefore *object-dependence*)—compliant with complementarity. Results are "*intersubjective*" if they are "invariant with respect to the substitution of the involved subjects" (Mari et al., 2017)—thus, the persons generating and using them (therefore *subject-independence*). In psychology, by contrast, "objectivity" commonly denotes intersubjectivity in terms of independence from the investigator. It refers to the results not the process, thus confounding two metrological criteria of measurement.

An important way of establishing object-dependence and subject-independence is to implement *traceability*. Traceability requires measurement results to be systematically connected

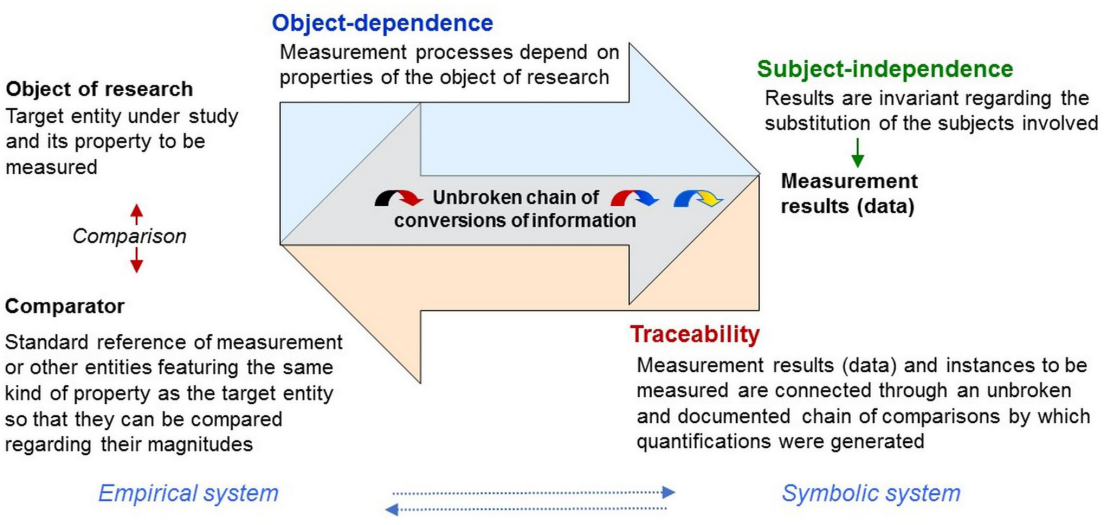

FIGURE 6 | Traceability. Metrological concepts stipulate basic and testable elements of measurement procedures linking measurement results (data) with the phenomena and properties under study through traceable conversions of information. They provide key concepts by which symbolic (e.g., numerical) relational systems can be mapped onto empirical relational systems, which were left undefined in representational measurement theory.

through an unbroken and documented chain of comparisons to a reference (comparator; **Figure 6**), which can be a measurement standard or the definition of a measurement unit through its practical realization (JCGM200:2012, 2012). This allows measurement results to be traced back to the particular instances of the properties measured (objects of research) and the particular comparisons and standards by which quantifications were obtained (empirical examples below).

These concepts stipulate basic and testable elements of measurement procedures by which 'numbers can be mapped onto empirical relational structures', thus allowing to establish evidence of measurability and intersubjectivity of the results obtained—key elements, left undefined in representational measurement theory (**Figure 6**). In the TPS-Paradigm, numerical data that fulfill these metrological criteria are called *scientific quantifications* as opposed to (subjective) quantifications in which these are not fulfilled.

PERSON-GENERATED MEASUREMENT AND SCIENTIFIC QUANTIFICATION: BASIC PRINCIPLES FOR FULFILLING METROLOGICAL CRITERIA IN PSYCHOLOGY AND SOCIAL SCIENCES

This section elaborates principles by which metrological concepts of measurement, although developed for physical phenomena, can also be met in investigations of "non-physical" phenomena, highlighting challenges and limitations.

Establishing Object-Dependent Measurement Processes and Subject-Independent Results: Some Challenges

To connect objects of research (empirical relational structures) and measurement results (symbolic relational structures) through unbroken documented chains of comparisons, suitable operational processes must be established including explanations of how unbroken chaining is achieved (for issues of measurement uncertainty, not discussed here, see Giordani and Mari, 2012, 2014; Mari et al., 2017).

Constructs: Defining Theoretical Ideas

Psychological and social-science objects of research can be conceived very differently (e.g., behaviors, attitudes). Therefore, researchers must *theoretically define* the phenomena and properties of interest. Theoretical definitions describe the objects of research—in representative measurement theoretical terms, the empirical entities under study and their relational structures. Theoretical concepts are abstract and generalized ideas, which necessarily differ from their perceivable referents (Daston and Galison, 2007; Uher, 2015a). Abstract concepts (e.g., 'personality', 'extraversion', 'social status') describe complex constellations of phenomena that *cannot be directly perceived* at any moment but that are only theoretically *construct*ed as entities (therefore called *constructs*; Cronbach and Meehl, 1955). Hence, their theoretical definition is a matter of decision, which can but need not be intersubjectively agreed (see, e.g., different definitions and theories of 'personality').

Measurement Variables: Encoding Perceivable Qualities

As abstract ideas, constructs cannot be measured in themselves. Therefore, constructs are often called 'latent' in terms of 'underlying' and not directly perceivable, which often misleads people to reify constructs as real entities internal to individuals (e.g., 'traits' as psychophysical mechanisms; Uher, 2013).

To enable quantification, constructs must be *operationally defined*, thus, be related systematically to *specific indicators* that are directly measurable and used to quantify a construct indirectly. Erroneous analogies are sometimes drawn to indirect physical measurement, where the target quantity is derived from measurement of other directly measurable quantities (see above). But indirect measurement builds on natural connections among different kinds of quantities, which are experimentally identifiable whereas construct operationalization is a matter of decision, which may, but need not, be intersubjectively agreed (see, e.g., the different models and operationalizations of 'personality').

The complexity of constructs requires multiple indicators; but no set of indicators, however, large, can be all-inclusive (e.g., comprehensively operationalize 'personality'). Constructs imply more meaning (*surplus meaning*) than the indicators by which they are operationalized. To ensure sufficient coverage, researchers specify a construct's meanings in a theoretical framework of more specific sub-constructs and establish links to an empirical framework comprising sets of indicators. For example, popular models of the abstract construct of 'personality' comprise various more specific constructs (e.g., 'extraversion', 'neuroticism', 'agreeableness', and 'conscientiousness'), each of which, in turn, comprises various sub-constructs (e.g., 'gregariousness', 'assertiveness') operationalized with various variables (e.g., rating items).

Psychotechnical engineering, where variables are purposefully chosen to operationalize theoretically defined constructs (following representational measurement theory), is aimed at generating aggregate scores for defined sets of variables (Cronbach and Meehl, 1955; Messick, 1995; Vautier et al., 2012). This differs from *psychometric engineering*, where construct definitions are derived from empirical interrelations among variables (following operationist assumptions; Thissen, 2001; Vautier et al., 2012). The Big Five personality constructs, for example, were derived from ratings on person-descriptors taken from the lexicon and are defined by these ratings' empirical interrelations as studied with factor analysis (therefore, commonly called 'personality' factors; Uher, 2015d).

While these issues are well-known and intensely discussed, psychometricians hardly ever specify how the data-generating persons can actually identify the empirical relational system and execute the assignments to the symbolic relational system. This likely results from the deficiencies of representational measurement theory and psychometric theories but also from the "non-physical" objects of research and language-based data

generation methods. Specifically, constructs are abstract ideas that 'exist' as entities only in people's mind and language. When concepts constituted by words are explored with methods constituted by words, it is difficult to distinguish the methods from the measures of the object of research (Lahlou, 1998; Uher, 2015d). Rating scales serve both as descriptors of the empirical relational system and as elements of the symbolic relational systems, thus *confounding two key elements of representational measurement theory*. Psychometricians provide raters neither with clear definitions of each relational system nor with specifications of the assignments to be made between them, leaving their interpretation and execution to raters' intuitive judgments and decisions (**Figure 7**).

As data generation requires interaction with the objects of research, persons must be able to *directly perceive* them. Consequently, data generation methods must be used that match the study phenomena's modes of perceptibility (see above). Researchers must define the study phenomena in terms of their *perceivable qualitative properties* and must specify the variables in which they are (commonly lexically) encoded (**Figure 8**).

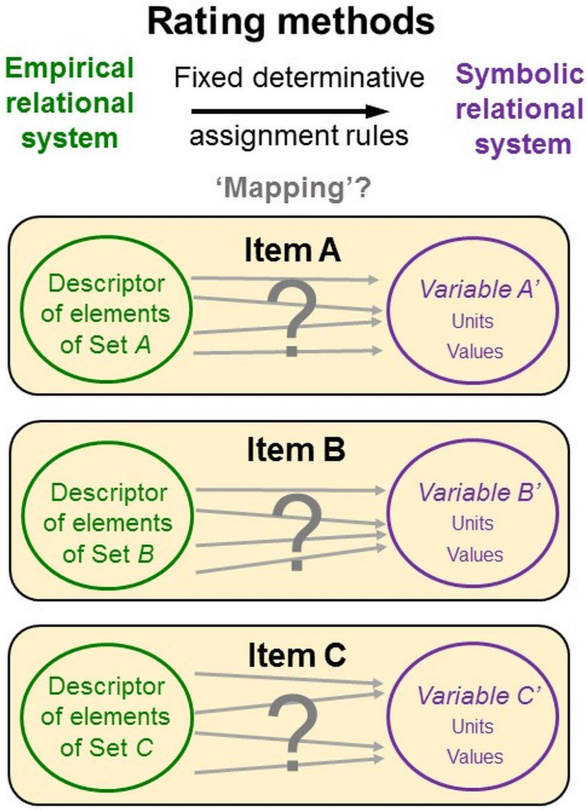

FIGURE 7 | Rating items confound empirical and relational system. Rating scales serve both as descriptors of the empirical relational system and as elements of the symbolic relational system, thus confounding two key elements of representational measurement theory. Psychometricians provide neither definitions of each relational system nor specifications for assignments between them, leaving their interpretation and execution to raters' intuitive judgments and decisions.

Measurement Units: Encoding Perceivable Quantities

For each measurement variable, researchers must then define *measurement units*. As they belong to the *same* variable, units refer to properties conceived as identical or at least sufficiently similar—thus, of the *same quality*.

Different types of units are used. Nominal units encode either more specific qualities or, as binary units, absence versus presence of the quality of interest, whereas rational, interval and ordinal units encode divisible properties of the quality studied, thus quantitative properties. For each unit type, permissible transformations are specified that maintain the mapping to the empirical relational system under study (Stevens, 1946). Hence, the empirical relational system's properties determine which unit type can be used. For person-generated quantification, researchers must define divisible properties of the study phenomena that are or can be made *directly perceivable* during data generation (**Figure 8**).

Encoding Procedure: Defining Fixed and Unchanging Assignment Rules

For measurement, the *same* properties must *always* be encoded with the *same* signs so that the data obtained always represent the *same* information and can be understood and used by others in the *same* way, thus subject-independently. This presupposes explicit assignment rules (e.g., many-to-one mappings), variables, units and values that are *fixed and unchanging*. Metatheoretically speaking, the symbolic systems (e.g., observational encoding scheme) must be intersubjectively understood with regard to the referents and meanings they are meant to semiotically encode.

Decisions to Be Made by the Person Generating the Data During Measurement Execution

To execute a measurement task, persons must have certain abilities and make various decisions (**Figure 9**), which form inherent parts of *data generation methods* (Uher, 2018a). Extroquestive accessibility of study phenomena enables multiple persons to jointly perceive the same entity. This facilitates establishing intersubjective consensus in making these decisions (i.e., subject-independence).

Demarcating the Entities of Interest Using Perceivable Properties

First, in the multitude of a study phenomenon's perceivable properties, data-generating persons must be able to demarcate the entities of interest in reliable and systematic ways. They must decide which pieces of information should be demarcated in what ways using perceivable similarities and dissimilarities. Variations in perceivable properties (e.g., in spatio-temporal extensions in behaviors) complicate these decisions (e.g., which demarcable entities are sufficiently similar to count as being of the same kind; **Figure 9**).

Categorizing Demarcated Entities Using Theoretical Considerations

Then, data-generating persons must decide how to categorize the demarcated entities using perceivable properties but also

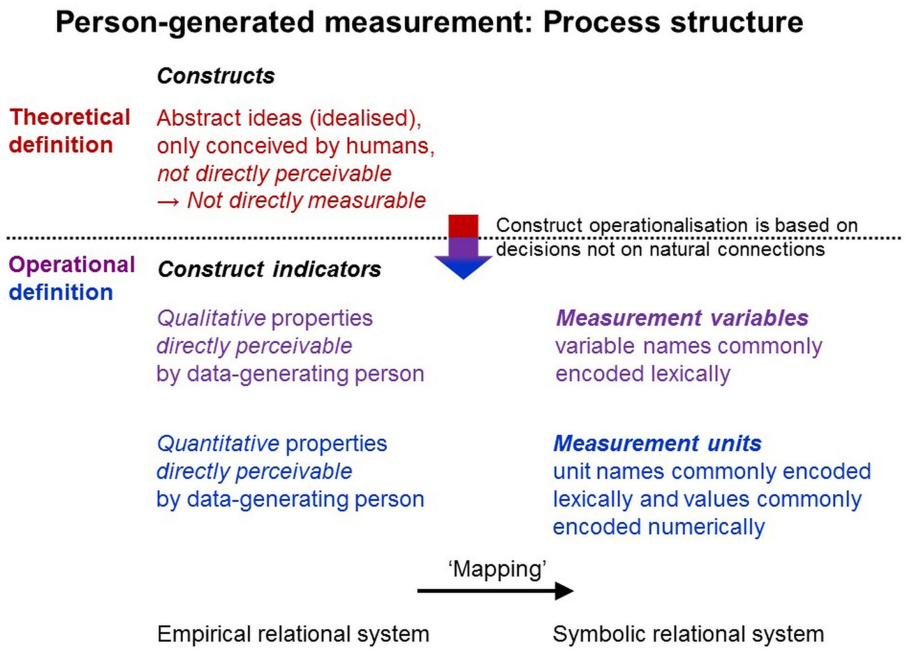

FIGURE 8 | Theoretical definition and empirical operationalization. Process structure for measurement directly by persons in research on individuals.

similarities and differences in their *known or assumed functions and meanings*—thus, theoretical and contextual considerations. For example, the behavioral acts of slapping someone to kill a mosquito and to bully that individual feature almost identical perceivable properties but differ in function and meaning; whereas smiling, talking and shaking hands have similar social functions but differ in their perceivable spatio-temporal forms (**Figure 9**). This shows why data generation is always theory-laden; as Einstein already said "it is the theory which decides what can be observed" (Heisenberg, 1989, p. 10). When researchers provide no system for categorizing the entities under study, as in many so-called 'data-driven' approaches, then the data-generating persons must use their own implicit theories to accomplish this task.

Converting Information About Categorized Entities Into Semiotically Encoded Information

Thereafter, data-generating persons must represent perceived occurrences of the thus-categorized entities into the signs used as data. When information from one kind of phenomenon is represented in another one, this is called *conversion* in the TPS-Paradigm (Uher, 2018a). For systematic and standardized conversions of information from the empirical into the symbolic relational system, scientists must specify which pieces of information from the study phenomena should be demarcated, categorized and semiotically encoded in what ways. That is, scientists must define the three constituents of the signs used as data (see **Figure 3**) such that the data-generating persons can execute the measurement operations.

Scientific Quantifications Generated Directly by Persons: General Preconditions and Challenges

Psychometricians are rather unconcerned with all these decisions raters have to make during data generation. Instead, psychometricians apply sophisticated methods of data modeling (e.g., Rasch modeling) to demonstrate that the data—once raters have produced them—exhibit quantitative structures. But data analysis cannot *add* fundamental properties that have not been encoded in the raw data. To what extent are persons actually able to *directly generate* scientific quantifications (i.e., quantitative data that are object-dependent and subject-independent) *during data generation*?

For interval and ratio-scaled direct quantifications, spatial standard units of measurement are widely used (e.g., yard sticks for length measurement). Distinct entities (i.e., multitudes; e.g., rope jumps) can be directly counted. If not applicable, persons can compare several entities with one another—provided these can be perceived in close spatial *and* temporal proximity together—to determine their relative magnitude regarding the quality of interest (e.g., body height, intensity), thus enabling ordinal-scaled quantifications[8] (e.g., highest, second highest, third highest; **Figure 10**).

But persons' abilities to count or directly compare the entities of interest with one another or with spatial standards of measurement are often compromised in momentary and highly fluctuating phenomena featuring variable properties. For example, the dynamics of behaviors often hinder applications of spatial standards of measurement (e.g., to quantify movements). Direct comparisons between behavioral acts are complicated

[8]In metrology, ordinal scaled data do not constitute quantifications (BIPM, 2006).

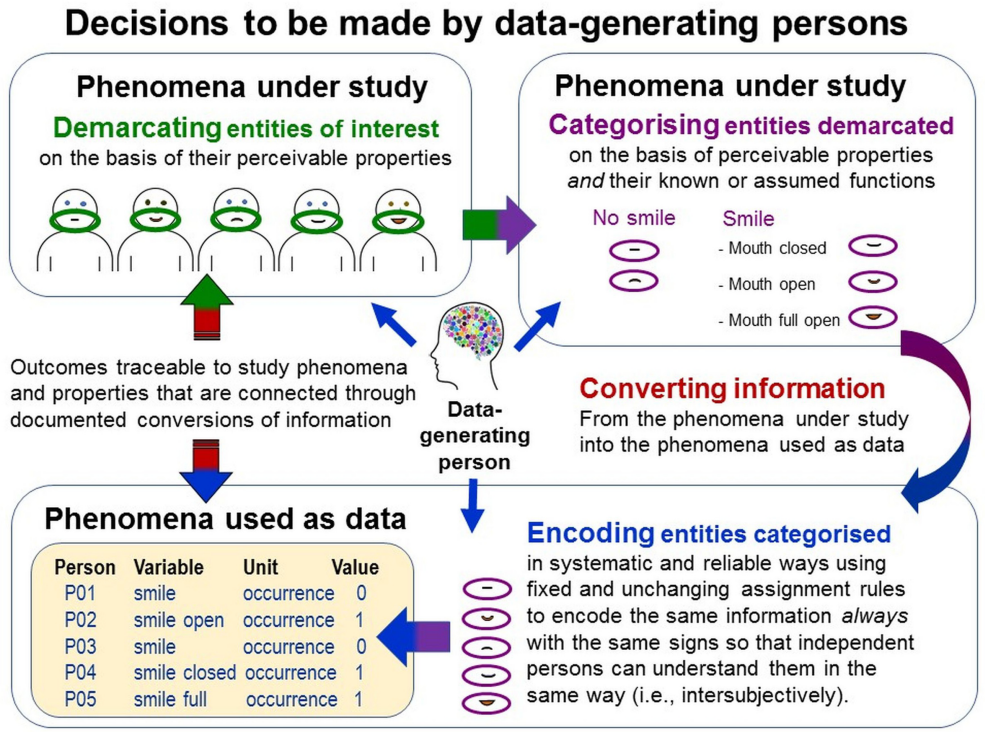

FIGURE 9 | Decisions to be made by data-generating persons. Data generation requires complex decisions about demarcating, categorizing and encoding the entities of interest. They are particularly challenging when study phenomena feature variable perceivable properties (e.g., variable spatio-temporal extensions of behaviors). For example, smiles vary in spatial extensions; humans can turn their mouth corners up in various ways, with mouth open or closed, with or without laughter lines around their eyes, and all this at various levels of intensity. Given this, what entity can be demarcated and categorized as one (n = 1) smile? Smiles also vary in temporal extension. Are quick and long-lasting smiles events of the same kind? When does one smile end and another one start? Thus, which particular demarcable entities can be considered to be sufficiently similar to categorize them as being of the same kind? Making these decisions explicit is important for establishing traceability of the generated data.

both within individuals because previous acts have already ceased to be and between individuals because individuals seldom behave spatio-temporally in parallel with one another (as arranged in races). To solve this problem, behavioral scientists (e.g., biologists) apply observational methods enabling time-based measurement, whereas psychologists and social scientists primarily use rating methods. These two methods are now explored in detail and compared with one another.

QUANTITATIVE DATA GENERATION WITH RATING METHODS VERSUS OBSERVATIONAL METHODS: POSSIBILITIES AND LIMITATIONS FOR FULFILLING METROLOGICAL CRITERIA

The TPS-Paradigm's frameworks and the metrological criteria of scientific quantification are now applied to deconstruct the demands that different methods of quantification impose on data-generating persons, contrasting rating methods with behavioral observations (starting with the latter). These elaborations are illustrated by the example of individual-specific behaviors as study phenomena (behavioral parts of 'personality').

To be specific to individuals, behavioral patterns must vary among individuals and these differences must be stable over some time (Uher, 2013, 2018b). But neither differential nor temporal patterns can be directly perceived at any moment. As behaviors are transient, fluctuating and dynamic, individual behavior patterns cannot be straightforwardly measured either (Uher, 2011). This considerably complicates quantifications of individual-specific behaviors.

Demands Placed on Observers

Targeted Perception and Real-Time Demarcation, Categorization and Encoding

Observation, unlike looking or watching, involves targeted and *systematic perception* of the phenomena and properties under study. As behaviors are transient, observers must target their perceptions to relevant properties and must demarcate and categorize behavioral events *in the brief* moments while they occur, thus real-time using nunc-ipsum methods (see above). To achieve this in standardized ways *while observing* the continuous flow of dynamically changing events, observers must know by heart all elements of the empirical relational system under study (e.g., all definitions of behavioral acts specified in the ethogramme), all assignment rules and all variables, units and

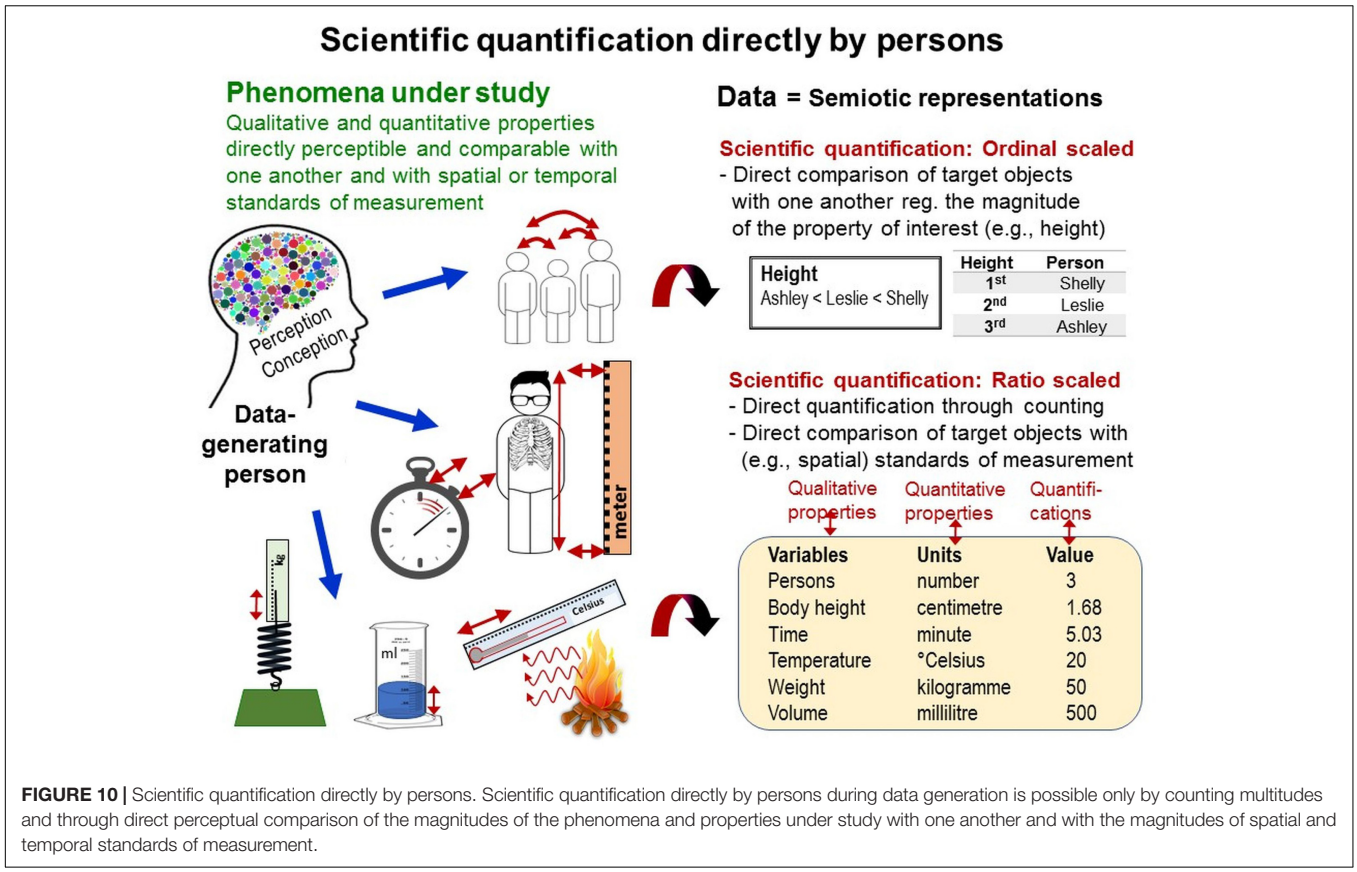

FIGURE 10 | Scientific quantification directly by persons. Scientific quantification directly by persons during data generation is possible only by counting multitudes and through direct perceptual comparison of the magnitudes of the phenomena and properties under study with one another and with the magnitudes of spatial and temporal standards of measurement.

values that constitute the symbolic relational system, thus the data.

To meet these demands, observers are instructed and trained. Training is possible because behaviors are extroquestively accessible, which facilitates intersubjective perception and discussion about the decisions required to generate data. Observers' performances are studied as agreement between codings generated by independent persons observing the *same* behaviors in the *same* individuals and situations at the *same* occasions. This subject-independence is statistically analyzed as inter-observer (inter-coder) reliability. Behaviors' extroquestive accessibility also facilitates the design of object-dependent observation processes involving unbroken documented chains of comparisons (see **Figure 6**). Video-based coding software allows observers to mark the video sequences in which particular behaviors occur so that their demarcation, categorisation and encoding can be traced to the specific behaviors observed (as recorded on video) and to the ways they were quantified.

Defining the Empirical Relational System: Specifying All Studied Elements of the Sets B, S, I, and T

To enable observers to perceive, demarcate, categorize and encode behaviors in systematic and standardized ways, researchers must specify all phenomena to be quantified (i.e., all elements of the empirical relational system) in terms of their perceivable qualitative and quantitative properties. For investigations of individual-specific behaviors, this involves the set *B* of all behaviors studied and the set *S* of all situations in which they are observed (considering the context-dependent meanings of behaviors; Uher, 2016b). Researchers must also specify the set *I* of individuals studied as well as the set *T* of occasions and periods of time in which their behaviors are recorded (**Figure 11A**).

The sets of individuals (e.g., sample characteristics) and times studied (e.g., observation time per individual) are specified in every method section. Situations can be defined on more abstract levels as nominal situations (e.g., location, test condition) or on more fine-grained levels such as regarding specific interpersonal situations (e.g., being approached by others). This requires observers to demarcate and categorize situational properties *in addition to* the behavioral properties studied, thus further increasing the demands placed on them. For such fine-grained analyses, researchers often use video-based techniques enabling deceleration of the flow of events and repeated observations of the same instances.

The perceivable qualities of behaviors can often be interpreted differently regarding their possible functions and meanings. To enable categorisation, researchers must specify the theoretical interrelations of the behaviors studied as well as their possible contexts and observers must know these by heart. Observational designs must be developed that are practically feasible given the used settings (e.g., restricted or unrestricted), sampling methods (e.g., behavior or time sampling) and recording techniques (e.g., manual or computerized recording). Defining

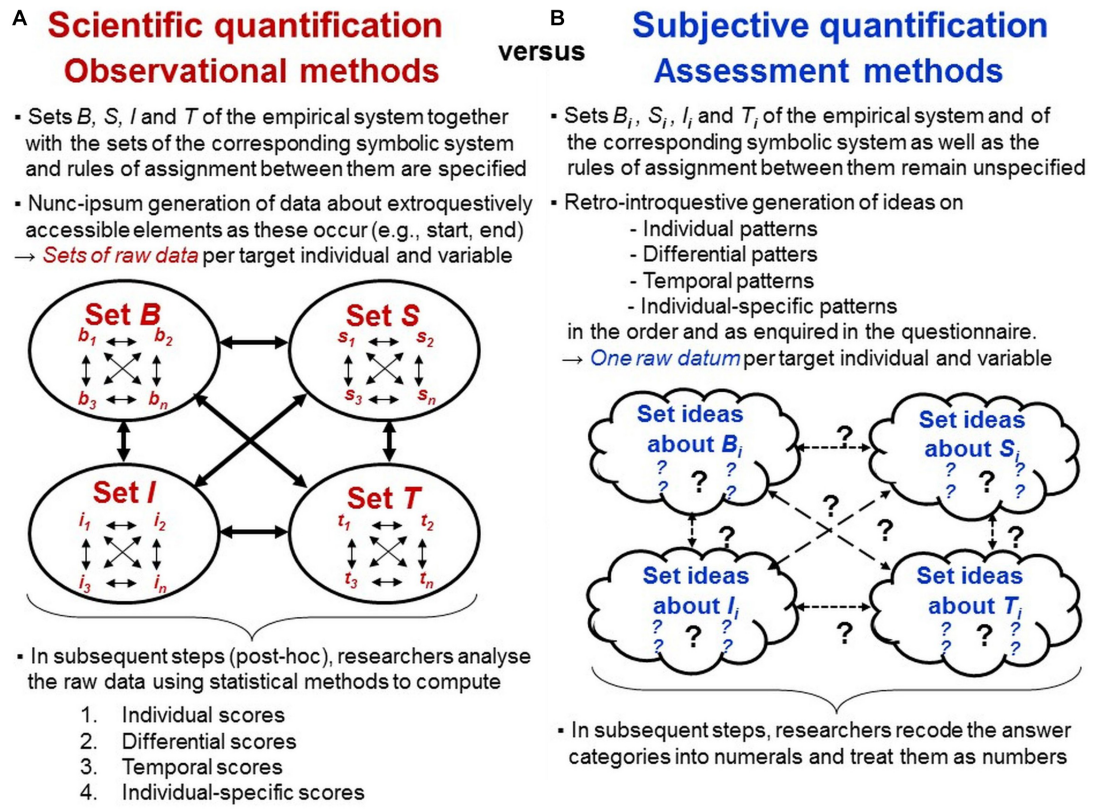

FIGURE 11 | Scientific versus subjective quantification. Processes involved in the generation of quantitative data using **(A)** observational methods versus **(B)** assessment methods by the example of investigations of individual-specific behaviors (habitual behaviors forming part of an individual's 'personality').

the studied elements' theoretical interrelations within and between the empirical sets *B, S, I,* and *T* studied is prerequisite for specifying the corresponding symbolic relational system (therefore indicated with primes) involving the sets *B′*, *S′*, *I′*, and *T′* as well as the rules for assignments between both (**Figures 4**, **11A**).

Measuring and Quantifying Behavioral Events Real-Time – Possibilities and Limitations

Transience and pronounced spatio-temporal variations of behaviors require that observers decide flexibly about how to demarcate events, thus often precluding comparisons with *spatial standards of measurement*. Therefore, behavioral events—with all perceivable spatial variations as defined by the researchers—are often encoded only in their occurrence or non-occurrence using binary units (see **Figure 9**). Scientific quantifications are then generated through comparison with *temporal standards of measurement*. Such quantifications, as they are based on behaviors' temporal properties, may differ from quantifications that could be obtained from their spatial properties. Time-based measurement puts high demands on observers because they must monitor time in addition to the behavioral and situational properties studied. It also requires clear specification of the perceivable divisible properties used for quantification (see **Figure 9**). Software- and video-based observation technologies facilitate the nunc-ipsum recording of occurrences, onsets and ends of binarily encoded behavioral events, producing time-based log-files (Uher, 2013, 2015b; Uher et al., 2013a).

Summarizing, behavioral observations place high demands on the data-generating persons. They show that persons' abilities to directly generate quantifications that meet axioms of quantity and measurement are very limited. This often confines observational data to nominal formats; but these data are clearly defined and traceable and thus suited to generate scientific quantifications *post hoc* (see next). By recording the events of interest in nominal units while or immediately after they occur (nunc-ipsum), observers have already completed their task. They are required neither to memorize events observed, nor to directly quantify them nor to mentally compute their empirical occurrences or interrelations within and across the empirical sets *B, S, I,* and *T*. Such computations are made by researchers in subsequent steps of data *analysis* using the elements of the symbolic relational system generated.

After Observations Are Completed: *Post hoc* Generation of Ratio-Scaled Data From Nominal-Scaled Raw Data

Nominal data indicate classification, which is essential for measurement but not yet quantification. Because nominal units disjunctively encode occurrence or non-occurrence of qualitative

properties, nominal-scaled raw data can be used to generate ratio-scaled quantifications, which meet the axioms of quantity and quantification, *post hoc—after raw data generation has been completed*. Generating ratio-scaled quantifications of individual-specific behaviors ('personality' scores) requires three steps (**Figure 11A**; Uher, 2011, 2013).

First, to generate data reflecting *individual patterns*, each individual's raw data are aggregated over specified time periods, thus, they are *temporally standardized*. Formally speaking, in the symbolic relational system, the nominal-scaled data (multitudes) of each studied element b'_n for each studied element i'_n within each studied element s'_n are aggregated (counted) and then related to all studied elements t'_n of the set T'. Because non-occurrence of events defines an absolute zero point, the data thus-generated are *ratio-scaled*. Most behavioral coding software executes this step automatically (e.g., computing durations and frequencies).

Second, to generate data reflecting patterns of *individual differences*, individuals' data must be *differentially standardized* within the sample and each situation studied (e.g., using *z*-standardization). Formally speaking, the data generated in step 1 for each element i'_n within each element b'_n and each element s'_n are statistically standardized across the entire set I' of individuals (**Figure 11A**). *Differential standardization* transforms data reflecting absolute quantifications into data reflecting *relative* between-individual differences. As these transformations are made explicitly and *post hoc*, individuals' absolute scores can always be traced for interpretation and possible re-analyses as well as for comparisons with other sets of individuals, thus fulfilling the criterion of object-dependence of measurement outcomes. Differential standardization enables direct comparison of individuals' relative scores among behavioral variables of different kind (e.g., frequencies, durations) both within and between individuals. It also enables statistical aggregation into more abstract and composite variables on the basis of algorithms (e.g., different weighting of behaviors) that are specified in the theoretical definitions of the empirical relational system. Importantly, these comparisons and aggregations are always made with regard to the differential patterns reflected in the data, not with regard to individuals' absolute scores (computed in step 1) because these may generally vary across behaviors and situations (Uher, 2011).

Third, differential patterns can reflect *individual-specificity* only if they are stable across time periods longer than those in which they were first ascertained and in ways considered to be meaningful (e.g., defined by test-retest correlation strength; Uher, 2018b). Hence, identifying individual-specificity in behavior requires *temporal analyses* of differential patterns that are defined by certain temporal patterns in themselves (Uher et al., 2013a). Therefore, the symbolic set T' of occasions and spans of time is divided into subsets (e.g., t'_1 and t'_2), and steps 1 and 2 are performed separately on these subsets to enable between-subset comparisons for test–retest reliability analysis. Sufficient test–retest reliability provided, the differential patterns obtained in step 2 are then aggregated across the subsets t'_n to obtain data reflecting *ratio-scaled quantifications* of *individual-specificity* in behaviors (**Figure 11A**).

Importantly, this *post hoc* data processing is done by researchers and constitutes first steps of data analysis, which is independent of observers' data generation task. These analytical steps are described here to highlight the complexity of the comparisons required to scientifically quantify individual-specific behaviors. This puts into perspective the demands placed on raters.

Demands Placed on Raters

Quantifying Individual-Specific Behaviors Directly – An Impossible Requirement

To quantify individual-specific behaviors with rating methods, relevant behaviors are described in sets of statements, called *items*, that constitute a rating 'instrument' (e.g., questionnaire, inventory, and survey). Persons, the raters, are asked to judge the behaviors described (e.g., "tends to be lazy") regarding, for example, their occurrences, intensity or typicality for a target individual. Raters are asked to indicate their judgements on rating scales comprising a fixed set of *answer categories* often labeled lexically (e.g., "agree" or "neither agree nor disagree" and "disagree"). Hence, raters are asked to *directly quantify* individual-specific behaviors.

But in everyday life and without recording technologies, persons often cannot *directly* quantify even single behavioral events (see section "Demands Placed on Observers"). Quantifying individual specificity in behaviors (or other kinds of phenomena) requires quantifying not only single events and individual patterns in many behaviors but also differences among individuals and over time. But in transient, dynamic and fluctuating phenomena, differential and temporal patterns cannot be directly perceived and thus cannot be quantified at any moment. Individual specificity is not an entity one could directly perceive but an abstract idea constructed by humans. For this construction, human language is essential.

Language – Essential for Abstract Thinking but Also Misleading

Language allows persons to semiotically represent perceivable phenomena (e.g., concrete behavioral acts) in single words (e.g., "shout", "kick"). This allows perceivable qualities to be made independent of their immediate perception and to abstract them into objects, thus *reifying* them ("aggression"). This so-called *hypostatic abstraction* (Peirce, 1958, CP 4.227) enables people to develop not only concrete words that refer to directly perceivable phenomena but also abstract words that refer to ideas and concepts describing phenomena that are distant from immediate perception (Vygotsky, 1962) or complex and imperceptible in themselves—such as constructs of 'personality' (e.g., "aggressive*ness*"). Signs (e.g., rating items) therefore cannot reflect the referents they denote (**Figure 3**) in the same ways as individuals can perceive them.

People (including scientists) often tend to mistake linguistic abstractions for concrete realities. This so-called *fallacy of misplaced concreteness* (Whitehead, 1929) misleads people to assume that the complex phenomena described with abstract terms (e.g., in rating items) could be directly perceived. It also occurs when people encode their ideas about individual

specificity in abstract terms (e.g., 'personality', 'traits', 'character', or 'dispositions'), and then treat these abstractions as real entities that they assume to underlie individuals' feeling, thinking and behaving and thus to be located internally. This entails explanatory circularity (Uher, 2013, 2018b).

Further challenges occur because semiotic representations contain implicit structures in both their external physical constituents (e.g., phonetics) and the particular meanings of the referents assigned to them (e.g., semantics). These implicit structures and meanings are not readily apparent because no physical border demarcates a sign's three constituents as an entity (see above). This entails intricacies for language-based methods like ratings.

Standardized Rating Items Do Not Reflect Standardized Meanings—Instead, Their Meanings Vary Within and Between Individuals

Like in observations, variables and units of rating scales (as elements of the symbolic relational system) are predetermined and fixed. But unlike in observations, raters are commonly neither instructed nor trained to interpret and use them in standardized ways—thus in how to understand the given symbolic relational system and to relate it to the empirical relational system under study (in fact, both systems are confounded in rating methods; see **Figure 7**).

Rating scales are worded in everyday language in abstract and generalized ways to make them applicable to diverse phenomena, events and contexts without specifying any particular ones. Therefore, raters must use their common-sense knowledge to interpret and contextualize the given scale and to construct specific meanings for the rating task at hand. Common-sense categories are, however, not as well-elaborated and disjunctive as scientific categories but often fuzzy and context-sensitive, enabling flexible demarcations (Hammersley, 2013). To reduce cognitive effort, raters may interpret items on the abstract level on which they are worded—the semantic level (Shweder and D'Andrade, 1980; Block, 2010). Semantic processing may be triggered especially by highly inferential items requiring judgments of the social valence, appropriateness, and normativity of individual behaviors (e.g., "respectful", "socially adapted") or of their underlying aims and motivations (e.g., "helping").

As meanings are not inherent but only as*sign*ed to the physical constituents of signs (e.g., phonemes, graphemes), meanings vary. For popular personality questionnaires, substantial within- and between-individual variations in item interpretations have meanwhile been demonstrated, highlighting that—contrary to common assumptions—standardized items represent not standardized meanings but broad and heterogeneous *fields of meaning* (Valsiner et al., 2005; Rosenbaum and Valsiner, 2011; Arro, 2013; Lundmann and Villadsen, 2016; Uher and Visalberghi, 2016).

In personality psychology, broadly worded rating items are known to be related to broader ranges of more heterogeneous and less specific behaviors and situations, thus representing more diverse aspects of given constructs (Borkenau and Müller, 1991). So far, such fidelity-bandwidth trade-offs were not regarded problematic because more abstract items have higher predictive validity for broader ranges of behaviors (e.g., job performance; Ones and Viswesvaran, 1996). Following instrumentalist epistemologies, broad rating items were even considered necessary to match the breadth of the criteria to be predicted (Hogan and Roberts, 1996).

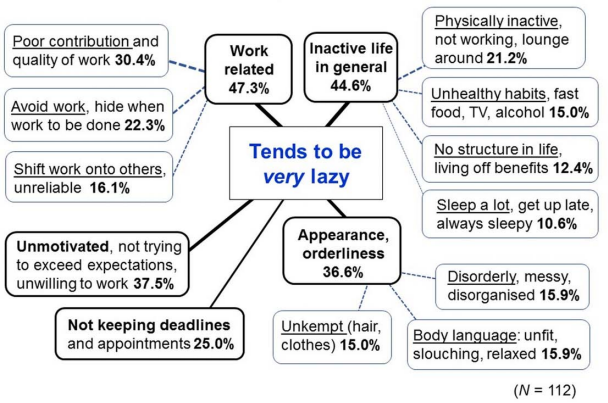

FIGURE 12 | Diverse item interpretations spanning a field of meaning. Interpretations provided by *N* = 112 raters for a fictitious target person scoring high on the item. Main-themes (bold frames) summarize sub-themes of similar interpretations (non-bold frames). Numbers indicate percentages of raters who provided given interpretations; for the main-themes, each rater was counted just once even when they provided multiple similar interpretations as specified in the corresponding sub-themes.

From a measurement perspective, however, fidelity-bandwidth trade-offs are highly problematic because they entail lack of traceability of what has actually been encoded in the data. An example illustrates this. Interpretations of "tends to be lazy", operationalizing the construct Conscientiousness in a popular personality inventory (BFI-10; Rammstedt and John, 2007), varied considerably within and among 112 raters. Raters variously associated this item with different behaviors and situations related to work, an inactive life style, appearance and orderliness, not keeping deadlines and lack of motivation (**Figure 12**; Uher and Dharyial, unpublished). This diversity may reflect the well-known fidelity-bandwidth trade-offs of broadly worded items. But importantly, every rater provided on average only two different interpretations (M = 2.08; SD = 0.92; range = 1–5). Thus, the single raters did not consider the item's broad field of meaning that it may generally have in their socio-linguistic community. Instead, when judging the target person, different raters thought of very different behaviors and contexts; some considered "sleeping a lot", others "shifting work on others", still others "eating fast food", or "not keeping deadlines" (**Figure 12**). This may explain the substantial variations in internal consistencies of personality scales across countries (see above).

Moreover, raters' item interpretations can also go beyond the bandwidth of meanings that researchers may consider. A study involving five-method comparisons showed that, despite expert-based item generation, raters' item interpretations

clearly referred also to constructs other than those intended to be operationalized; raters' and researchers' interpretations overlapped to only 54.1–70.4% (Uher and Visalberghi, 2016).

Variations in item interpretation are an unavoidable consequence of the abstract and generalized wording of items and the necessity for raters to apply them to specific cases, and therefore occur despite careful iterative item selection (Uher and Visalberghi, 2016). They show that, for different raters, the same item variables do not represent the same meanings (symbolic relational system); consequently, raters do not have the same empirical relational system in mind. This precludes subject-independence and also limits possibilities to establish object-dependence. These issues are ethically problematic because 'personality' ratings are used not only for making predictions (following instrumental epistemologies) but also to identify properties that are attributable to the target individuals (following naïve-realist epistemologies).

Unknown Demarcation, Categorization and Encoding Decisions: The Referents Raters Consider for Their Ratings Remain Unspecified

A key feature of rating methods is the *introquestive data generation in retrospect*. This allows persons to generate data any time, in any situation (e.g., online), and even in complete absence of the phenomena (e.g., behaviors) and individuals under study. This contributes to the enormous efficiency of ratings (Uher, 2015e)—but has numerous methodical implications.

Persons can encode in data relevant information about the study phenomena *only if* they can directly perceive the phenomena and properties under study during data generation. Direct perceptibility is prerequisite for establishing object-related measurement processes (see above). But quantitative ratings inherently involve also comparisons among individuals or over time or both. To rate (one's own or others') individual-specific behaviors, raters must consider past behaviors and situations, thus phenomena no longer extroquestively accessible. Raters can form such judgments only by retrieving pertinent information from memory; therefore, ratings *are long-term memory-based introquestive methods*[9] (Uher, 2018a).

Human abilities to reconstruct memorized events are generally constrained, susceptible to various fallacies and influenced by situational contexts (Shweder and D'Andrade, 1980; Schacter, 1999; Schacter and Addis, 2007). Therefore, assessments are influenced by raters' motivations and goals (Biesanz and Human, 2010). Moreover, in individuals' psychical systems, past events are stored not in the forms as once perceived but only in abstracted, integrated and often lexically encoded form (Le Poidevin, 2011; Valsiner, 2012). Individuals can base their ratings only on the outcomes of their past processing of past perceptions and conceptions; thus, on the beliefs, narratives and knowledge they have developed about individuals in general and the target individual in particular. Ratings cannot encode (habitual) behaviors in themselves, as sometimes assumed, but only the psychical and semiotic representations raters have developed about them—which are phenomena very different from behaviors (see above; Uher, 2013, 2015d, 2016a,b).

When raters' responses are restricted to ticking boxes on standardized scales, not only remain differences in item interpretations unknown but also raters' decisions on how to demarcate, categorize and encode the phenomena and properties that *they* consider as the referents of their ratings (elements of the empirical relational system). Formally stated, the elements of the set B_i of ideas about behaviors, the set S_i of ideas about behavioral situations, the sets I_i of ideas about individuals, the set T_i of ideas about occasions and spans of time as well as the ideas about these elements' empirical occurrences and interrelations that raters implicitly consider cannot be specified (**Figure 11B**). Consequently, raters' decisions during data generation and their degree of standardization and reliability cannot be analyzed. With inter-rater reliability, psychometricians analyze only agreement in the data sets produced (symbolic relational system) but not agreement in the ways in which raters demarcate and categorize information from the study phenomena (empirical relational system) and convert and encode them in the data (representational mapping between both systems). Insufficient or even lacking specification of the objects of research (**Figure 7**) hinders the design of object-dependent measurement processes and compromises the interpretability of data and findings.

In rating methods, every item variable is commonly used only once to generate one single datum per target individual and rater. In extroquestive nunc-ipsum methods like observations, by contrast, measurement variables can be used without limitation to encode defined elements of the empirical relational system as often as these may empirically occur—observational methods are matched to the study phenomena (object-dependence). The same variable can be used to generate entire data sets per target individual and observer (**Figures 11A,B**). In rating methods, by contrast, variables are presented in a fixed order predetermined by the researcher and that is mostly random with regard to the empirical relational systems they are meant to operationalize. Consequently, occurrences of events in raters' minds are triggered by and adapted to the given rating scale—*the study phenomena are matched to the methods rather than vice versa* (Westen, 1996; Toomela and Valsiner, 2010; Omi, 2012; Uher, 2015d,e).

Unknown and Changing Units and Assignment Rules

Scale categories are commonly labeled with global terms (e.g., "strongly", "often"), icons (e.g. ☺, ☹), numerals or segmented lines. Given the well-known fidelity-bandwidth trade-offs of broadly worded rating items, researchers commonly assume that, for any given rating, raters consider a broader range of evidence to form an overall judgment that they then indicate in a single score on the scale. But how do raters actually choose the answer box on the scale? How do they interpret and use quantitative scale categories at all?

To constitute measurement units, scale categories must refer to the same quality as defined by the item variable.

[9]Not all introquestive methods are based on long-term memory recall. Further methods involve *nunc-ipsum introquestion* (e.g., thinking aloud methods) and *retro-introquestion* (e.g., diary methods; for details and criteria, see Uher, 2018a).

The different scale categories assigned to each variable must represent divisible properties, thus different quantities of that quality. These categories' interrelations must adequately reflect the interrelations among the quantities of the empirical relational system that they encode. Thus, a lower score on the scale must indicate a lower quantity of the quality under study than a higher score on that same scale. Statistical aggregation across different items—a common practice in psychometrics—presupposes that the scale units have the same meaning for *all* the item variables for which they are used; similarly, metric units of weight (e.g., kg, lb) have the same meaning for all kinds of objects, whether stones, persons or feathers. Statistical aggregation across different raters—another common practice in psychometrics—presupposes that the different raters interpret and use the same scale in the same way. Similarly, different weighing machines, no matter how constructed, should be standardized (i.e., calibrated) and provide the same results for the same object. Hence, measurement outcomes should be subject-independent.

But similar to item interpretations, raters' interpretation and use of scale categories vary substantially within and between raters (Rosenbaum and Valsiner, 2011). When asked to judge a film protagonist's personality on a five-stage agreement scale and to explain their choice of the answer category, 78 raters provided very different reasons (**Figure 13** depicts reasons provided for the BFI-10 item "is outgoing, sociable" operationalizing the construct Extraversion; Uher, unpublished). Only 10.7% of all explanations indicated that raters considered and weighted various pieces of evidence as commonly assumed (highlighted in red). About 15% indicated that raters based their ratings on the occurrence of one single instance, such as a key indicator or their first impression (highlighted in blue), ignoring all other pieces of evidence available. Most explanations (67.7%) showed that raters found there was not enough evidence to make a judgment, that they missed key indicators, attributed the behaviors observed to the situation or found them not genuine, and thus not indicative of the target's personality, among further reasons. But all raters had ticked a box.

The diversity of reasons for ticking rating scales and the triviality of many decisions that raters reported to have made shows that raters' interpretations of scale categories vary considerably. Quite many interpretations do not refer to quantitative considerations at all. Moreover, raters assigned the same reasons to different categories, which have thus not distinct but overlapping meanings (**Figure 13**). This shows that raters do not interpret and use these scales in standardized ways, and that they do not apply fixed determinative assignment rules to indicate their judgments in the scale units. Neither object-dependence nor subject-independence can be established. But all this remains unknown because raters are commonly not asked to explain how they have generated their ratings.

Raters' reasons for selecting answer categories of agreement scales

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Saw multiple key indicators (3)	Did not see enough evidence (15)	Did not see enough evidence (15)	Did not see enough evidence (30)	Saw key indicator for high sociability (8)
	Found behaviour was due to situation not the target person (3)	Found behaviour was due to situation not the target person (15)	Missed key indicator (15)	First impression was crucial (2)
	Saw key indicator for low sociability (2)	Averaged what I saw (6)	Saw key indicator of high sociability (8)	Saw multiple key indicators (1)
	Found behaviour not genuine (2)	Missed key indicator (2)	Found behaviour not genuine (7)	
	Keep extreme columns for really extreme cases (2)	Unsure if what I saw is related to item (2)	Averaged what I saw (5)	
	Find word "strongly" problematic (1)	Relevant behaviours were moderate (1)	Found behaviour was due to situation not the target person (3)	
		Found behaviour not genuine (1)	First impression crucial (3)	
		Imagined how I would get along with target person (1)	Imagined how I would get along with target person(2)	
		Felt in a rush (1)	Use "strongly" only for extreme cases (2)	(*N* = 78)
1	2	3	4	5

and their common recoding into numerals by researchers

FIGURE 13 | Raters' interpretations of scale categories. Reasons provided by *N* = 78 raters for their ratings of a target person seen in a film on the BFI-10 item "... is outgoing, sociable". Numbers in parentheses indicate absolute frequencies of reasons provided; multiple nominations possible.

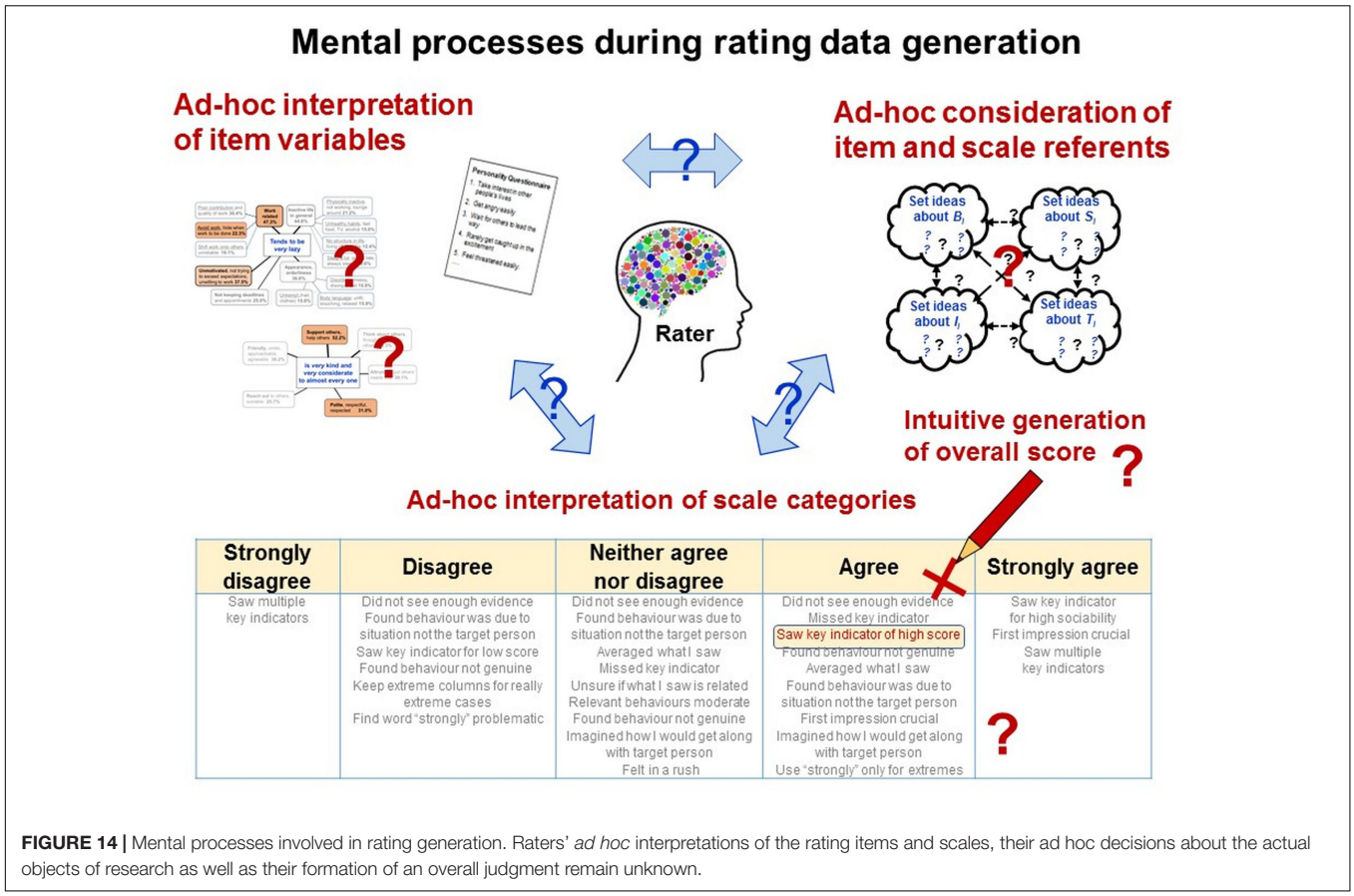

FIGURE 14 | Mental processes involved in rating generation. Raters' *ad hoc* interpretations of the rating items and scales, their ad hoc decisions about the actual objects of research as well as their formation of an overall judgment remain unknown.

Lack of Traceability of Intuitive Ratings Cannot Be Overcome by Converting Rating Scale Categories *post hoc* Into Numerals

Once raters have completed raw data generation, researchers commonly recode the (often lexically encoded) answer categories into numerals and treat these as numbers. This means that "have not seen enough evidence" can be recoded into the same numerical score as "missed a key indicator" or "found the behavior not genuine" (e.g., "4"). Likewise, "found behavior was due to situation not the target person" and "unsure if what I saw is related to the item" can be recoded into a higher Extraversion score for the target person (e.g., "3") than "have seen a key indicator for low sociability" (e.g., "2"). Hence, the interrelations of the numbers into which researchers recode scale categories (e.g., order of magnitude) do not match the interrelations of the answer categories as raters have interpreted and used them (**Figures 13**, **14**). Instead of constituting quantities of the quality specified in the item, raters' interpretations of scale units rather constituted further *qualities* in themselves, such as ideas of the considered behaviors' authenticity, relevance and situation-dependence.

Summarizing, the requirement to generate data introquestively and long-term memory-based, the lack of information about the representational system under study and the constraint response format prevent that researchers come to know about how raters actually understand and use rating scales. This precludes intersubjective discussion about the interpretation of the data generated and thus establishing subject-independence. Instead, researchers commonly interpret rating data with regard to the meanings that *they themselves* assign to the item variables, supported by countless validation studies. But for any single rating, raters obviously do not consider the broad fields of meaning an item may generally have in their sociolinguistic community. Instead, for the specific case at hand, they construe specific meanings, which constitute only a fraction of the item's overall field of meaning (**Figures 14**, **15**).

Moreover, researchers interpret the units and values of rating data with regard to the meanings of numbers that *they themselves* assign to the scale categories. This recoding of units constitutes a conversion of information that, in itself, is based on well-documented and unbroken chains of comparisons from raters' ticks on the scales, thus creating perfect traceability. But the quantifications thus-obtained cannot be traced to the referents (empirical relational system) that raters have aimed to encode in their ratings (symbolic relational system), thus also precluding the establishment of object-dependent data generation processes. Researchers' rigid recoding of answer categories breaks the chain of traceability that could be established if raters' judgment and encoding processes were systematically explored (**Figures 14**, **15**).

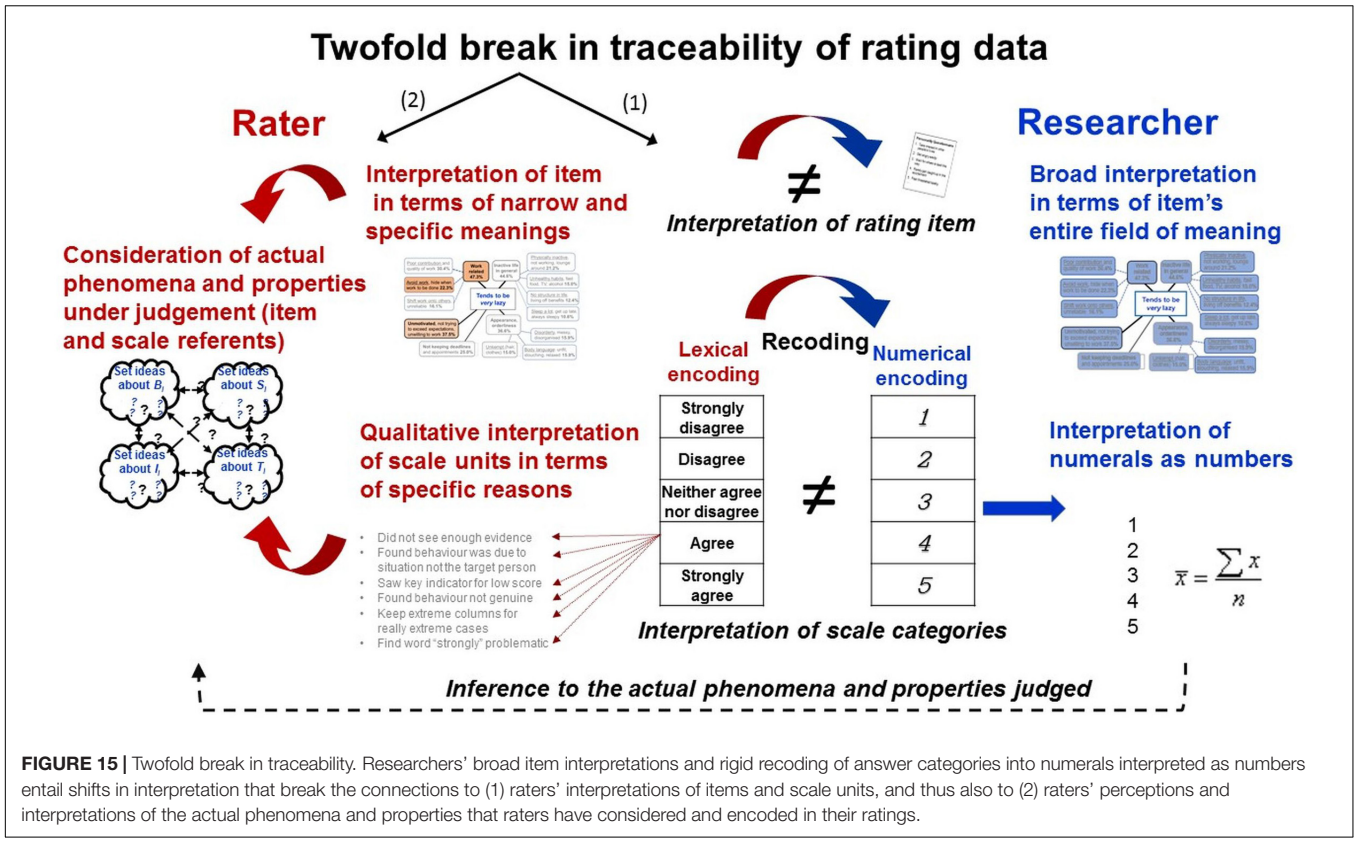

FIGURE 15 | Twofold break in traceability. Researchers' broad item interpretations and rigid recoding of answer categories into numerals interpreted as numbers entail shifts in interpretation that break the connections to (1) raters' interpretations of items and scale units, and thus also to (2) raters' perceptions and interpretations of the actual phenomena and properties that raters have considered and encoded in their ratings.

CONCLUSION

Application of the TPS-Paradigm's metatheoretical and methodological frameworks opened up novel perspectives on methods of quantitative data generation in psychology and social sciences. They showed that concepts from metrology can be meaningfully applied even if the objects of research are abstract constructs. But they also revealed serious limitations of rating methods.

Psychological and social-science concepts of 'measurement' were not built on metrological theories and not meant to meet metrological criteria. But when ratings are treated as 'quantitative' data and subjected to statistical analysis, and when their results are used to make inferences on and decisions about individuals and real-world problems, then the generation of these numerical data must conform to the principles of scientific (metrological) measurement. In the times of replication crises and Big Data, these methodological mismatches can no longer be ignored.

Quantitative Versus Qualitative Methods – An Inaccurate and Misleading Divide

The analyses showed that all quantitative methods presuppose qualitative categorizations because "quantities are *of* qualities" (Kaplan, 1964, p. 207). Objects of research can only be identified by their qualities. The common polarization of so-called 'quantitative methods' versus 'qualitative methods', reflects misconceptions of the measurement-theoretical foundations of scientific quantification.

Raters' explanations of their scale responses revealed that they consider rating scale units not as quantitative but rather as qualitatively different categories. Researchers increasingly consider this, such as by reporting percentages of raters who ticked particular categories rather than calculating averages or medians over rigidly assigned numbers. Unlike rating methods, many so-called 'qualitative methods', feature operational structures to establish object-dependent data generation processes and traceable outcomes (Uher, 2018a). Qualitative data thus-generated can be used to derive *post hoc* ratio-scaled quantifications, such as by computing frequencies of the occurrences of key themes in textual data (e.g., content analysis; Flick, 2014). In summary, all methods inherently explore qualitative properties of their objects of research and only some of them *additionally* enable these qualitative properties to be quantified.

The concept of semiotic representations illuminates a further controversy underlying the qualitative-quantitative debate. So-called quantitative researchers (using rating methods) focus on the interrelations between the signs' physical constituents (signifier; e.g., item statements) and their referents (e.g., target persons' behaviors), whereas so-called qualitative researchers focus on the signifiers' interrelations with the meanings (the signified) that particular persons construct for them (**Figure 3**). The former researchers tend to ignore the composite's psychical constituent, the latter its referent (see similarly Bhaskar, 1994).

This shows that the different epistemologies underlying so-called qualitative and quantitative methods are not *per se* incommensurate with one another. Rather, their proponents only focus on different aspects in the triadic relations inherent to semiotic representations. This metatheoretical concept will therefore be useful to help find common ground and develop integrative concepts and methodologies in the future.

Assessments on Rating Scales Are Not Measurements—Rating Data do Not Constitute Scientific Quantifications

Not every quantification is an outcome of measurement. For valid inferences from quantifications generated to properties of the actual phenomena under study, measurement processes must be established that are object-dependent, producing results that are subject-independent and thus traceable.

Key to scientific measurement and quantification is standardization. But not any kind of standardization fulfills the necessary requirements. Standardized scale presentation, administration, instruction and scoring are fundamental to rating methods (Walsh and Betz, 2000). But they standardize only the format of data encoding, not the ways in which raters actually generate the data. Therefore, assessments do not constitute measurements and should not be labeled as such. The current use of the term measurement in psychology and social sciences largely constitutes a cross-disciplinary jingle fallacy (same term denotes different concepts; Thorndike, 1903), which creates misunderstandings and hampers exchange and development.

Problematic Assumptions in Psychometrics

The numerals into which psychometricians rigidly recode raters' ticks on the scales do not constitute measurement-based quantifications. Rasch analysis and conjoint measurement, often assumed to enable quantitative measurement with rating data (Borsboom and Mellenbergh, 2004; Michell, 2014), are only methods for modeling data *once they have been generated*. These methods show that rating data, *as recoded and interpreted by the researchers* (i.e., units interpreted as reflecting numbers, items as reflecting broad fields of meanings) can exhibit particular quantitative properties (e.g., additivity). But these properties are obtained through rigorous psychometric variable selection that align the data generation process to statistical assumptions rather than to properties of the actual objects of research, thus precluding object-dependence.

This entails a *twofold break in traceability in the triadic interactions involved in human-generated data generation*—first, to raters' interpretation and use of the rating scales as methods, and second, to their perceptions and interpretations of the actual phenomena and properties under study. As a consequence, quantitative properties ascertained in psychometric analyses cannot be attributed to the actual referents of the raw data (e.g., target persons' properties) as conceived by the raters who have generated these data (**Figure 15**).

Consequences for the Replicability and Transparency of Data

The methodological problems involved in rating methods, especially the inability to establish traceable chains of information conversions from the objects of research to the outcomes of data generation, may constitute a major reason for the lack of replicability in psychology and social sciences not yet considered. "Robust measures", often proposed as a solution to this problem (Carpenter, 2012; Yong, 2012; Asendorpf et al., 2013), are unlikely to be attained with rating-based quantifications. On the contrary, the standardisations implemented in rating methods may lead to systematic errors because consistency in data structure is achieved at the cost of data accuracy in terms of standardized and traceable relations to the actual phenomena under study and the ways in which they were quantified (see Hammersley, 2013).

Consequences for the Validity and Utility of Data: Interpretability Presupposes Traceability

Nowadays, rating data can be generated quickly and at large scale (e.g., online-questionnaires; Buhrmester et al., 2011; Chandler and Paolacci, 2017; Buchanan and Scofield, 2018) producing floods of data—Big Data. But to answer research questions and to find solutions for real-world problems, scientists must eventually interpret the data produced. This article showed that, in the process of *data generation*, information must be converted from perceptions and conceptions of the study phenomena into signs. But in the process of *data interpretation*, information must be converted in the reverse direction from the signs back to conceptions and ideas about the actual phenomena under study. Such backward conversions of information may not be straightforwardly possible because signs, especially mathematical ones, can be abstracted, processed and changed in ways not applicable to the properties of the actual study phenomena (Brower, 1949; Trierweiler and Stricker, 1998), highlighting the importance of traceability not only in data generation but also in data analysis.

Major Tasks Still Laying Ahead

As interpretations of rating scales are based on everyday knowledge with its fuzzy and flexible categories, any interpretation of rating data can appear plausible (Laucken, 1974). But the purpose of scientific measurement is to quantify phenomena in the real world—not to construe a possible match with data that can be generated even in absence of the persons, phenomena and properties under study. Therefore, traceability is a fundamental requirement for scientific quantification that should be implemented systematically also in the methods used to generate quantitative data in psychology and the social sciences. This article started to elaborate some principles by which this can be achieved.

Psychologists and social scientists must finally investigate how people actually understand and use rating scales to generate quantitative data in research and applied contexts. Exploring raters' mental processes and the meanings they

attribute to items and scale categories is key to specifying the representational systems underlying rating data, which, in many fields, make up much of the current empirical data basis.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The author gratefully acknowledges funding from a Marie Curie Fellowship (EC Grant Agreement Number 629430).

ACKNOWLEDGMENTS

The author thanks the editor and the three reviewers for their thoughtful comments on previous drafts of the manuscript.

REFERENCES

Abran, A., Desharnais, J.-M., and Cuadrado-Gallego, J. J. (2012). Measurement and quantification are not the same: ISO 15939 and ISO 9126. *J. Softw. Evol. Process* 24, 585–601. doi: 10.1002/smr.496

Arro, G. (2013). Peeking into personality test answers: inter- and intraindividual variety in item interpretations. *Integr. Psychol. Behav. Sci.* 47, 56–76. doi: 10.1007/s12124-012-9216-9

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *Eur. J. Pers.* 27, 108–119. doi: 10.1002/per.1919

Baumeister, R. F., Vohs, K. D., and Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: whatever happened to actual behavior? *Perspect. Psychol. Sci.* 2, 396–403. doi: 10.1111/j.1745-6916.2007.00051.x

Berglund, B. (2012). "Measurement in psychology," in *Measurement with Persons?: theory, Methods, and Implementation Areas*, eds B. Berglund and G. B. Rossi (New York, NY: Taylor and Francis), 27–50.

Bhaskar, R. (1994). *Plato etc.: Problems of Philosophy and their Resolution*. London: Verso.

Bhaskar, R., and Danermark, B. (2006). Metatheory, interdisciplinarity and disability research: a critical realist perspective. *Scand. J. Disabil. Res.* 8, 278–297. doi: 10.1080/15017410600914329

Biesanz, J. C., and Human, L. J. (2010). The cost of forming more accurate impressions: accuracy-motivated perceivers see the personality of others more distinctively but less normatively than perceivers without an explicit goal. *Psychol. Sci.* 21, 589–594. doi: 10.1177/0956797610364121

BIPM (2006). *BIPM: The International System of Units (SI)*, 8th Edn. Paris: Organisation Intergouvernementale de la Convention du Mètre.

Blanton, H., and Jaccard, J. (2006). Arbitrary metrics in psychology. *Am. Psychol.* 61, 27–41. doi: 10.1037/0003-066X.61.1.27

Block, J. (2010). The Five-Factor framing of personality and beyond: some ruminations. *Psychol. Inq.* 21, 2–25. doi: 10.1080/10478401003596626

Bohr, N. (1937). Causality and complementarity. *Philos. Sci.* 4, 289–298. doi: 10.1086/286465

Boring, E. G. (1923). Intelligence as the tests test it. *New Repub.* 36, 35–37.

Borkenau, P., and Müller, B. (1991). Breadth, bandwidth, and fidelity of personality-descriptive categories. *Eur. J. Pers.* 5, 309–322. doi: 10.1002/per.2410050404

Borsboom, D., and Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory Psychol.* 14, 105–120. doi: 10.1177/0959354304040200

Bridgman, P. W. (1927). *The Logic of Modern Physics*. New York, NY: Macmillan.

Bringmann, L. F., and Eronen, M. I. (2015). Heating up the measurement debate: what psychologists can learn from the history of physics. *Theory Psychol.* 26, 27–43. doi: 10.1177/0959354315617253

Brody, N., and Oppenheim, P. (1969). Application of Bohr's principle of complementarity to the mind-body problem. *J. Philos.* 66, 97–113. doi: 10.2307/2024529

Brower, D. (1949). The problem of quantification of psychological science. *Psychol. Rev.* 56, 325–331. doi: 10.1037/h0061802

Bruschi, A. (2017). Measurement in social research: some misunderstandings. *Qual. Quant.* 51, 2219–2243. doi: 10.1007/s11135-016-0383-5

Buchanan, E. M., and Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behav. Res. Methods* 50, 2586–2596. doi: 10.3758/s13428-018-1035-6

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Buntins, M., Buntins, K., and Eggert, F. (2016). Psychological tests from a (fuzzy-)logical point of view. *Qual. Quant.* 50, 2395–2416. doi: 10.1007/s11135-015-0268-z

Campbell, N. R. (1920). *Physics: The Elements*. Cambridge: Cambridge University Press.

Capra, F. (1997). *The Web of Life: A New Synthesis of Mind and Matter*. New York, NY: Anchor Books.

Carpenter, S. (2012). Psychology's bold initiative. *Science* 335, 1558–1561. doi: 10.1126/science.335.6076.1558

Chalmers, A. F. (2013). *What is this Thing Called Science?* Indianapolis, IN: Hackett Publishing.

Chandler, J. J., and Paolacci, G. (2017). Lie for a dime. *Soc. Psychol. Pers. Sci.* 8, 500–508. doi: 10.1177/1948550617698203

Chang, H. (2004). *Inventing Temperature?: Measurement and Scientific Progress*. Oxford: Oxford University Press. doi: 10.1093/0195171276.001.0001

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Creswell, J. W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 2nd Edn. Thousand Oaks, CA: Sage.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

Daston, L., and Galison, P. (2007). *Objectivity*. New York, NY: Zone Books.

Dingle, H. (1950). A theory of measurement. *Br. J. Philos. Sci.* 1, 5–26. doi: 10.1093/bjps/I.1.5

Doliński, D. (2018). Is psychology still a science of behaviour? *Soc. Psychol. Bull.* 13:e25025. doi: 10.5964/spb.v13i2.25025

Ellis, B. (1966). *Basic Concepts of Measurement*. Cambridge: Cambridge University Press.

Fahrenberg, J. (1979). The complementarity principle in psychophysiological research and somatic medicine. *Z. Klin. Psychol. Psychother.* 27, 151–167.

Fahrenberg, J. (2013). *Zur Kategorienlehre der Psychologie: Komplementaritätsprinzip; Perspektiven und Perspektiven-Wechsel*. Lengerich: Pabst Science Publishers.

Fahrenberg, J., Myrtek, M., Pawlik, K., and Perrez, M. (2007). Ambulatory assessment - monitoring behavior in daily life settings. *Eur. J. Psychol. Assess.* 23, 206–213. doi: 10.1027/1015-5759.23.4.206

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., et al. (1940). Quantitative estimates of sensory events: final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Adv. Sci.* 1, 331–349.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement* 34, 39–48. doi: 10.1016/S0263-2241(03)00018-6

Flick, U. (2014). *An Introduction to Qualitative Research*. Thousand Oaks, CA: Sage.

Frigerio, A., Giordani, A., and Mari, L. (2010). Outline of a general model of measurement. *Synthese* 175, 123–149. doi: 10.1007/s11229-009-9466-3

Giordani, A., and Mari, L. (2012). Measurement, models, and uncertainty. *IEEE Trans. Instrum. Meas.* 61, 2144–2152. doi: 10.1109/TIM.2012.2193695

Giordani, A., and Mari, L. (2014). "Modeling measurement: error and uncertainty," in *Error and Uncertainty in Scientific Practice*, eds M. Boumans, G. Hon, and A. Peterson (London: Pickering & Chatto), 79–96.

Grzyb, T. (2016). Why can't we just ask? The influence of research methods on results. The case of the "bystander effect." *Pol. Psychol. Bull.* 47, 233–235. doi: 10.1515/ppb-2016-0027

Guyon, H., Kop, J.-L., Juhel, J., and Falissard, B. (2018). Measurement, ontology, and epistemology: psychology needs pragmatism-realism. *Theory Psychol.* 28, 149–171. doi: 10.1177/0959354318761606

Hammersley, M. (2013). *The Myth of Research-Based Policy and Practice*. London: SAGE Publications, doi: 10.4135/9781473957626

Hanel, P. H. P., and Vione, K. C. (2016). Do student samples provide an accurate estimate of the General Public? *PLoS One* 11:e0168354. doi: 10.1371/journal.pone.0168354

Hartmann, N. (1964). *Der Aufbau der Realen Welt. Grundriss der Allgemeinen Kategorienlehre (3. Aufl.)*. Berlin: Walter de Gruyter. doi: 10.1515/9783110823844

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Z. Phys.* 43, 172–198. doi: 10.1007/BF01397280

Heisenberg, W. (1989). *Encounters with Einstein: And other Essays on People, Places, and Particles*. Princeton, NJ: Princeton University Press.

Hirschberger, J. (1980). *Geschichte der Philosophie Band II Neuzeit und Gegenwart*. Frankfurt am Main: Zweitausendeins.

Hoche, H.-U. (2008). *Anthropological Complementarianism. Linguistic, Logical, and Phenomenological Studies in Support of a Third Way Beyond Dualism and Monism*. Paderborn: Mentis.

Hogan, J., and Roberts, B. W. (1996). Issues and non-issues in the fidelity–bandwidth trade-off. *J. Organ. Behav.* 17, 627–637. doi: 10.1002/(SICI)1099-1379(199611)17:6<627::AID-JOB2828>3.0.CO;2-F

Hölder, O. (1901). *Die Axiome der Quantität und die Lehre vom Mass (Band 53). Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig*. Leipzig: Mathematisch-Physische Classe.

Hossenfelder, S. (2018). *Lost in Math: How Beauty Leads Physics Astray*. New York, NY: Basic Books.

Humphry, S. M. (2017). Psychological measurement: theory, paradoxes, and prototypes. *Theory Psychol.* 27, 407–418. doi: 10.1177/0959354317699099

JCGM200:2012 (2012). *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*, 3rd Edn. Paris: Joint Committee for Guides in Metrology.

Kant, I. (1781/1998). *Kritik der Reinen Vernunft*, ed. J. Timmermann (Hamburg: Felix Meiner Verlag). doi: 10.28937/978-3-7873-2112-4

Kant, I. (1786/2016). *Metaphysische Anfangsgründe der Naturwissenschaft*, ed. M. Holzinger (Scotts Valley, CA: CreateSpace).

Kaplan, A. (1964). *The Conduct of Inquiry: Methodology for Behavioral Science*. Scranton, PA: Chandler Publishing Co.

Koffka, K. (1935). *Principles of Gestalt Psychology*. New York, NY: Harcourt, Brace, & World.

Krantz, D., Luce, R. D., Tversky, A., and Suppes, P. (1971). *Foundations of Measurement Volume I: Additive and Polynomial Representations*. San Diego, CA: Academic Press.

Lahlou, S. (1998). *Penser Manger*. Paris: Les Presses Universitaires de France. doi: 10.3917/puf.insti.1998.01

Laucken, U. (1974). *Naive Verhaltenstheorie*. Stuttgart: Klett.

Le Poidevin, R. (2011). *The Experience and Perception of Time*, 2011 Edn, ed. E. N. Zalta (Stanford, CA: The Stanford Encyclopedia of Philosophy).

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 22, 1–55.

Locke, J. (1999). *An Essay Concerning Human Understanding. Electronic Classics Series*. University Park, PA: The Pennsylvania State University.

Logan, R. K. (2007). *The Extended Mind: The Emergence of Language, the Human mind, and Culture*. Toronto: University of Toronto Press. doi: 10.3138/9781442684911

Ludeke, S. G., and Larsen, E. G. (2017). Problems with the Big Five assessment in the World Values Survey. *Pers. Individ. Dif.* 112, 103–105. doi: 10.1016/j.paid.2017.02.042

Lundmann, L., and Villadsen, J. W. (2016). Qualitative variations in personality inventories: subjective understandings of items in a personality inventory. *Qual. Res. Psychol.* 13, 166–187. doi: 10.1080/14780887.2015.1134737

Mari, L. (2013). A quest for the definition of measurement. *Measurement* 46, 2889–2895. doi: 10.1016/J.MEASUREMENT.2013.04.039

Mari, L., Carbone, P., Giordani, A., and Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Stud. Hist. Philos. Sci.* 65–66, 46–56. doi: 10.1016/j.shpsa.2017.08.001

Mari, L., Carbone, P., and Petri, D. (2012). Measurement fundamentals: a pragmatic view. *IEEE Trans. Instrum. Meas.* 61, 2107–2115. doi: 10.1109/TIM.2012.2193693

Maul, A., Mari, L., Torres Irribarra, D., and Wilson, M. (2018). The quality of measurement results in terms of the structural features of the measurement process. *Measurement* 116, 611–620. doi: 10.1016/J.MEASUREMENT.2017.08.046

McGrane, J. A. (2015). Stevens' forgotten crossroads: the divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Front. Psychol.* 6:431. doi: 10.3389/fpsyg.2015.00431

Menon, G., and Yorkston, E. A. (2000). "The Use of memory and contextual cues in the formation of behavioral frequency judgments," in *The Science of Self-Report: Implications for Research and Practice*, eds A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, and V. S. Cain (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 63–80.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *Br. J. Psychol.* 88, 355–383. doi: 10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in Psychology. A Critical History of a Methodological Concept*. Cambridge: Cambridge University Press, doi: 10.1017/CBO9780511490040

Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: response to Humphry and Sijtsma. *Theory Psychol.* 24, 111–123. doi: 10.1177/0959354313517524

Morin, E. (2008). *On Complexity*. Cresskill, NJ: Hampton Press.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture: author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science* 348, 1422–1425. doi: 10.1126/science.aab2374

Omi, Y. (2012). Tension between the theoretical thinking and the empirical method: is it an inevitable fate for psychology? *Integr. Psychol. Behav. Sci.* 46, 118–127. doi: 10.1007/s12124-011-9185-4

Ones, D. S., and Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *J. Organ. Behav.* 17, 609–626. doi: 10.1002/(SICI)1099-1379(199611)17:6<609::AID-JOB1828>3.0.CO;2-K

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716

Peirce, C. S. (1958). *Collected Papers of Charles Sanders Peirce, Vols. 1–6, Vols. 7–8*, eds C. Hartshorne, P. Weiss and A. W. Burks (Cambridge, MA: Harvard University Press).

Pendrill, L. (2014). Man as a measurement instrument. *NCSL Int. Meas.* 9, 24–35. doi: 10.1080/19315775.2014.11721702

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Prigogine, I., and Stengers, I. (1984). *Order Out of Chaos?: Man's New Dialogue with Nature*. New York, NY: Bantam Books.

Rammstedt, B., and John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. *J. Res. Pers.* 41, 203–212. doi: 10.1016/j.jrp.2006.02.001

Rosenbaum, P. J., and Valsiner, J. (2011). The un-making of a method: from rating scales to the study of psychological processes. *Theory Psychol.* 21, 47–65. doi: 10.1177/0959354309352913

Rothschuh, K. E. (1963). *Theorie des Organismus. Bios – Psyche – Pathos (2. erw. Aufl.)*. München: Urban & Schwarzenberg.

Schacter, D. L. (1999). The seven sins of memory. Insights from psychology and cognitive neuroscience. *Am. Psychol.* 54, 182–203. doi: 10.1037/0003-066X.54.3.182

Schacter, D. L., and Addis, D. R. (2007). Constructive memory: the ghosts of past and future. *Nature* 445:27. doi: 10.1038/445027a

Schwitzgebel, E. (2016). *Introspection. In Stanford Encyclopedia of Philosophy. (Winter 2016)*. Stanford, CA: Metaphysics Research Lab.

Shweder, R. A., and D'Andrade, R. G. (1980). "The systematic distortion hypothesis," in *Fallible Judgment in Behavioral Research: New Directions for Methodology of Social and Behavioral Science*, Vol. 4, ed. R. A. Shweder (San Francisco, CA: Jossey-Bass), 37–58.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science* 103, 667–680. doi: 10.1126/science.103.2684.677

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. Stanford, CA: CSLI Publications.

Tafreshi, D., Slaney, K. L., and Neufeld, S. D. (2016). Quantification in psychology: critical analysis of an unreflective practice. *J. Theor. Philos. Psychol.* 36, 233–249. doi: 10.1037/teo0000048

Thissen, D. (2001). Psychometric engineering as art. *Psychometrika* 66, 473–486. doi: 10.1007/BF02296190

Thorndike, E. L. (1903). *Notes on Child Study*, 2nd Edn. New York, NY: Macmillan.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* 34, 273–286. doi: 10.1037/h0070288

Thurstone, L. L. (1928). Attitudes can be measured. *Am. J. Soc.* 33, 529–554. doi: 10.1086/214483

Titchener, E. B. (1905). *Experimental Psychology: A Manual of Laboratory Practice: Quantitative Experiments, Part 1, Students Manual*, Vol 2. New York, NY: MacMillan Co. doi: 10.1037/13683-000

Toomela, A., and Valsiner, J. (2010). *Methodological Thinking in Psychology?: 60 Years Gone Astray?* Charlotte, NC: Information Age Publishing.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory Psychol.* 19, 579–599. doi: 10.1177/0959354309341926

Trendler, G. (2018). Conjoint measurement undone. *Theory Psychol.* 1–29. (in press). doi: 10.1177/0959354318788729

Trierweiler, S. J., and Stricker, G. (1998). *The Scientific Practice of Professional Psychology*. Boston, MA: Springer. doi: 10.1007/978-1-4899-1944-1

Uher, J. (2011). Individual behavioral phenotypes: an integrative meta-theoretical framework. Why "behavioral syndromes" are not analogs of "personality." *Dev. Psychobiol.* 53, 521–548. doi: 10.1002/dev.20544

Uher, J. (2013). Personality psychology: lexical approaches, assessment methods, and trait concepts reveal only half of the story-Why it is time for a paradigm shift. *Integr. Psychol. Behav. Sci.* 47, 1–55. doi: 10.1007/s12124-013-9230-6

Uher, J. (2015a). "Agency enabled by the psyche: explorations using the transdisciplinary philosophy-of-science paradigm for research on individuals," in *Constraints of Agency: Explorations of theory in Everyday Life. Annals of Theoretical Psychology*, Vol. 12, eds C. W. Gruber, M. G. Clark, S. H. Klempe, and J. Valsiner (New York, NY: Springer International Publishing), 177–228. doi: 10.1007/978-3-319-10130-9_13

Uher, J. (2015b). "Comparing individuals within and across situations, groups and species: metatheoretical and methodological foundations demonstrated in primate behaviour," in *Comparative Neuropsychology and Brain Imaging*, Vol. 2, eds D. Emmans and A. Laihinen (Berlin: Lit Verlag), 223–284. doi: 10.13140/RG.2.1.3848.8169

Uher, J. (2015c). Conceiving "personality": psychologist's challenges and basic fundamentals of the transdisciplinary philosophy-of-science paradigm for research on individuals. *Integr. Psychol. Behav. Sci.* 49, 398–458. doi: 10.1007/s12124-014-9283-1

Uher, J. (2015d). Developing "personality" taxonomies: metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integr. Psychol. Behav. Sci.* 49, 531–589. doi: 10.1007/s12124-014-9280-4

Uher, J. (2015e). Interpreting "personality" taxonomies: why previous models cannot capture individual-specific experiencing, behaviour, functioning and development. Major taxonomic tasks still lay ahead. *Integr. Psychol. Behav. Sci.* 49, 600–655. doi: 10.1007/s12124-014-9281-3

Uher, J. (2016a). "Exploring the workings of the Psyche: metatheoretical and methodological foundations," in *Psychology as the Science of Human Being: The Yokohama Manifesto*, eds J. Valsiner, G. Marsico, N. Chaudhary, T. Sato, and V. Dazzani (New York, NY: Springer International Publishing), 299–324. doi: 10.1007/978-3-319-21094-0_18

Uher, J. (2016b). What is behaviour? And (when) is? language behaviour? A metatheoretical definition. *J. Theory Soc. Behav.* 46, 475–501. doi: 10.1111/jtsb.12104

Uher, J. (2018a). Data generation methods across the empirical sciences: differences in the study phenomena's accessibility and the processes of data encoding. *Qual. Quant.* 1–26. (in press). doi: 10.1007/s11135-018-0744-3

Uher, J. (2018b). Taxonomic models of individual differences: a guide to transdisciplinary approaches. *Philos. Trans. R. Soc. B* 373:20170171. doi: 10.1098/rstb.2017-0171

Uher, J. (2018c). "The transdisciplinary philosophy-of-science paradigm for research on individuals: foundations for the science of personality and individual differences," in *The SAGE Handbook of Personality and Individual Differences: Volume I: The Science of Personality and Individual Differences*, eds V. Zeigler-Hill and T. K. Shackelford (London: SAGE), 84–109. doi: 10.4135/9781526451163.n4

Uher, J., Addessi, E., and Visalberghi, E. (2013a). Contextualised behavioural measurements of personality differences obtained in behavioural tests and social observations in adult capuchin monkeys (*Cebus apella*). *J. Res. Pers.* 47, 427–444. doi: 10.1016/j.jrp.2013.01.013

Uher, J., Werner, C. S., and Gosselt, K. (2013b). From observations of individual behaviour to social representations of personality: developmental pathways, attribution biases, and limitations of questionnaire methods. *J. Res. Pers.* 47, 647–667. doi: 10.1016/j.jrp.2013.03.006

Uher, J., and Visalberghi, E. (2016). Observations versus assessments of personality: a five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments. *J. Res. Pers.* 61, 61–79. doi: 10.1016/j.jrp.2016.02.003

Valsiner, J. (2012). *A Guided Science: History of Psychology in the Mirror of Its Making*. New Brunswick, NJ: Transaction Publishers.

Valsiner, J. (2017). *From Methodology to Methods in Human Psychology*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-61064-1

Valsiner, J., Diriwächter, R., and Sauck, C. (2005). "Diversity in Unity: standard questions and nonstandard interpretations," in *Science and Medicine in Dialogue*, eds R. Bibace, J. Laird, K. Noller, and J. Valsiner (Westport, CT: Praeger-Greenwood), 289–307.

van der Maas, H., Kan, K.-J., and Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *J. Intell.* 2, 12–15. doi: 10.3390/jintelligence2010012

Vautier, S., Veldhuis, M., Lacot, É., and Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative foundation of socially relevant avatars. *Theory Psychol.* 22, 810–822. doi: 10.1177/0959354312450093

Vessonen, E. (2017). Psychometrics versus representational theory of measurement. *Philos. Soc. Sci.* 47, 330–350. doi: 10.1177/0048393117705299

von Bertalanffy, L. (1937). *Das Gefüge des Lebens*. Leipzig: Teubner.

von Glasersfeld, E. (1991). "Knowing without metaphysics: aspects of the radical constructivist position," in *Research and Reflexivity*, ed. F. Steier (London: Sage), 12–29.

von Helmholtz, H. (1887). *Zählen und Messen, Erkenntnistheoretisch Betrachtet*. Leipzig: Fuess Verlag.

Vygotsky, L. S. (1962). *Thought and Language*. Cambridge, MA: MIT Press. doi: 10.1037/11193-000

Vygotsky, L. S., and Luria, A. (1994). "Tool and symbol in child development. Reprinted," in *The Vygotsky Reader*, eds R. van der Veer and J. Valsiner (Oxford: Blackwell), 99–174.

Wagenmakers, E. J., Verhagen, J., and Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behav. Res. Methods* 48, 413–426. doi: 10.3758/s13428-015-0593-0

Wagoner, B., and Valsiner, J. (2005). "Rating tasks in psychology: from a static ontology to a dialogical synthesis of meaning," in *Contemporary Theorizing in Psychology?: Global Perspectives*, eds A. Gülerce, I. Hofmeister, G. Saunders, and J. Kaye (Toronto: Captus), 197–213.

Walach, H. (2013). *Psychologie: Wissenschaftstheorie, Philosophische Grundlagen Und Geschichte (3. Aufl.)*. Stuttgart: Kohlhammer.

Walach, H., and Römer, H. (2011). Complementarity is a useful concept for consciousness studies: a reminder. *Neuroendocrinol. Lett.* 21, 221–232.

Walsh, W. B., and Betz, N. E. (2000). *Test and Assessment*, 4th Edn. Upper Saddle River, NJ: Prentice-Hall.

Westen, D. (1996). A model and a method for uncovering the nomothetic from the idiographic: an alternative to the Five-Factor Model. *J. Res. Pers.* 30, 400–413. doi: 10.1006/jrpe.1996.0028

Whitehead, A. N. (1929). *Process and Reality*. New York, NY: Harper.

Wundt, W. (1863). *Vorlesungen Über die Menschen- und Thierseele*. Hamburg: Voss.

Wundt, W. (1894). *Grundriss der Psychologie*, Vol. 1. Leipzig: Engelmann. doi: 10.1037/h0067923

Wundt, W. (1907). *Logik der Exakten Wissenschaften, Band II (3. Umgearb. Aufl.)*. Stuttgart: Enke.

Yong, E. (2012). Replication studies: bad copy. *Nature* 485, 298–300. doi: 10.1038/485298a

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2017). Making replication mainstream. *Behav. Brain Sci.* doi: 10.1017/S0140525X17001972

Conflict of Interest Statement: 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.

frontiers in Psychology

BOOK REVIEW
published: 04 April 2018
doi: 10.3389/fpsyg.2018.00455

Book Review: Another Science Is Possible

Jose D. Perezgonzalez[1]*, *Dolores Frías-Navarro*[2] and *Juan Pascual-Llobell*[2]

[1] Business School, Massey University, Palmerston North, New Zealand, [2] Department of Methodology of the Behavioral Sciences, Universitat de València, Valencia, Spain

Keywords: science, society, slow science, philosophy, error statistics, Bayesian inference

A Book Review on
Another Science Is Possible. A Manifesto for Slow Science

Isabelle Stengers (Cambridge, UK: Polity Press), 2018, 163 pages, ISBN: 9781509521807.

The philosopher of science Isabelle Stengers provides some food for thought regarding both the way we are doing science and the need for an alternative approach likened to the slow movement in other spheres of life.

The title of the book already promises a dialectical contrast between contemporary and another form of science, and between fast and slow science. The remainder of the book does not disappoint in such strategy. Indeed, Stengers does a good job in focusing on different contrasts in the five main chapters comprising the book (the sixth and last chapter mostly wraps up what had been said before).

Stengers's chief contrast is between Science and Society: Science pursuant of knowledge, of facts, of right answers to specific problems by specialist people; Society as the net beneficiary of Science's work but also as a mass which confuses facts and values because it often lacks the scientific literacy to spot the difference (Ch. 1). Stengers argues against Science's technocratic mindset and in favor of Society's democracy, which needs from Science contextualized answers to its social concerns and the cultivation of a public intelligence of connoisseurs.

Stengers next uses gender in lieu of "marked" scientists to identify a second contrast, that between "hard" (or "sound") sciences and "soft" sciences (Ch. 2). For Stengers, Science is mostly about mimicking the hard sciences, about scientists having the "right stuff," focused on facts and laboratory objectivity, mobilized in serving industrial interests. "Marked" scientists are those who deviate from above ideal to become concerned with social matters, either historically (women) or contemporarily (youth avoiding the hard sciences, and scientists inclined toward "soft" matters).

As the book progresses, Stengers tackles contemporary research autonomy and evaluation, identified as "fast" science and intimately correlated with competitive evaluation, publication in high-impact-factor journals, inbreeding review by peers, and industrial capture of financial research resources (Ch. 3). By contrast, Stengers calls for a contested evaluation, a slow-down of publications and peer-review, and a reclaiming of social interdependency as a definition for scientific excellence.

She follows such call by explicitly linking to the 2010 "Slow Science Manifesto" (The Slow Science Academy, 2010), which she contrasts against her own idea of slow science (Ch. 4). For Stengers, slow science is not about returning to the (fast science) golden era where scientists were autonomous and respected, but about creating a collective awareness and appreciation for Society among scientists (i.e., for them to "become civilized").

OPEN ACCESS

Edited by:
Ulrich Dettweiler,
University of Stavanger, Norway

Reviewed by:
Andrew Edgar,
Cardiff University, United Kingdom

***Correspondence:**
Jose D. Perezgonzalez
j.d.perezgonzalez@massey.ac.nz

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *22 February 2018*
Accepted: *19 March 2018*
Published: *04 April 2018*

Citation:
Perezgonzalez JD, Frías-Navarro D and Pascual-Llobell J (2018) Book Review: Another Science Is Possible. Front. Psychol. 9:455. doi: 10.3389/fpsyg.2018.00455

Finally, Stengers brings her slow science plea to academia, and tasks the university with creating such a future for slow science, of complementing the reliability of the laboratory with the reliability of the context of application, and of bringing value to facts (Ch. 5).

Ultimately, the book delivers a different idea than its title promised. It is not about another science, but about contemporary science communicated and applied differently, more attuned to Society's milieu. Nor is the book in line with the Berlin manifesto for slow science but about Science slowing down so that it can be successful in the above form of attunement.

The book also has two small drawbacks. One is stylistic: Stengers did not apply to her own philosophy her criticisms of what Science is doing, insofar her book has not left her own "Ivory Tower" of circumloquacious writing and conceptual detours ending in cul-de-sacs, possibly highly appreciated by her peers but taxing other readers unnecessarily (indeed, about 80% of the text could be safely dismissed without affecting the main ideas in the book).

The second drawback is implementation: Stengers takes herself out of the fight by book's end, in a way reminiscent of a criticism she had earlier laid onto scientists, as it seems she equally "[does] not feel there is an option at all" (p. 110). Her calls are, thus, "only suggestions…to try to activate the imagination" (p. 124), "a little derisory" (p. 142), "a philosopher['s]…dream, for such a counterfactual story" (p. 144).

And yet, all the time we have spent reading (and re-reading) Stengers's book, we kept wondering about a related contrast, that of the statistics wars between frequentists and Bayesians. Indeed, not long ago, another philosopher of science, Deborah Mayo, lashed out against Bayesians in what parallels a defense—by Mayo—of current practices of laboratory research for "warranting a scientific research claim, or learning about a substantive phenomenon of interest" (Mayo, 2017a). She correctly argued that "in an adequate account [of severity testing], the improbability of a claim must be distinguished from its having been poorly tested. (You need to be able to say things like, 'it's plausible, but that's a lousy test of it.')" (Mayo, 2017b). The relevance of Mayo's stance in favor of research objectivity and severe testing needs to be defended. However, Mayo did not tackle the alternative consequence to her claim, an alternative which underlies Stengers's ideas: that you also need to be able to say things like, "it may have been reliably tested, but its social reliability is nonetheless lousy."

This contrast between claims that need to be severely tested (e.g., Mayo and Spanos, 2010) and applications that need to be reliably assessed in the wider context of application thus suggests a method for scientists to move from the laboratory to the social milieu: Bayesian inference (e.g., Kruschke, 2011). With a Bayesian inference built upon error statistics, Stengers's contextual reliability would combine with scientific reliability to respond to the important question regarding the (subjective) value of an (objective) fact, both before implementation as well as throughout the life-cycle of those solutions already implemented. The initial advantage of this method rests on the preference scientists already have toward quantification and formulation, yet forces them to further consider those social "matters of concern" that may escape them in their daily scientific milieu. This method may, thus, provide substance to Stengers's slow science manifesto and a practical solution to its implementation.

AUTHOR CONTRIBUTIONS

JP acquired the funding, and initiated and drafted the review. DF-N, JP-L and JP revised and edited the manuscript. All authors approved the final version of the manuscript for submission.

FUNDING

This publication was financially supported by the Massey University Research Fund (MURF) Mini Grant 2017, Massey University, New Zealand.

REFERENCES

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis.* Oxford, UK: Academic Press.

Mayo, D. G. (2017a). "*A Megateam of Reproducibility-Minded Scientists" Look to Lowering the p-value [Web log post]*. Available online at: https://errorstatistics.com/2017/07/25/a-megateam-of-reproducibility-minded-scientists-look-to-lowering-the-p-value

Mayo, D. G. (2017b). *New Venues for the Statistics Wars [Web log post]*. Available online at: https://errorstatistics.com/2017/10/05/new-venues-for-the-statistics-wars

Mayo, D. G., and Spanos, A. (eds.). (2010). *Error and Inference*. New York, NY: Cambridge University Press.

The Slow Science Academy (2010). *The Slow Science Manifesto*. Available online at: http://slow-science.org/

Conflict of Interest Statement: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership