Research Problem Validity in Primary Research:
Precision and Transparency in Characterizing Past Knowledge

Martin Schweinsberg
ESMT Berlin

Stefan Thau
INSEAD

Madan Pillutla
Indian School of Business

Acknowledgments: The authors would like to thank Kelly A. Nault, Marko Pitesa, Phanish Puranam, Michael Schaerer, Kaisa Snellman, Rafael Wittek, and seminar participants at the University of Groningen for helpful comments on an earlier version of this paper. The authors would like to thank Katherine L. McLain for her help as research assistant.

Corresponding author: Martin Schweinsberg, ESMT Berlin, Schlossplatz 1, 10178 Berlin, Germany

Abstract

Four validity types evaluate the approximate truth of inferences communicated by primary research. However, current validity frameworks ignore the truthfulness of empirical inferences that are central to research problem statements. Problem statements contrast a review of past research with other knowledge that extend, contradict, or call into question specific features of past research. Authors communicate empirical inferences, or quantitative judgments about the frequency (e.g., "few," "most") and variability (e.g., "on the one hand, on the other hand") in their reviews of existing theories, measures, samples, or results. We code a random sample of primary research articles and show that 83% of quantitative judgments in our sample are both vague and their origin non-transparent, making it difficult to assess their validity. We review validity threats of current practices. We propose that documenting the literature search, how the search was coded, along with quantification facilitates more precise judgments and makes their origin transparent. This practice enables research questions that are more closely tied to the existing body of knowledge and allows for more informed evaluations of the contribution of primary research articles, their design choices, and how they advance knowledge. We discuss potential limitations of our proposed framework.

Word Count Abstract: 196

Keywords: validity; reproducibility; open science; transparency; research process

The goal of this paper is to define, examine, and discuss the validity of research problems in primary psychological research. The psychological research process starts with (0) an idea about the phenomenon of interest, followed by (1) a research problem statement which includes a literature review of past research on the phenomenon and the research question the studies seek to answer (ideas can also follow from a literature review); followed by (2) theory and in the case of confirmatory research, predictions to answer the question; (3) study designs which use sampling, manipulation and measurement; and (4) data analyses and discussion of study results to assess the extent to which they solve the research problem and answer the research question (see e.g., Kerlinger, 1986; Trochim, 2006). Exploratory research follows a similar process, with the goal of generating, rather than testing predictions and hypotheses (Swedberg, 2020). Collectively, empirical research articles advance the literature by prompting new research problems and questions.

Currently, an evaluation of validity takes place at Steps 3 (study design) and 4 (data analysis, results, and discussion) of the research process, in which researchers both document and communicate their inferences, or judgments, about issues involving, for example, causality, effect sizes, measurement, or generalizability. Based on the documentation provided, readers of such research can scrutinize and evaluate the validity of researcher judgments and assess the extent to which relevant evidence supports the communicated inferences as true or correct (Shadish et al., 2002, p. 34).

We argue that validity standards can also be meaningfully applied to what methodologists refer to as the *problem statement* (Campbell et al., 1982; Gall et al., 1996; Kerlinger, 1986), in which an idea is justified as worth studying by contrasting it with what is already known in the literature. In problem statements, authors characterize aspects of past research through quantitative judgments, which are inferences about quantities or amounts (e.g., "*most* theories of past research on X-Y have….;" "the findings on X-Y are *inconsistent*”). These judgments are then contrasted with knowledge on the phenomenon (e.g., findings, theories, assumptions, or unstudied aspects of the phenomenon) which extend, contradict, or call into question the reviewed body of knowledge. The “contribution” of a primary research article is to offer an answer to the resulting research question and solution to the research problem (Gall et al., 1996). The characterization of past research in quantitative terms involves empirical inferences that can be, to varying degrees truthful, accurate, or *valid* but we currently pay, as our empirical analyses show, *little* attention to it.

Our main argument is that the quantitative judgments used in problem statements in primary research are both vague (instead of precise) and the origin of these judgments is obscured (instead of made transparent). In contrast, systematic reviews in secondary research (Denyer & Tranfield, 2009; Siddaway et al., 2019) make transparent the search parameters of the literature review. Authors of systematic reviews are called upon to detail which literature was included in the review and explain how the reviewed articles were counted, coded, and classified to arrive at a precise judgment about past research. We believe that the reviews made to justify the purpose of primary research would benefit from a similarly precise and transparent approach. Note that we are not suggesting that researchers document and quantify where their *ideas* came from, which could come from personal experience, watching a movie, reading social commentary, studying past research, or any combination of these activities (Glueck & Jauch, 1975; Zechmeister et al., 2001, pp. 22-25). What we are proposing is that it is important to transparently and precisely describe how researchers reviewed past research that is used to justify that pursuing an idea through research is a contribution to knowledge.

We report an analysis of randomly selected articles in psychological science journals to examine what type of quantitative judgments are used to communicate research problems and whether they are supported by any systematic documentation on its origins. We discuss the validity threats that result from using different standards to evaluate reviews in primary versus secondary research and offer solutions based on existing methodological practices. We explain the knowledge gains this approach promises and discuss its potential use and limitations.

Validity

Four types of validity (Cook & Campbell, 1979, Chapter 2; Shadish et al., 2002, Chapters 2-3) are typically discussed to evaluate Steps 3 (study design) and 4 (data analysis, results, and discussion) of the research process (see Figure 1). *Internal* validity refers to the researcher inference that the manipulation of the independent variable is the sole cause of variation in the dependent variable. A key question to evaluate internal validity is whether alternative explanations are ruled out and/or whether the hypothesized mechanism is ruled in. Issues such as successful randomization, operationalization of the independent variable, and whether the dependent variable was measured consistently with its definition are relevant to evaluate the internal validity of primary research (Campbell, 1957). *External* validity refers to the question of whether the sample, design and measures correspond to real-world features of the phenomenon (Campbell, 1957) . *Construct* validity inferences assess whether the relevant constructs are measured and operationalized consistent with their definition (Cronbach & Meehl, 1955), and whether the mechanisms relevant to the construct’s measurement have been identified (Whitely, 1983). *Statistical conclusion* validity describes whether the statistical model is consistent with the variance structure of the data (Cook & Campbell, 1979, pp. 39-44).

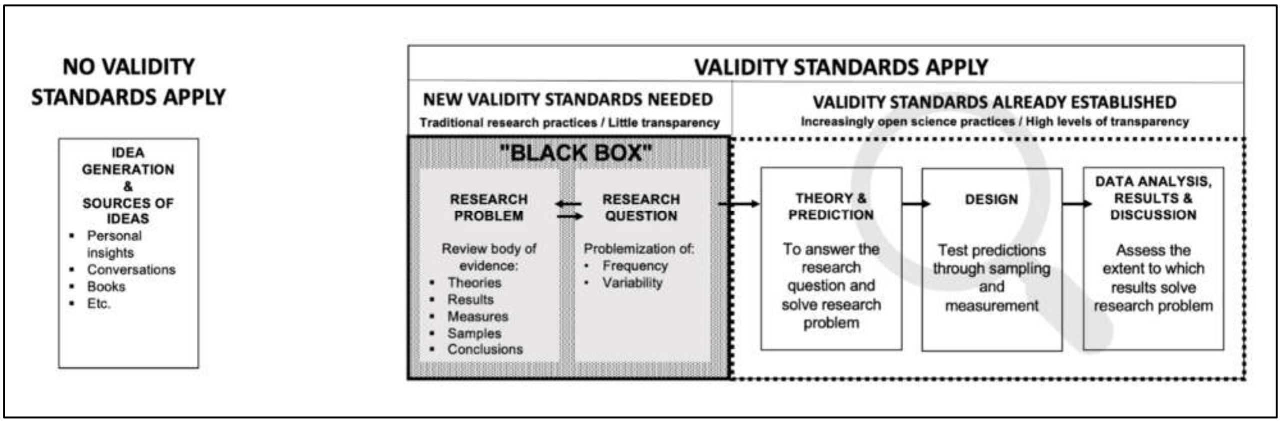
A careful discussion of validity threats, or specific reasons why inferences about causality, constructs, statistics, or generalizations could be more or less correct , along with the knowledge generated by the research process contributes to the emergence of new research problems and advances scientific progress (Chan & Arvey, 2012). Every scientific enterprise seeks to generate cumulative knowledge that is as accurate and as truthful as possible (Bird, 2007; Giner-Sorolla, 2012; Meehl, 1978; Sismondo, 2004). Evaluating the validity of inferences made in the research process supports this goal by assessing the quality of the research and the informational value it generates (Brewer & Crano, 2014).

Despite the many articles and books on ways to define, evaluate, and reduce threats to validity (Campbell, 1957; Cronbach & Meehl, 1955; Shadish et al., 2002), current research practices frequently lack validity – be it low construct validity by the use of non-validated scales (Flake et al., 2017), internal validity by the use of non-validated manipulations (Chester & Lasko, 2021) or external validity by using biased samples, measures or settings (Loyka et al., 2020; Yarkoni, 2020). This has led to calls for current validity standards to be given more attention and improved (Kenny, 2019; Vazire et al., 2022) because they otherwise impede scientific progress (Eronen & Bringmann, 2021).

We do not seek to improve existing validity types here, but instead propose *research problem* validity as an effort to define, scrutinize, and improve the extent of correctness of research problems in primary research. Research problem validity helps scholars to articulate more precise and transparent research problems, which should also enable a more precise contribution to knowledge (Gall et al., 1996). With research problems directing many decisions of the subsequent research process (Kerlinger, 1986), the existing four validity types become meaningful once valid research problems are established. The construction of research problems and establishing their validity can be conceptualized as a superordinate step to development of theory and predictions, just like theory and prediction can be thought of as superordinate steps to research design and statistical analyses (c.f., Fiedler et al., 2021).

Figure 1

The psychological research process: Open science practices have increased the validity and transparency of predictions, design, data analyses, and results.


NO VALIDITY STANDARDS APPLY
IDEA GENERATION & SOURCES OF IDEAS
Personal insights
Conversations
Books
Etc.
VALIDITY STANDARDS APPLY
NEW VALIDITY STANDARDS NEEDED
Traditional research practices / Little transparency
VALIDITY STANDARDS ALREADY ESTABLISHED
Increasingly open science practices / High levels of transparency
"BLACK BOX"
RESEARCH PROBLEM
Review body of evidence:
Theories
Results
Measures
Samples
Conclusions
RESEARCH QUESTION
Problemization of:
Frequency
Variability
THEORY & PREDICTION
To answer the research question and solve research problem
DESIGN
Test predictions through sampling and measurement
DATA ANALYSIS, RESULTS & DISCUSSION
Assess the extent to which results solve research problem

Research Problem Validity Defined

A research problem in primary research consists of a review of past research which reveals two or more factors that bring about a contradiction or undesirable consequence (Clark et al., 1977, p. 6; Kerlinger, 1986, p. 17; Pillutla & Thau, 2013). Central to this process is a review of past research[1]. *Typically*, and as our analysis of a sample of published articles shows, this review involves making quantitative judgments about features of past research with regards to the frequency (e.g., “most,” “few”) and/or variability (e.g., “on the one hand, on the other hand”) of theories, measures, samples, tasks, analyses, results, or conclusions. The reviewed features are then problematized by making apparent one or multiple contradictions with other scientific knowledge and by pointing out the undesirability of not having an answer to these contradictions. The problematized review of past work is a crucial step that directly determines the research question and the professed magnitude of the

[1] Certain literatures in experimental social psychology such as cognitive dissonance theory (Harmon-Jones & Harmon-Jones, 2007), conformity pressures (Cialdini & Goldstein, 2004) or intergroup bias (Hewstone et al., 2002) have now accumulated findings for 40-60 years, contributing to psychology as a cumulative science (but also see Meehl, 1978). However, even in younger research domains there will be structurally similar phenomena that could be reviewed.

potential contribution of the research. It informs the theory, predictions, and all other subsequent aspects of the research process, including design and data analyses (Hernon & Schwartz, 2007).

Consider the following example. Authors reviewing the literature on group size and risk-taking may juxtapose the direction of effects of group size on risk-taking found in past research, stating that existing results are “mixed” (i.e., some studies find positive effects, others negative effects, others none), a judgment about the variability of results in past research. The problem would then be that we do not understand the source of variability. Frequency judgments are statements that characterize past research to rely “too much” on, for example, one specific risk-taking measure and a problem may be that this measure lacks generalizability. Similarly, past research may be characterized as having “mostly” relied on a dominant theoretical paradigm, which could be a problem because alternative theories may account more accurately for results. Perhaps a handful of references are offered to support the quantitative judgments. These problem statements have in common that they present vague summaries about features of past research and the search parameters of the literature review are unknown. To what extent are these summaries truthful? Are past findings on group size and risk taking truly mixed? How mixed are the findings? To answer this, we would need to know what precisely is meant by “mixed”? And how did the researchers arrive at this conclusion? To answer this, we would need to know how the literature was reviewed. In each instance, the answer concerns the validity of the research problem. We define research problem validity as *the extent to which judgments about past research informing the research problem are approximately truthful*.

Like other validity definitions, ours highlights the degree of truthfulness of an inference as central to evaluating validity. More specifically, by truthful we mean the extent to which the judgment is correct given the available evidence in past research. Just like with other validity types (Cook & Campbell, 1979, Chapter 2; Shadish et al., 2002, Chapters 2-3), for a judgment to be evaluated as correct, multiple criteria need to be considered. We argue here that two key considerations are the extent to which the judgment is precise, and how transparent it is how judgment came about.

Precision refers to the degree to which the inference is exact and accurately judges the available research. A judgment such as “most studies” could refer to anything between 50.01% and 99.99% of all studies and is therefore vague (Partee, 1989; Rett, 2018). The judgment conceals the distribution of past research[2]. A numerical statement such as “70% of reviewed studies,” on the other hand is precise and clearly summarizes past research. The imprecision of summary statements in primary research has the potential to bias the interpretation of quantitative judgments. Medical research suggests that consumers of imprecise verbal descriptions tend to make extreme inferences about the meaning of verbal labels. For example, “rare” or “common” side effects are understood as extreme numerical estimates both by physicians and laypeople, although the underlying data are often not extreme (Andreadis et al., 2021). It is possible that readers of scientific communication make similarly extreme conclusions when they read about “most” or “few” results that have shown a specific pattern and choose, for example, not to pursue certain research interests. The practice to communicate summaries about past research in vague terms could also lead to collective misunderstandings. To the research community, it is unclear what type of summary the body of evidence warrants, creating a knowledge gap. It is also unclear what authors mean when they judge the body of evidence as “most” or “few” or “mixed,” and the interpretation of these terms varies substantially. In sum, imprecise quantitative judgments can lead to multiple misunderstandings and potentially cause poor scientific decisions.

[2] Vagueness conceals the distribution of evidence, but also conceals features such as different quality levels of evidence. Precision reveals these features and allows quality differences in evidence to be coded (see Cochrane reviews for examples of coding quality differences in evidence; Higgins & Thomas, 2020).

Transparency refers to how clear the authors communicate how they arrived at their conclusions about past research, or whether the authors were open and explicit about the processes and methods they used to search, code, and characterize past research (Denyer & Tranfield, 2009). Transparency is of overarching importance because it provides both an incentive to be more precise and to be more accurate. When researchers describe how their assessment of "most studies" came about, they are called upon to show the sample of studies that was reviewed, how these studies were coded, and what this implies for certain features of past work. This process enables precision. Researcher could also state that "no other study" has certain features, used a particular paradigm, or found evidence for a given result. This judgment would be precise (because "no" equates to a numeric estimate of zero), but it is not transparent because it is unknown how the judgment came about. The correctness of a judgment on past research can only be evaluated when the judgment is both precise and transparent. Transparency should increase, then, on average, truthfulness because researchers are called upon to document the process they used to arrive at the summary of past research. We note that transparency is not a sufficient condition for truthfulness of research problems, but it is necessary to be able to scrutinize them (cf. Vazire, 2020). The counterfactual of simply providing a short list of references to past work does, in our view, jeopardizes truthfulness.

Current Practices

Although methodologists recognize that an "adequate statement of the research problem is one of the most important parts of research" (Kerlinger, 1986, pp. 16-17), the construct is not debated even in recent comprehensive frameworks for building better research and theory in psychology (Borsboom et al., 2021). Research problems are scrutinized during the review process by editors and reviewers who work under unprecedented pressure (Aczel et al., 2021), which may compromise their ability to review submissions thoroughly (Tsui & Hollenbeck, 2008). Moreover, they may not always have the time or even the specific domain expertise to assess the validity of inferences made about a specific literature or have access to the appropriate information to do so.

This lack of attention is problematic because empirical inferences are made about a large body of data generated by past research. Research problem statements may correctly or incorrectly claim that past research is “one-sided” (e.g., Schaerer et al., 2018, p. 73) or has yielded “mixed” results (Wong & Howard, 2017, p. 216), or that it relies too heavily on a particular experimental paradigm (Schweinsberg et al., 2012). These examples illustrate inferences that make broad judgments on aspects of past research. While such judgments may not be entirely false, it is possible that they are in some instances and that false statements take on a life of their own and perpetuate false beliefs (Carney et al., 2010; Letrud & Hernes, 2019).

Instead of being altogether false, it is more likely that the correctness of research problems varies, because there is little attempt to make them as precise as they could be. Authors support inferences on the body of knowledge by offering a few references to past research, but this lacks precision, particularly when the field of research has received considerable academic attention, as many phenomena central to psychology have (Jones, 1998). If a body of past research yields mixed results question, it would be more accurate to quantify “mixed” results through a meta-analysis of past data studies using one task versus another task that is a useful and fair comparison (Gerlach et al., 2019).

Not all research problems require meta-analyses to increase precision. Other transparent quantification practices to review past work exist, which have the potential to be more informative than imprecise and possibly even biased statements. Take, for example, the judgment that past research relies “too heavily” on a particular experimental paradigm. This inference could be substantiated by counting the usage of the paradigm relative to others in the existing publication record on the phenomenon of interest, along with some coding of other interesting characteristics of this research (e.g. Schaerer et al., 2018). Similarly, the inference that prior literature is "one-sided" could be tested by a count to establish the frequency of the one-sidedness of past research with a description of what those one-sided studies have in common. By systematically engaging with the body of past research, more nuanced conclusions would be forthcoming and more specific research questions would result.

Not only would this practice increase transparency, but it would provide readers who wish to evaluate this part of the research process with structured information that could inform their evaluations. Only 4% of the papers in our sample contained precise and transparent inferences, but even these papers did not provide structured information on how literature was searched, coded, and classified. If this information was provided by authors, then reviewers, editors, and subsequently readers, would have the information relevant to evaluate the problem's accuracy. For example, if authors document they have reviewed literature published between 1980 and 2022, then this may remind a reviewer of relevant papers from the 1970s, and that could change the quantitative judgment in the current formulation of the research problem.

We believe that the counterfactual of not documenting the literature review involved in research problem statements puts too much faith in self-correcting mechanisms. It is possible that eventually readers of published articles conclude that the research problem described in an article is not entirely inaccurate or even false. But self-correcting mechanisms tend to operate slowly (Piller, 2022), do not always work (Vazire & Holcombe, 2022), and papers can impact research literatures, long after their claims have been falsified. For example, Hardwicke et al. (2021) examined how citations of five prominent original studies changed by disconfirming replication evidence: for four out of the five original studies, the percentage of subsequent citations that also cited the disconfirming replication study never exceeded 50%. Only one original study had a somewhat balanced citation pattern, with >88% of subsequent citations also citing the replication study. Similarly, Kelley and Blashfield (2009) present the citation history of an influential paper on sex bias among mental health professionals, considered to be a "citation classic" that has "impacted the thinking of a generation of psychologists and mental health professionals" (Broverman et al., 1970), even though the conclusions have been shown to be wrong repeatedly (Phillips & Gilroy, 1985; Stricker, 1977; Widiger & Settle, 1987). These case studies on self-correction lower our faith in the self-correcting capabilities and speed of the scientific process.

An Analysis of Leading Psychology Journals

What are the current practices describing research problems in primary research in psychological science? We reviewed a random sample of 100 papers published between 1st January 2011 to 31st December 2020 in six leading psychology journals (*Journal of Personality and Social Psychology, Journal of Experimental Social Psychology, Psychological Science, Journal of Experimental Psychology: General, Organizational Behavior and Human Decision Processes, Personality and Social Psychology Bulletin).* Details on our random selection process are provided on OSF: https://osf.io/bq68u/wiki/home/?view_only=de923858992545e2891ed3daf21b3442. One paper in our sample was a meta-analysis, and one paper was a correction. We did not code these two observations because our analysis focusses on primary research.

Coding Process

First, we coded which of these papers made inferences, or summary statements about past research informing the research problem. We also recorded the summary terms used to summarize past research and categorized these terms into frequency (i.e., “most studies in literature X show Y”, or “most studies on topic X use paradigm Y”) and variability terms[3] (i.e., “on the one hand X, on the other hand Y”). We also coded which features of the literature these summary statements described (results; theories; aspects of the phenomenon; study design, methodology, and measures). Second, we coded whether these judgments can be considered *precise and transparent*, which we consider necessary but not sufficient conditions to evaluate the truthfulness, or validity of research problems. We adapted the Oxford Dictionary of English definition of precision as “the quality, condition, or fact of being exact and accurate” and the Merriam Webster definition of transparency as “characterized by visibility or accessibility of information”.

The two first authors and a research assistant reviewed the 100 randomly selected papers and coded all text sections that made “inferences about past research informing the research problem”, following a pre-registered coding scheme (available at https://osf.io/79qdr/?view_only=7c0776b29a674777af6f20704a0c7a7c). Any coding disagreements were resolved through discussions.

We coded text that both made empirical inferences about past research *and* that was also used by the authors to generate a research problem. For example, we coded the following sentence in West et al. (2014, p. 826) because it both includes a frequency-type judgment about past research *and* informs the definition of the research problem: “Although previous studies have documented negative effects of perceived anxiety on cross-group relationships, to date, few studies have explored underlying psychological mechanisms that may account for these effects.” The term “few studies” constitutes an inference about the frequency of mechanism-testing study designs in the literature, *and* this informs the research problem as

[3] Terms that describe the central tendency of past research (i.e., “typical,” “most,” etc.) were coded as part of the frequency category.

the authors "sought to isolate one mechanism" of these relationships (West et al., 2014, p. 839). "Few" is an imprecise term, and how the inference came about is not transparent but remains unclear because no systematic literature review is provided. As an example of a variability judgment that informed the research problem, we coded the following section in Hilbig et al. (2014, p. 529): "However, at closer inspection, the extant findings also reveal a noteworthy degree of variability, such that some individuals actually behave very much in line with self-interested individual utility maximization, whereas others display other-regarding preferences (e.g., Engel, 2011; Fischbacher, Gächter, & Fehr, 2001)." The inference is not precise because the degree of variability is not quantified, and it is also not clear how the inference came about because the authors do not share whether they systematically reviewed the existing literature.

We ignored statements that only reviewed past research, but that did not inform the research problem. For example, we did not code statements that merely summarized certain studies, were used as stylized facts, or that did not directly feed into the research problem. For example, Beckmann et al. (2013, p. 681) verbally summarized the literature on muscle contractions. The review helps the reader understand why the studies in Beckmann et al. (2013, p. 681) are methodologically sound, but we did not code this section because the review did not inform the paper's research problem.

Results From Coding 100 Papers in Leading Psychology Journals

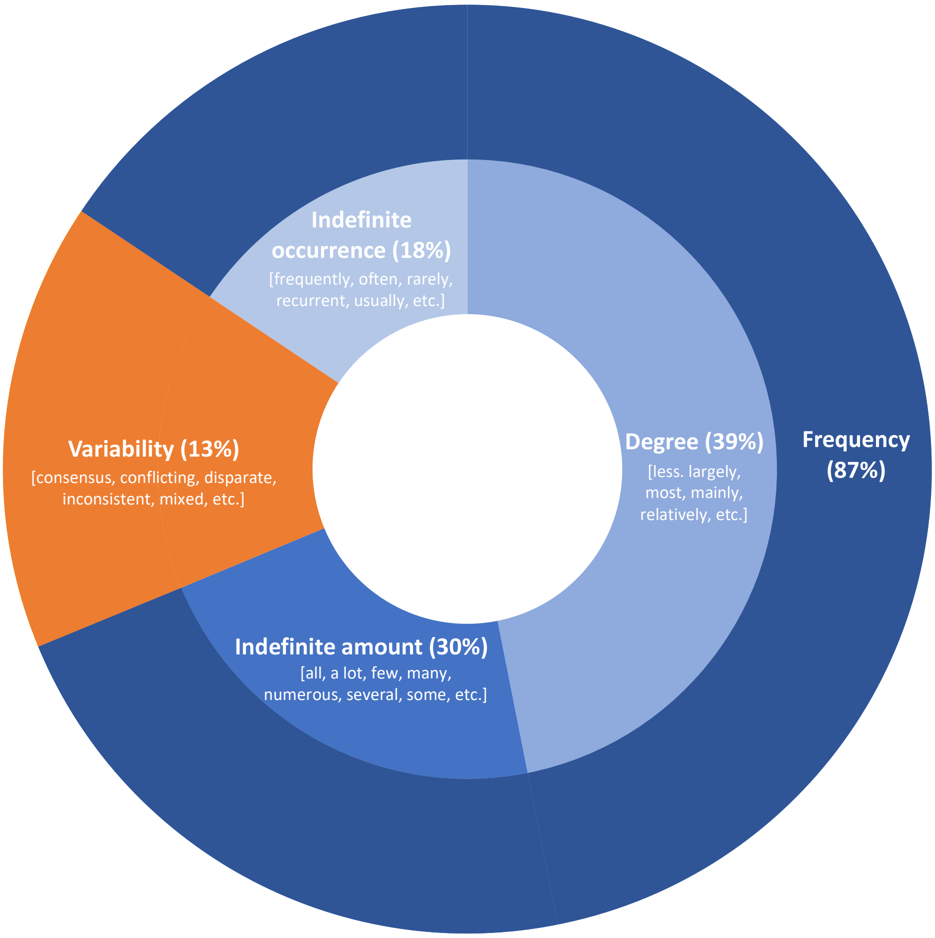
First, our coding shows that 78 out of the 100 randomly selected papers made summary statements about past research informing the research problem. The 100 randomly selected papers contained 133 inferences on past research that informed their research problems and questions.

Frequency judgments are most common (114/133), followed by variability judgments (12/133), and judgments relating to both frequency and variability (4/133). Three additional inferences (3/133) were based on the summary term "unclear" and were not coded on this dimension, because they do not represent a quantitative judgment. We coded 77 different summary terms, 67 of which related to frequency judgments. To better understand the nature of these 67 frequency terms, we categorized them according to their function (Aarts, 2011): 30/67 were coded as *degree terms* that describe the intensity of an action or quality (e.g. largely, primarily), 23/67 were coded as *indefinite amount terms*, which do not specify how many things are being referred to (e.g. few, many), and 14/67 were coded as *indefinite occurrence terms* which describe how often something takes place in indefinite terms (e.g. frequently, often), (see Figure 2). We did not subcategorize the 10 variability terms.

We also coded which parts of the research process these inferences relate to. An inference could be coded as relating to more than one part of the research process. The inferences we coded predominantly relate to aspects of the phenomenon (59/133) and results (47/133), and comparatively less to the study design, methodology and measures (44/133), and to theories (16/133).

Figure 2

Inference type, summary term function and prevalence, and example summary terms in our review of 100 papers. The coding for each individual paper can be seen here: https://airtable.com/shr5NNHGVuQFVRq8j



Second, we assessed to what extent these claims can be considered precise and transparent. Only 15% of the 133 inferences made in the 100 papers we coded were precise statements about past research. For example, Re and Rule (2016, p. 87) claim that "no study" has examined how internal features of the face relate to leadership ability. Another example of a precise inference can be found in Murphy et al. (2015, p. 200) when the authors claim that "only one study" has investigated a particular question. Both "no study" and "one study" are precise summary terms. The remaining 113 inferences were imprecise judgments on the frequency or variability of past research that do not articulate a precise numeric estimate (Bass et al., 1974). For example, one inference in a paper on the nonconscious priming of communication (Pickering et al., 2015, p. 78) suggested that “many studies have found that unconscious goal pursuit produces the same outcomes as conscious goal pursuit (Chartrand & Bargh, 1996; Dijksterhuis & Aarts, 2010; Dijksterhuis, Chartrand, & Aarts, 2007). However, there appear to be many mediators to effects of priming on goal pursuit (e.g., Locke & Latham, 2006).” We coded both the frequentist characterization of “many studies” and “many mediators” as imprecise. Another paper that examines “the chills” as a psychological construct (Maruskin et al., 2012, p. 136) included the inference that “The literature is a jumble of disparate claims and findings, a fact that has not been apparent to date because the literature has never been reviewed thoroughly.” We coded this inference coded as imprecise because the degree of inconsistency in the literature has not been evaluated and it is not clear what the label “jumble of disparate findings” describes.

Only 5% of the 133 inferences made in the 100 papers we coded transparently explained how the inferences about past research came about. For example, van Scheppingen et al. (2019) transparently show that their assessment of the literature is that the effects of personality similarity on attraction and satisfaction are small and inconclusive, and that they base this judgment on the results of two meta analyses they cite. Ashworth et al. (2019) claim that the endowment effect is one of the most robust and well-studied phenomena in the behavioral sciences, and they make it transparent that this claim is based on a Google Scholar search they conducted.

Threats to Research Problem Validity, Possible Solutions, and Knowledge Gains

Threats to Research Problem Validity

When research problems communicate imprecise, non-transparent, quantitative judgments, psychological science holds them to lower standards of truthfulness, transparency, and replicability than other parts of the research process, and this threatens the validity of psychological science (Vazire, 2017, 2018). We found that 85% (113/133) of the quantitative judgments we coded in research problem statements were imprecise. These inferences could be incomplete, inaccurate, and perhaps in some cases altogether incorrect. What explains this lack of precision?

Motivated cognition is likely one explanation for imprecise quantitative judgments. Confirmation bias can lead to a search process or to claims about past research that are one-sided, incomplete, or altogether false (Duyx et al., 2017; Rosenthal, 1979). Motivated cognition can also selectively shape evidence into dichotomies that are not warranted (Garcia-Marques & Ferreira, 2011). Carelessness or lack of processing depth is a second likely explanation for imprecision in quantitative judgments. Even when bias is not at play, unsystematic literature searches may omit entire relevant research fields from the problem statement, creating inefficiency in the scientific progress (Beaman, 1991). Generalized overall impressions of the state of past research may also be inconsistent with a structured and quantitative assessment of that research (Stanley, 2001).

Reporting norms are another strong explanation for the lack of precision. The current norm that literature searches and the corresponding quantitative judgments remain undocumented prevents other scholars from reproducing the search and, by extension, the characterization of past research in research problem statements. For example, 95% of the articles we coded did not transparently describe how they selected the parts of the literature they brought to bear on the research problem. Readers of these articles cannot evaluate whether the quantitative judgments that inform the research problem correspond to the body of knowledge they are based upon.

We do not put all our faith in tightening reporting norms. Research problem validity is also threatened when authors are either sufficiently motivated, or just happen to be careless, causing them to sidestep "actual" precision with incorrect or misleading summary statements that *seem* precise and transparent. Although our suggestions cannot completely prevent this, we believe that the alternative of not providing any attempts to increase transparency and precision in research problem statements is worse for reviewers, editors, and readers. Our proposal (or similar ones that could be developed), can help reviewers, editors, and other readers of a primary research article to scrutinize the documented literature search, their results, and the criteria that were defined for the search and coding of search results. Without this documentation, the reader's judgments on the search are basically criterion-free, beyond the references that are provided, and the expertise knowledge that is applied. Our suggested approach provides readers with the information they need to evaluate the search and the subsequent judgments on search results.

Possible Solutions

Possible solutions to quantify inferences on past research more precisely and transparently could be implemented relatively easily. For example, (Fazio & Sherry, 2020, p. 1150; Hughes et al., 2020, p. 2265; West et al., 2014, p. 825) are articles from our sample that supported claims about the size of evidence by citing outcomes of literature reviews and meta analyses . Wölfer et al. (2017, p. 1567) is another paper from our sample that directly cite prevalence statistics from such reviews ("Previous studies primarily relied on self-reports to assess intergroup contact (81% of the studies included in Pettigrew & Tropp's, 2006, meta-analysis used this approach)”). Zhou and Fishbach (2016) is a paper in our sample, and they present a precise research problem when they argue that unattended, selective attrition can bias studies with online samples (e.g., MTurk). They pursued easily implemented strategies to increase the transparency and precision of their review of past work. For example, they substantiated their claim that “dropouts are *rarely* disclosed in published papers” by examining all papers published within a certain timeframe in a specific journal for search terms that indicate both data collected online and the disclosure of dropout information and found that only 4 out of 289 papers reported this information. This transparency helps the reader make an informed judgment whether they agree with the conclusion (Zhou & Fishbach, 2016, p. 495).

Using Systematic Review Guidelines to Improve Research Problem Validity

Implementing transparency in documenting the literature review that led to quantitative judgments about past research is simple, and reporting standards already exist in secondary research that could be borrowed or used as templates. For example, a recent paper in *Perspectives on Psychological Science* (Antonoplis, 2022) transparently describes the search parameters in a systematic review of socioeconomic status: the data base used, publication time period, search phrases, number of articles found, number of duplicates removed, number of articles screened and assessed for whether they fit the eligibility criteria for inclusion in the summary of past research. The guidelines that Antonoplis (2022) followed are already used in systematic reviews in epidemiology (Page et al., 2021), and can improve the transparency of research problem statements in primary research.

We believe that it is useful to precisely and transparently document literature reviews because the current practice remains closed to scientific scrutiny. Perhaps reviews motivating primary research are systematic, perhaps not, perhaps they follow a particular system, or another; we simply do not know. Just as data-analytic decisions for the same hypothesis test vary widely and cause heterogeneity in conclusions about data (Botvinik-Nezer et al., 2020; Schweinsberg et al., 2021), variation in how literature is searched and coded will yield heterogenous conclusions about the same body of research. By not documenting this research step, we are unable to evaluate the comprehensiveness and quality of a review of the literature and readers are unable to *learn* which review practices are superior to others.

Another solution could be to establish conventions for when to use specific summary terms to describe specific frequency or variability observations in the literature (Bass et al., 1974). These conventions could offer consistent rules for translating numeric estimates into verbal summary statements (e.g., use “few” when referring to quantities below five). Similar terminology conventions are used to communicate risk in national security (Kent, 1964), or probabilities in medicine (Andreadis et al., 2021). Cohen (1988) suggested simple conventions to generate consistency for verbal descriptions of effect sizes, and when they should be described as small, medium, or large. Although global conventions are not perfect (Cohen, 1962), similar conventions could help clarify which numeric estimates are underlying the vague verbal summary terms we identified such as “several,” “few,” “many,” or “some” (Bass et al., 1974; Borges & Sawyers, 1974). Consistent terminology could also prevent instances in which authors strategically use ambiguity (Eisenberg, 1984) to advance their objectives (Frankenhuis et al., 2022; Rohrer, 2021) in research problem construction.

Finally, recent calls to make distinct aspects of the scientific process machine-readable (Lakens & DeBruine, 2021; Spadaro et al., 2022) could enhance transparency, standardize the literature review process, and reduce time and effort expenditure (Brisebois et al., 2017; Sabharwal & Miah, 2022)

Open Science Practices to Improve Research Problem Validity

A systematic literature review we conducted showed that research practices related to research questions and problems (Step 1 in Figure 1) are still confined to traditional research practices. When we searched the Web of Science, we found *n* = 1,781 publications on open science and search terms relevant to Step 2 (“open science” & theory; “open science” & prediction), *n* = 1,720 publications for Step 3 search terms ("open science” & design), *n* = 9,286 publications for Step 4 search terms ("open science” & analysis; “open science” & results; “open science” & discussion), but only *n* = 23 publications for Step 1 search terms (“open science” & “research question”; “open science” & “research problem”). When we read these *n* = 23 publications in detail, not one publication discussed how to establish a research problem or research question using open science practices (for a detailed review of these *n* = 23 publications, see here: https://airtable.com/shrRZWuX6bBQlbmkK).

Open science practices make transparency the default choice (Klein et al., 2018) for the steps in the scientific research process, from sharing materials that inform study design choices (Landy et al., 2020) and raw data (Simonsohn, 2013) to data analyses (Botvinik-Nezer et al., 2020; Schweinsberg et al., 2021). Open science practices could also be implemented when reviewing the literature in primary research: scholars could share the search terms they used, and search results, along with coding criteria, and the results of this coding process on an online repository such as OSF.

Possible Knowledge Gains

We believe that there could be several knowledge gains from adopting a more rigorous and systematic approach in the communication of research problems. First, the practice could lead to more informed debate. Authors, reviewers, and readers may understand verbal summary terms of quantitative judgments differently (Bass et al., 1974), but seemingly agree when they do not. For example, “some” heterogeneity may mean 20% for Scholar A, but 60% for Scholar B. Scholar A (who understands the term to mean 20%) might not see this amount of heterogeneity as large enough to plan a new study that would warrant a contribution. Scholar A may also think that a particular moderator would not produce enough variation to systematically affect the existing variance. However, Scholar B (who understands "some" heterogeneity to mean 60%) does see the heterogeneity judgments as informative for subsequent judgments about the potential contribution of a new study, moderator selection, or for study design choices.

Both scholars might agree with the other's thresholds for what amount of heterogeneity is large enough to affect the decision to pursue a study or a specific design. But they disagree on the actual extent of heterogeneity. Quantifying heterogeneity shows the amount of dispersion in findings. Larger amounts of heterogeneity might justify additional empirical investigations to identify (multiple) moderators and in that way have a greater scope for advancing the context-dependence of knowledge (Tipton et al., 2022). Precise and transparent literature reviews can reveal these sources of disagreement and make them explicit instead of concealing them behind verbal summary statements. We believe that an *informed* debate should improve the scholarly discourse and the quality of beliefs after a discussing the disagreements (Duke, 2020, Chapter 5).

A second knowledge gain is to help readers and authors evaluate the amount of uncertainty in knowledge on a given phenomenon. Whether "the majority" of research refers to a majority in seven studies or in seven hundred studies communicates a different degree of confidence in the judgment itself because small samples contain more unusual information than large samples. Related, precise information clarifies the weight of evidence relative to vague statements. Four studies show positive effects, and three studies show negative effects is a more precise characterization than "several" studies showing positive effects and "few" studies showing negative effects. Precise numeric estimates also help the reader evaluate the impact of minor changes: not considering just one of the four studies documenting positive effects changes the pattern of results from a "majority" of findings documenting positive effects to a pattern where "half" the study show positive effects and "half" the studies show negative effects. Precise estimates of variability help readers assess the nature of an effect and can help authors make their study design choices on a structured and precise knowledge base, and not just on their subjective and potentially biased impression of the literature. For example, authors can benefit from precise estimates of heterogeneity to evaluate whether an effect is moderated by a third factor and less variability may necessitate stronger manipulations. Finally, transparency can help reveal straw man arguments, by calling on authors to replace vague claims such as "critics argue", or "many people believe that" with precise evidence on the nature and origin of these claims.

Limitations

Our proposal to improve research problem validity is not without limitations. First, we acknowledge that neither our nor any other framework of this type can fully eliminate misleading characterizations of past research without running the risk of excessive tightening of the research process (Fiedler, 2018). A radical alternative to what we propose here is to altogether abandon current practices, in which research problem statements justify cumulative contribution. The 199th study on the same phenomenon with similar methods can still be considered useful knowledge because psychological phenomena are highly variable and context-dependent (McGuire, 1973). Perhaps what matters is the correctness of methods and conclusions alone, but then we need to abandon the current practice of justifying primary research based on problem statements that claim to have reviewed the literature. Another limitation is that sufficiently motivated authors could seemingly review research with precision and transparency but do so in misleading ways. However, we believe that transparency and precision will make such mischaracterizations easier to identify, limit the scope for strategic ambiguity (Eisenberg, 1984; Frankenhuis et al., 2022; Rohrer, 2021) in the construction of research problems, and thereby contribute to a more truthful cumulative psychological science. Second, a variation of this risk is an only partial, potentially strategic adoption of our recommendations: for example, simply replacing “few papers” with “four papers” increases the precision of a summary statement, but not its transparency. However, if authors also increase transparency and specify their search parameters, readers and reviewers can better assess which literatures were considered, and which were ignored, compared to an otherwise vague and non-transparent literature review. Third, implementing our (or variations of our) suggestions do take time, although the time required can be minimized in simple ways: authors are not called upon to conduct meta-analyses for primary research papers, but simply sharing transparently how they arrive at their characterization of past research and increasing the precision with which they characterize this past research will bring methodological standards in secondary and primary research closer together.

Conclusion

Why should the review of past literature and how researchers identify their research problems remain a black box? Why develop sophisticated methodologies to evaluate the validity of research designs and data analysis in primary studies, but not for the inferences and judgments on past research that justifies this study in the first place? We proposed here that existing tools such as quantification and a documented, reproducible literature search and coding can increase the truthfulness of judgments that are central to the research problem a primary research article attempts to solve. Ignoring research problem validity means that while the study design may be reproducible, externally valid and truthful, the research problem may not be, resulting in the right answer to the wrong problem (Kimball, 1957). More broadly these loose practices today may undermine the scientific goal of building an accurate and truthful cumulative body of knowledge.

References

Aarts, B. (2011). *Oxford modern english grammar*. Oxford University Press.

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*(1), 14. https://doi.org/10.1186/s41073-021-00118-2

Andreadis, K., Chan, E., Park, M., Benda, N. C., Sharma, M. M., Demetres, M., Delgado, D., Sigworth, E., Chen, Q., Liu, A., Grossman Liu, L., Sharko, M., Zikmund-Fisher, B. J., & Ancker, J. S. (2021). Imprecision and preferences in interpretation of verbal probabilities in health: A systematic review. *Journal of General Internal Medicine*, *36*(12), 3820-3829. https://doi.org/10.1007/s11606-021-07050-7

Antonoplis, S. (2022). Studying socioeconomic status: Conceptual problems and an alternative path forward. *Perspectives on Psychological Science*. https://doi.org/https://doi.org/10.31234/osf.io/29v6b

Ashworth, L., Darke, P. R., McShane, L., & Vu, T. (2019). The rules of exchange: The role of an exchange surplus in producing the endowment effect. *Organizational Behavior and Human Decision Processes*, *152*, 11-24. https://doi.org/https://doi.org/10.1016/j.obhdp.2019.03.012

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology*, *59*(3), 313-320.

Beaman, A. L. (1991). An empirical comparison of meta-analytic and traditional reviews. *Personality and Social Psychology Bulletin*, *17*(3), 252–257.

Beckmann, J., Gröpel, P., & Ehrlenspiel, F. (2013). Preventing motor skill failure through hemisphere-specific priming: Cases from choking under pressure. *Journal of Experimental Psychology: General*, *142*(3), 679-691. https://doi.org/http://dx.doi.org/10.1037/a0029852

Bird, A. (2007). *Nature's metaphysics: Laws and properties*. Oxford University Press.

Borges, M. A., & Sawyers, B. K. (1974). Common verbal quantifiers: Usage And Interpretation. *Journal of Experimental Psychology*, *102*(2), 335-338.

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756-766. https://doi.org/10.1177/1745691620969647

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., . . . Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84-88. https://doi.org/10.1038/s41586-020-2314-9

Brewer, M. B., & Crano, W. D. (2014). Research design and issues of validity. In C. M. Judd & H. T. Reis (Eds.), *Handbook of research methods in social and personality psychology* (2 ed., pp. 11-26). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511996481.005

Brisebois, R., Abran, A., Nadembega, A., & N'techobo, P. (2017). An assisted literature review using machine learning models to identify and build a literature corpus. *International Journal of Engineering Science Invention*, *6*(7), 72-84.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P. S., & Vogel, S. R. (1970). Sex-role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology*, *34*(1), 1-7.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312.

Campbell, J. P., Daft, R. L., & Hulin, C. L. (1982). *What to study: Generating and developing research questions*. Sage Publications.

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*(10), 1363-1368.

Chan, M. E., & Arvey, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science*, *7*(1), 79-92. https://doi.org/10.1177/1745691611429355

Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science*, *16*(2), 377-395. https://doi.org/10.1177/1745691620950684

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591-621.

Clark, D., Guba, E., & Smith, G. (1977). Functions and definitions of functions of a research proposal. *Bloomington: College of Education Indiana University*.

Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of Abnormal Social Psychology*, *65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis for field settings*.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 671-689). Sage Publications.

Duke, A. (2020). *How to decide: Simple tools for making better choices*. Penguin Random House.

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of clinical epidemiology*, *88*, 92-101. https://doi.org/https://doi.org/10.1016/j.jclinepi.2017.06.002

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227-242. https://doi.org/10.1080/03637758409390197

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779-788. https://doi.org/10.1177/1745691620970586

Fazio, L. K., & Sherry, C. L. (2020). The effect of repetition on truth judgments across development. *Psychological Science*, *31*(9), 1150-1160. https://doi.org/10.1177/0956797620939534

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, *13*(4), 433-438. https://doi.org/10.1177/1745691617745651

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*(4), 816-826. https://doi.org/10.1177/1745691620970602

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Frankenhuis, W., Panchanathan, K., & Smaldino, P. E. (2022). Strategic ambiguity in the social sciences. *Social Psychological Bulletin*. https://doi.org/https://doi.org/10.31222/osf.io/kep5b

Gall, M. D., , Borg, W. R., & Gall, J. P. (1996). *Educational research*. Longman.

Garcia-Marques, L., & Ferreira, M. B. (2011). Friends and foes of theory construction in psychological science: Vague dichotomies, unified theories of cognition, and the new experimentalism. *Perspectives on Psychological Science*, *6*(2), 192-201. https://doi.org/10.1177/1745691611400239

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1-44.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*(6), 562-571. https://doi.org/10.1177/1745691612457576

Glueck, W. F., & Jauch, L. R. (1975). Sources of research ideas among productive scholars: Implications for administrators. *The Journal of Higher Education*, *46*(1), 103-114. https://doi.org/10.2307/1980926

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. A. (2021). Citation patterns following a strongly contradictory replication result: Four case studies from psychology. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211040837. https://doi.org/10.1177/25152459211040837

Harmon-Jones, E., & Harmon-Jones, C. (2007). Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, *38*(1), 7-16.

Hernon, P., & Schwartz, C. (2007). What is a problem statement? *Library & Information Science Research*, *29*(3), 307-309. https://doi.org/10.1016/j.lisr.2007.06.001

Higgins, J. P. T., & Thomas, J. (Eds.). (2020). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). Wiley-Blackwell.

Hilbig, B. E., Glöckner, A., & Zettler, I. (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, *107*(3), 529-539. https://doi.org/http://dx.doi.org/10.1037/a0036074

Hughes, S., De Houwer, J., Mattavelli, S., & Hussey, I. (2020). The shared features principle: If two objects share a feature, people assume those objects also share other features. *Journal of Experimental Psychology: General*, *149*(12), 2264-2288. https://doi.org/http://dx.doi.org/10.1037/xge0000777

Jones, E. E. (1998). Major developments in five decades of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 3–57). McGraw-Hill.

Kelley, L. P., & Blashfield, R. K. (2009). An example of psychological science's failure to self-correct. *Review of General Psychology*, *13*(2), 122-129. https://doi.org/10.1037/a0015287

Kenny, D. A. (2019). Enhancing validity in psychological research. *American Psychologist*, *74*(9), 1018-1028.

Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, *8*(4), 49-65.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). Harcourt Brace Jovanovich College Publishers.

Kimball, A. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association*, *52*(278), 133-142.

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1). https://doi.org/10.1525/collabra.158

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science*, *4*(2), 2515245920970949. https://doi.org/10.1177/2515245920970949

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., . . . Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*. https://doi.org/10.1037/bul0000220

Letrud, K., & Hernes, S. (2019). Affirmative citation bias in scientific myth debunking: A three-in-one case study. *PLoS ONE*, *14*(9), e0222213. https://doi.org/10.1371/journal.pone.0222213

Loyka, C. M., Ruscio, J., Edelblum, A. B., Hatch, L., Wetreich, B., & Zabel, A. (2020). Weighing people rather than food: A framework for examining external validity. *Perspectives on Psychological Science*, *15*(2), 483-496. https://doi.org/10.1177/1745691619876279

Maruskin, L. A., Thrash, T. M., & Elliot, A. J. (2012). The chills as a psychological construct: Content universe, factor structure, affective composition, elicitors, trait antecedents, and consequences. *Journal of Personality and Social Psychology*, *103*(1), 135-157. https://doi.org/http://dx.doi.org/10.1037/a0028117

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven Koan. *Journal of Personality and Social Psychology*, *26*(3), 446-456.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(113), 806-834. https://doi.org/https://doi.org/10.1016/j.appsy.2004.02.001

Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., Roter, D. L., & Nguyen, L. (2015). Reliability and validity of nonverbal thin slices in social interactions. *Personality and Social Psychology Bulletin*, *41*(2), 199-213. https://doi.org/10.1177/0146167214559902

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Partee, B. H. (1989). Many quantifiers. ESCOL 89: Proceedings of the Eastern States Conference on Linguistics, Columbus, OH.

Phillips, R. D., & Gilroy, F. D. (1985). Sex-role stereotypes and clinical judgments of mental health: The Brovermans' findings reexamined. *Sex Roles*, *12*(1), 179-193.

Pickering, M. J., McLean, J. F., & Krayeva, M. (2015). Nonconscious priming of communication. *Journal of Experimental Social Psychology*, *58*, 77-81. https://doi.org/https://doi.org/10.1016/j.jesp.2014.12.007

Piller, C. (2022). Blots on a field? *Science*, *377*(6604), 358-363.

Pillutla, M. M., & Thau, S. (2013). Organizational sciences' obsession with "that's interesting!". *Organizational Psychology Review*, *3*(2), 187-194. https://doi.org/10.1177/2041386613479963

Re, D. E., & Rule, N. O. (2016). The big man has a big mouth: Mouth width correlates with perceived leadership ability and actual leadership performance. *Journal of Experimental Social Psychology*, *63*, 86-93. https://doi.org/https://doi.org/10.1016/j.jesp.2015.12.005

Rett, J. (2018). The semantics of many, much, few, and little. *Language and Linguistics Compass*, *12*(1), e12269. https://doi.org/https://doi.org/10.1111/lnc3.12269

Rohrer, J. M. (2021). Who would win, 100 duck-sized strategic ambiguities vs. 1 horse-sized structured abstract? *The 100%CI*. http://www.the100.ci/2021/12/08/who-would-win-100-duck-sized-strategic-ambiguities-vs-1-horse-sized-structured-abstract/

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641.

Sabharwal, R., & Miah, S. J. (2022). An intelligent literature review: Adopting inductive approach to define machine learning applications in the clinical domain. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00605-3

Schaerer, M., du Plessis, C., Yap, A. J., & Thau, S. (2018). Low power individuals in social power research: A quantitative review, theoretical framework, and empirical test. *Organizational Behavior and Human Decision Processes*, *149*, 73-96. https://doi.org/10.1016/j.obhdp.2018.08.004

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., . . . Uhlmann, E. L. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*(July), 228-249. https://doi.org/https://doi.org/10.1016/j.obhdp.2021.02.003

Schweinsberg, M., Ku, G., Wang, C. S., & Pillutla, M. M. (2012). Starting high and ending with nothing: The role of anchors and power in negotiations. *Journal of Experimental Social Psychology*, *48*(1), 226-231. https://doi.org/10.1016/j.jesp.2011.07.005

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*(1), 747-770. https://doi.org/10.1146/annurev-psych-010418-102803

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875-1888.

Sismondo, S. (2004). *An introduction to science and technology studies*. Blackwell Publishing.

Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Ten Teije, A., CoDa Team, & Balliet, D. (2022). The cooperation databank: Machine-readable science accelerates research synthesis. *Perspectives on Psychological Science*.

Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives*, *15*(3), 131-150.

Stricker, G. (1977). Implications of research for psychotherapeutic treatment of women. *American Psychologist*, 14-22.

Swedberg, R. (2020). Exploratory research. In C. Elman, J. Gerring, & J. Mahoney (Eds.), *The production of knowledge: Enhancing progress in social science*. Cambridge University Press.

Tipton, E., Bryan, C., Murray, J., McDaniel, M., Schneider, B., & Yeager, D. (2022). Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity. https://doi.org/10.13140/RG.2.2.34070.01605

Trochim, W. M. K. (2006). *The research methods knowledge base* (3rd ed.). Atomic Dog.

Tsui, A. S., & Hollenbeck, J. R. (2008). Successful authors and effective reviewers: Balancing supply and demand in the organizational sciences. *Organizational Research Methods*, *12*(2), 259-275. https://doi.org/10.1177/1094428108318429

van Scheppingen, M. A., Chopik, W. J., Bleidorn, W., & Denissen, J. J. A. (2019). Longitudinal actor, partner, and similarity effects of personality on well-being. *Journal of Personality and Social Psychology*, *117*(4), e51-e70.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*(1).

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411-417.

Vazire, S. (2020). Do we want to be credible or incredible? *Observer*, *33*(1), 35-37.

Vazire, S., & Holcombe, A. O. (2022). Where are the self-correcting mechanisms in science? *Review of General Psychology*, *26*(2), 212-223.

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162-168. https://doi.org/https://doi.org/10.31234/osf.io/bu4d3

West, T. V., Pearson, A. R., & Stern, C. (2014). Anxiety perseverance in intergroup interaction: When incidental explanations backfire. *Journal of Personality and Social Psychology*, *107*(5), 825-843. https://doi.org/http://dx.doi.org/10.1037/a0037941

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197.

Widiger, T. A., & Settle, S. A. (1987). Broverman et al. revisited: An artifactual sex bias. *Journal of Personality and Social Psychology*, *53*, 463– 469.

Wölfer, R., Jaspers, E., Blaylock, D., Wigoder, C., Hughes, J., & Hewstone, M. (2017). Studying positive and negative direct and extended contact: Complementing self-reports with social network analysis. *Personality and Social Psychology Bulletin*, *43*(11), 1566-1581. https://doi.org/10.1177/0146167217719732

Wong, R. S., & Howard, S. (2017). Blinded by power: Untangling mixed results regarding power and efficiency in negotiation. *Group Decision and Negotiation*, *26*(2), 215-245.

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1-37. https://doi.org/10.1017/s0140525x20001685

Zechmeister, J. S., Zechmeister, E. B., & Shaughnessy, J. J. (2001). *Essentials of research methods in psychology*. McGraw-Hill.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493-504.