The Long Way From α-Error Control to Validity Proper : Problems With a Short-Sighted False-Positive Debate

Klaus Fiedler, Florian Kutzner and Joachim I. Krueger

Perspectives on Psychological Science 2012 7: 661
DOI: 10.1177/1745691612462587

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 661–669

DOI: 10.1177/1745691612462587
http://pps.sagepub.com
SAGE

The Long Way From α-Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate

Klaus Fiedler[1], Florian Kutzner[1], and Joachim I. Krueger[2]
[1]University of Heidelberg, Germany, and [2]Brown University

Abstract

Several influential publications have sensitized the community of behavioral scientists to the dangers of inflated effects and false-positive errors leading to the unwarranted publication of nonreplicable findings. This issue has been related to prominent cases of data fabrication and survey results pointing to bad practices in empirical science. Although we concur with the motives behind these critical arguments, we note that an isolated debate of false positives may itself be misleading and counter-productive. Instead, we argue that, given the current state of affairs in behavioral science, *false negatives* often constitute a more serious problem. Referring to Wason's (1960) seminal work on inductive reasoning, we show that the failure to assertively generate and test alternative hypotheses can lead to dramatic theoretical mistakes, which cannot be corrected by any kind of rigor applied to statistical tests of the focal hypotheses. We conclude that a scientific culture rewarding strong inference (Platt, 1964) is more likely to see progress than a culture preoccupied with tightening its standards for the mere publication of original findings.

Keywords

replicability, false positives, false negatives, strong inference

False Positives as a Source of Bad Science

Recently, a growing number of methodological articles have painted a pessimistic picture of the state of the arts in behavioral science. From inappropriate significance testing (Bakker & Wicherts, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011) to questionable research practices (John, Loewenstein, & Prelec, 2012; LeBel & Peters, 2011; Simmons, Nelson, & Simonsohn, 2011) and even fraud, these articles have revived an interest in methodology. They have struck a chord, as evidenced by the flurry of conferences, Internet communications, and proposals to revise the peer-reviewing system (Simmons et al., 2011).

Although the underlying motives are worthy and certain symptoms of flawed methodology are incontestable, the new wave of critique has a cannibalistic aspect. Though meant to bring about reforms, these critiques may damage the mission and the social impact of behavioral science, which is so important in modern societies. We recognize the capacity of psychological science for self-examination and self-criticism as a core feature of its sustained success; so we are not suggesting to stifle it. Rather, our goal is to reorient the self-critical discourse by addressing deeper epistemological concerns of theory development and evaluation. In our view, statistical malpractice is subordinate to the superordinate criterion of validity.

Virtually all the critical arguments, and suggestions for improvement, that have been extracted from recent articles on "voodoo correlations" (Fiedler, 2011; Vul, Harris, Winkielman, & Pasher, 2009), inappropriate statistical tests (Nieuwenhuis, Forstmann, & Wagenmakers, 2011; Wagenmakers et al., 2011), questionable research practices (John et al., 2012; Simmons et al., 2011), and replication (this volume), are concerned with the problem of false positives or, in statistical jargon, α-error control. An α error is the incorrect rejection of a null hypothesis H_0 or, equivalently, the erroneous inference that a study supports an alternative hypothesis H_1 (see Fig. 1). Such erroneous inferences may simply reflect sampling error (chance) or inadequate data sampling procedures that violate the stochastic-independence of statistical tests (Simmons et al., 2011). As a result, virtually all suggested remedies to invalid scientific inferences have concentrated on α-error control in statistical significance testing, thus considering false positives as the most prevalent and most costly error in current behavioral science (Ioannidis, 2005; Simmons et al., 2011; Wagenmakers et al., 2011).

Corresponding Author:
Klaus Fiedler, University of Heidelberg, Hauptstrasse 47-51, Heidelberg, 69117, Germany
E-mail: kf@psychologie.uni-heidelberg.de

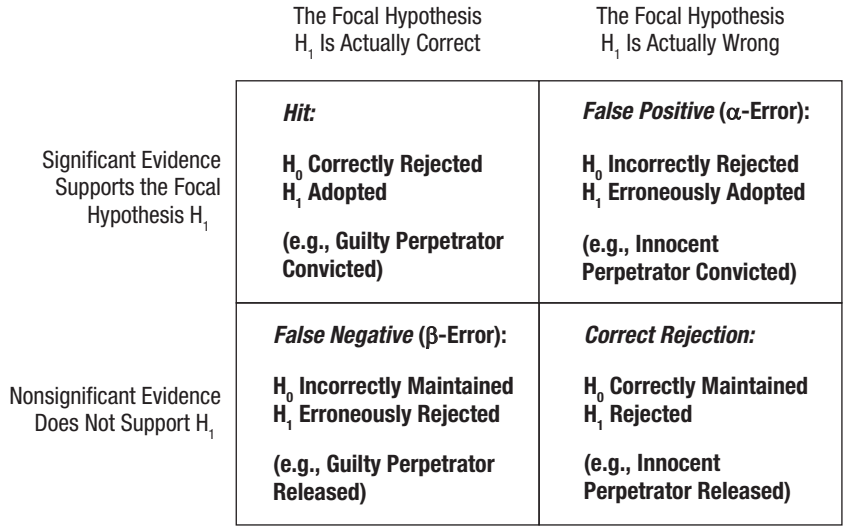

Fig. 1. False-positive errors and false-negative errors represent two out of four possible outcomes of statistical hypothesis testing (exemplified in parentheses with regard to juridical decisions).

The Aim of This Article

Although we fully agree with the goal of reducing false positives and the view that replication is a virtue of empirical science, we hasten to add that isolated discussions and interventions to reduce false positives, without any consideration of the importance of false negatives, may retard rather than support the growth of knowledge in psychological science. We believe that false negatives or β errors – failures to discover and substantiate correct hypotheses (see Fig. 1) – are a more fundamental problem than are false positives. We explain why false negatives are logically antecedent and superordinate to false positives, and we provide examples of how overlooking alternative hypotheses (β errors) renders all scrutiny intended to reduce α errors worthless. False positives can be corrected through replication whereas false negatives are less likely to be detected, corrected, and understood.

Reasons for the Primacy of False Negatives

There are two main reasons for the primacy of false negatives, one occurring in what Reichenbach (1938/1952) called the "context of justification "(i.e., hypothesis testing) and another in what he called the "context of discovery" (i.e., hypothesis selection).

Statistical false negatives: The lesson taught by Jacob Cohen

At the stage of significance testing, the primacy argument was memorably made by Jacob Cohen (1962, 1992). Given the conventionally small samples and effect sizes in the behavioral sciences, the expected rate of failures to obtain a significant result when H_1 is true can be expected to be high. In a review of representative psychological studies, Cohen (1962) found false-negative rates for small and medium effects (corresponding to mean differences of .25 or .50 standard deviations, respectively) ranging from $\beta = .52$ to $\beta = .82$.[1] A generation later, Sedlmeier and Gigerenzer (1989) found that things had not improved. Even the most pessimistic estimates of α errors reached by researchers who exploit all bad practices strategically (Simmons et al., 2011) can hardly be higher.

Another reason for high β-error rates is that most behavioral scientists use standard (omnibus) analyses of variance instead of specific theory-driven models that decompose the systematic variance into theoretically meaningful contrasts (Rosenthal, Rosnow, & Rubin, 2000). Such nonspecific tests of the existence of (any) differences between two of three or more factor levels suffer from low statistical power. Or, if only one cell deviates from the others in a 2 × 2 design (e.g., due to the multiplicative influence of a treatment and a personality trait), the true effect is concealed and distributed over equally weak and nonsignificant main effect and interaction results. Further, the common failure to correct for measurement error and for sampling error raises the rate of unsuccessful tests of valid hypotheses (Schmidt, 2010).

Apart from the sheer prevalence of α and β errors, the consequences of the latter are more likely to be irreversible. A false negative implies, pragmatically, that a hypothesis is discarded and thus no longer considered for testing, whereas a false positive makes it likely that continued research will fix the mistake. In other words, the scientific virtue of self-correction applies much more directly and more strongly to α errors than to β errors. Many errors of the latter type may never be discovered. Such truncated hypothesis testing is likely to produce irrational judgments and decisions (Denrell, 2005; Einhorn & Hogarth, 1978; Denrell & LeMens, 2012). The final outcome is that the bag of valid findings is much smaller than

it would be if the falsely positive and falsely negative results had the same probability of being followed up.

Cohen's (1962, 1992) analysis highlights the deep asymmetry between a high tolerance for β errors coupled with strict policing of α errors. Simmons et al. (2011) sought to justify this asymmetry by invoking a difference in value: "The most costly error is a *false positive*, the incorrect rejection of a null hypothesis … false positives waste resources: They inspire investment in fruitless research programs and can lead to ineffective policy changes" (p. 1359). Simmons et al. did not present a cost–benefit analysis to support this claim, nor did they consider the possibility of costs due to false negatives.

We refrain from stating that inappropriately committed actions are more costly than erroneously omitted interventions and that only false positives motivate decisions and actions. The failure to find empirical support for a valid hypothesis (statistical β error) or the failure to pursue a valid alternative hypothesis (theoretical β error) can have irreversible consequences. In legal decision making, for example, the failure of a linguistic truth test (criteria-based content analysis; Vrij, 2005) to exceed a threshold may prevent the court from recognizing the truth of an aggravating witness report and from convicting a guilty defendant.

Global assumptions about the particular danger of false positives are hard to justify. Notice that error types can be reframed. One may test the hypothesis that prime durations of less than 100 ms increase the strength of priming effects or, conversely, that strong priming effects previously found below 100 ms can also be obtained at durations greater than100 ms. Imposing stronger precautions on one type of error would only lead to strategic reframing of research questions.

Yarkoni (2009), a leading theoretical statistician, recently commented on the limitations of an isolated false-positive debate. Given the typical sample size of $n = 20$ participants in expensive fMRI research and given the low power of the correlation coefficients used to relate brain measures to manifest behaviors or traits, many published studies can be expected to reflect false positives. Wrongly identified crucial brain areas entail α errors. However, the same error logically implies the existence of many β errors. Other brain regions that may be better suited to account for the criterion behavior may not produce statistically significant signals or may simply be overlooked by the investigators. Every α error on a focal hypothesis entails β errors on alternative hypotheses, just as for every falsely convicted person one (or more) true criminals go free.

It makes little sense to pretend that liberal strategies to avoid these false negatives are less useful or effective than conservative strategies to reduce false positives. If anything, false negatives have logical precedence because whenever the correct hypothesis is dismissed, even the strictest tests of the remaining hypotheses can only create an illusion of validity (Einhorn & Hogarth, 1978). Conversely, false positives do not prevent researchers from replicating and correcting erroneously supported hypotheses or searching creatively for alternative hypotheses. The truncation of research on a valid hypothesis is more damaging and less reversible than the replication of research on a wrong hypothesis.

Theoretical false negatives

So far, we have disputed the claim that critical hypothesis testing in the context of justification should only be a matter of refining and tightening measures to reduce false positives. From the perspective of signal-detection theory, setting the criterion for the funding, investigation, and publication of findings at an extremely conservative level is indeed a strategy that reduces false positives, but this reduction comes at the price of an unknown increase in false negatives (Swets, Dawes, & Monahan, 2000). To lower the entry threshold for novel hypotheses and nonmainstream ideas, a more liberal criterion may be a superior strategy. This strategy purchases a reduced risk of false negatives with a greater tolerance for false positives.

Following Reichenbach, we now consider the context of discovery—that uncharted water where science intersects with courage and imagination (see also Huxley, 1963). Creativity in both arts and sciences requires low thresholds for detection, curiosity, and communication (publication) of new ideas. In the context of theory discovery, false positives are not really a problem. Indeed, the fertile production of interesting and nonredundant hypotheses is the *sine qua non* at this loosening stage (Kelly, 1955), even though most of these ideas will turn out to be false. Screening out bad hypotheses then occurs at multiple, staggered levels. Konrad Lorenz famously remarked that he generated and rejected a dozen hypotheses before breakfast to keep sharp.

The lesson taught by Peter Wason. Theoretical false negatives, that is, overlooking valid alternative hypotheses, is the graver threat, as illustrated in Peter Wason's (1960) seminal paper on the common failure to apply Popper's (1959) logic of scientific discovery. Given a sample sequence of numbers "2, 4, 6," participants were to find the underlying rule that generated these numbers. They had to propose other sequences and received feedback on whether these were rule consistent or inconsistent. Wason's rule-induction task is an experimental analogue of scientific discovery. The most serious psychological problem revealed by this task is not the lack of α-error control—that is, the uncertainty about the classification of given hypotheses (suggested sequences) as either correct or incorrect.

Instead, the main problem was a narrow focus on a few restrictive hypotheses and the corresponding failure to consider other, broader hypotheses that suggest more parsimonius explanations. Most participants focused on a single hypothesis (e.g., linearly increasing natural numbers with a slope of 2 per element). Various tests of this hypothesis (e.g., 6, 8, 10 or 237, 239, 241) would all draw feedback confirming that the sequence does conform to the rule. But participants failed to test other hypotheses (e.g., strictly monotonic increase, weakly

monotonic increase, nondecreasing, any natural numbers) that could have simpler, more parsimonious, and more adequate explanations.

To demonstrate the limited role of statistical inference in a probabilistic version of Wason's paradigm, one might apply a sign test or a Bayesian test to decide whether some error-prone feedback sample confirms the focal hypothesis. If the test is highly significant, confidence will increase that the tested hypothesis captures the underlying rule. However, as Wason (1960) noted, the critical problem does not lie in the existence of a few false positives but in the failure to consider alternative hypotheses offering different—and more general—explanations. A strategy that only focuses on strict tests of a focal and local hypothesis will hardly lead to better performance. Instead, it is likely to distract from a more fundamental feature of the inference task, namely, the search for more generally valid hypotheses.

Wason's lesson highlights the fact that the human mind is biased towards engaging in repeated and increasingly restrictive and technically sophisticated tests of focal hypotheses; it is reluctant to test alternative and innovative hypotheses (Fiedler & Walther, 2004; Koriat, Lichtenstein, & Fischhoff, 1980). People in general and scientists in particular prefer specific and conjunctive explanations over general and disjunctive explanations calling for a liberal criterion (Tversky & Kahneman, 1983; Zuckerman, Eghrari, & Lambrecht, 1986). This basic asymmetry favors the dogged pursuit of significance testing in the context of justification over the creative construction of alternative hypotheses in the context of discovery. Scientists, much like experimental participants, find it harder to reason about false negatives than about false positives. In a graduate seminar conducted by the first author, for example, 38 Dutch PhDs had no difficulty providing examples of false positives (e.g., surprising priming effects), but they were almost completely unable to generate compelling examples of false negatives.

A note on model testing. Roberts and Pashler (2000) showed that even highly significant correlations or other indices of model fit cannot prove that a tested model is valid or more valid than other models. Take the story of Clever Hans. Even the most compelling evidence (large n, small α) that the horse named Hans provided correct responses to calculation tasks could not prove that the horse could do math in its head. The genius of Oskar Pfungst (Pfungst, Stumpf, & Rahn, 1911) was to consider β errors. He tested Hans under novel conditions (e.g., varying the testers' body language) that others—among them famed psychologist Carl Stumpf—had not dreamed of as being relevant.

The exaggerated importance attributed to the statistical tests of causal models (while ignoring alternative theoretical models) is particularly evident and troubling in mediation analysis, a method now *de rigueur* in high-impact journal outlets. The pertinent literature is almost exclusively concerned with error terms in significance tests of mediation models (cf. Bullock, Green, & Ha, 2010). For example, to find out if the theoretical construct of fluency (Z) mediates the impact of repeated exposure (X) on attractiveness (Y), a statistical test is based on the rationale that the (partial) correlation between X and Y decreases significantly when Z is controlled (i.e., partialled out). However, independent of its statistical rigor, such a focal test of Z does not rule out the possibility that countless other mediators (Z', Z'', Z''', etc.) might also prove significant and provide more adequate causal models (cf. Fiedler, Schott, & Meiser, 2011). Perfect α control in testing a focal model is of little value if theoretical β errors conceal more adequate models.

The ubiquity of the Wason phenomenon

Blindness to the false-negative problem reaches well beyond these particular examples. This fundamental problem plagues some of the most prominent theories of the day. Although being explicit about highly involving topics may raise an issue of tact, we consider it essential to illustrate our argument with specific theories.

Terror management theory. Myriad studies suggest that exposure to stimuli such as funeral homes, hearses, 9/11, or other symbols of death makes the idea of one's own mortality salient. As a consequence, people shift towards conservative values, conventional cultural categories, and old habits that promise security and familiarity (cf. Greenberg, Solomon, & Pyszczynski, 1997). The success of the theory of terror management in scientific publications, research funding, career opportunities, and in the media rests on a hypothesis, which is intuitively appealing but restrictive in these assumptions: It assumes that there is something unique about the threat caused by mortality-related thoughts or symbols.

If they followed the recommendations of Simmons et al. (2011), researchers would collect at least 20 observations per cell, report all variables and conditions, determine sample size in advance, and report alternative analyses for excluded and included outliers and covariates. In turn, reviewers and editors would ask for the replication of studies based on suboptimal data collection. We doubt that such local tightening of the process would address the key challenges to the validity of terror-management theory. Even hundreds of well-conducted experiments that perfectly conform to the maxims of α control would lead to misleading theoretical conclusions if a few existence proofs of alternative hypotheses, motivated by β control, could demonstrate that the hypothesis being tested is too restrictive. Maybe neither physiological threat nor mortality proper is a necessary condition of the obtained findings.

Analogous to Wason (1960), false negatives may go unnoticed when a variety of more general and less restrictive hypotheses are not tested systematically and frequently enough. Are the stimuli used in terror management experiments unequivocal operationalizations of mortality salience? Could they imply the opposite—namely, survival (Nairne,

Pandeirada, & Thompson, 2008)? Could they represent existential values? Would stimuli referring to birth, religion, or miracles of nature induce the same effects? Even when a literature review may reveal a few findings meant to exclude such alternative hypotheses, this would hardly justify the final conclusion that all these potential false negatives have been ruled out sufficiently.

The uneasiness and the lack of interest in counterfactual reasoning about false negatives are evident in the reluctance of mortality-salience researchers to cite studies inspired by Gollwitzer, Wicklund, and Hilton's (1982) theory of self-completion. Originating in similar sources as terror-management theory, this theory predicts a number of similar results, though under much less restrictive conditions. To induce a conservative shift, a desire for order, and regained self-esteem, it is sufficient to manipulate subtle cues of incompleteness, reminding participants of their status as a learner, or of an incomplete task (Wicklund & Braun, 1987). Disentangling this cluster of alternative and supplementary factors would appear to represent a more comprehensive research goal than all the scrutiny applied to the control of false positives in testing a single privileged hypothesis.

Survival encoding and memory. On the basis of a recently published series of widely recognized memory experiments, Nairne and colleagues (e.g., Nairne et al., 2008) concluded that memory for a list of words improves considerably when the encoding task calls for judgments of the survival value of the stimuli. As this finding is of utmost interest to evolutionary approaches to cognitive psychology, it is important to secure its replicability and to rule out the possibility that it merely reflects an α error.

But what about theoretical β errors, or failures to consider alternative accounts for the seeming impact of survival encoding—that is, errors that are analogous to the failure to test alternative hypotheses in Wason's (1960) famous study? Indeed, a number of alternative interpretations suggest themselves (Klein, 2012; Kroneisen & Erdfelder, 2011). Because the specific manipulation of survival reference is complex and polysemous, it is inevitably confounded with a host of alternative explanations. Rather than reflecting survival reference as a necessary condition, the findings may be due to a memory advantage of self-referent encoding (Klein, 2012; Kuiper & Rogers, 1979), assuming that survival concerns are self-referent. An even broader alternative explanation would attribute the findings to the relevance or the degree of ego-involvement triggered during encoding.

As soon as any more general hypothesis is found to account for the enhanced memory performance, the survival-encoding theory is cast into doubt and replaced by a less restrictive theory. Within the method of multiple hypotheses (Chamberlin, 1944), every theoretical β error concerning one broader hypothesis entails a theoretical α error for a too narrow hypothesis and vice versa. Note also that a positive test result for a logically superior (more inclusive) hypothesis provides less equivocal evidence than does a negative (nonsignificant) test result for the focal hypothesis. Too conservative a criterion for new research findings can therefore decrease the chances of the most powerful remedies to wrong and premature theorizing: namely, strong inference (Platt, 1964) based on falsification and counter-evidence.

Confirmation bias. Since Rosenthal's (1964) famous demonstrations of experimenter effects and self-fulfilling prophecies, cognitive, social, and applied psychologists have been concerned with the bias to confirm rather than disconfirm one's hypothesis (Nickerson, 1998). Again, popular explanations of the confirmation bias assume that, in addition to sampling more information on a focal hypothesis (called *positive testing*), further motives must be at work, such as stereotypes (Snyder, 1992), wishful thinking (Munro & Stansbury, 2009), or selective search for confirmatory information (Jonas, Schulz-Hardt, Frey, & Thelen, 2001). However, reminiscent of Wason (1960), confirmation bias may be a much more general phenomenon, for which none of these conditions is necessary. For teachers to confirm that boys are better in science than girls, they need not believe this hypothesis to be true, nor do they have to focus on this hypothesis, nor do they have to like boys more than girls or to engage in selective search for confirmatory information (Fiedler & Walther, 2004; Moore & Small, 2007). Given completely unbiased information search, it is sufficient that teachers are exposed to a larger sample of boys' responses in science (Fiedler, Walther, Freytag, & Plessner, 2002; Fiedler, Walther, & Nickel, 1999).

Strict control of false-positives errors in significance tests of confirmation-bias studies has value only locally and will hardly contribute to the development of a comprehensive theory within which too restrictive accounts of confirmation biases can be understood as special cases of a more general principle. Sustained theoretical progress requires unorthodox researchers whose horizon is broad enough to be concerned with false negatives and counter-intuitive hypotheses (Roberts & Pashler, 2000). Ironically, entry criteria that are too strict may harm unconventional theories more than leading theories, for the latter have already established paradigmatic wisdom about the best stimuli, task settings, instructions, and context conditions that warrant large effect sizes and significant results (Fiedler, 2011).

More than a file-drawer problem

These examples demonstrate that failures of scientific discovery cannot be reduced to a file-drawer problem in publications or nonpublication of those hypotheses that become the focus of research. Rather, the internal and external validity (Campbell, 1957) of scientific findings critically depend on the failure to test a multitude of alternative hypotheses for which there is "no file in the drawer." The misleading consequences of these theoretical β errors are more fundamental than all the statistical α or β errors that pertain to specific (and often

overly narrow) hypotheses. What is perfect α control worth if it pertains to a hypothesis that turns out to be premature, false, or unnecessary? It would be a cardinal mistake to assume that statistical significance testing is the essential core of methodology and that it is the master rule on which both internal and external validity depend (Campbell, 1957; Krueger, 2001). A hierarchy of other aspects have logical precedence. Statistical tests depend on research designs, which in turn depend on the sampling and operationalization of variables. Sampling itself is a function of the underlying theory, which in turn is comprehensible only in the context of an implicit metatheory that explains why other theories can be excluded.

Why should the most subordinate levels in this hierarchy—technical issues of measurement and statistical testing—be most crucial for scientific progress? The point here is not to suggest that technical issues of psychometrics, scale transformation, or statistics should be ignored in behavioral science. Indeed, in highly developed paradigms, such as visual search, classical conditioning, or psychophysics, technical sophistication (such as scaling) lie at the heart of new insights and strong theory tests. Yet, one needs to be aware of the relative and conditional nature of all precision at the data level. Conclusions about statistical significance, effect size, or quantitative model fit cannot be valid if only one superordinate assumption is revised: for example, if some variable is transformed monotonically or operationalized differently, if a different control condition is used or a different rival hypothesis is included in a Bayesian test, or if one of many problematic assumptions that underlie every statistical test (e.g., stochastic independence) is given up. Moreover, an idle belief in a short-sighted theoretical account can become even more detrimental if it is based on a hypothesis established by the most rigorous statistical standards.

Strict α but lenient publication threshold? It is interesting to note that Simmons et al. (2011) proposed that "reviewers should be more tolerant of imperfections in results" [p. 1362]. The goodwill behind this suggestion notwithstanding, it can be anticipated that stricter α control alone will do more to strengthen conservative, risk-averse research than to encourage the publication of heuristically promising, innovative work. It remains a statistical fact that measures that decrease α will often increase β. Moreover, we suspect that stricter entry criteria will not reduce the likelihood that researchers exploit all degrees of freedom in research designs and selection of auspicious stimuli and task conditions. More likely, it will provide incentives to make effect-inflating voodoo strategies (Fiedler, 2011) more subtle, more ethically acceptable, and more difficult to detect. For example, researchers may run more pilot studies to select the very stimuli, task parameters, or instructions that happen to produce the statistically strongest results. The question is again whether stricter rules of the kind suggested by Simmons et al. (2011) will place a handicap on those false-positive defectors who engage in bad practices or maybe on those naive and honest false-negative players whose presentational skills are less developed.

Elimination of false positives as a remedy of bad practices? Recently, several articles have gained publicity because they have linked the methodological issue of replicability and quality of science to the serious ethical issue of data fabrication and fraud (John et al., 2012). The critical assumption underlying this unfortunate linkage—which can cause great harm to the image of behavioral sciences—is apparently that many false positives reflect researchers' bad practices and their deliberate strategies to deceive others and themselves.

However, the rationale underlying this pessimistic interpretation is not at all clear. Why should stricter α control imposed on significance testing prevent bad practices applied to the fabrication or correction of data prior to all statistical testing? Why should a Machiavellian researcher who fabricates data have a false-positive problem due to small *n*, outliers to be removed, or unwanted dependent variables to be ignored? The problem would arise only for incompetent data fabricators. In the current academic discourse, the question arises of whether stricter α control decreases the likelihood of bad practices. We suggest that it is equally possible that stricter α control increases bad practices. In the absence of pertinent evidence, we believe that the conspicuous way in which researchers struggle with significance and effect size is reflective of ethical scruples rather than common skills in faking data.

Conclusions

For all these reasons, we conclude that an isolated debate of false positives, with its emphasis on statistical hypothesis testing, is not fruitful for the long-term growth of science. A campaign to reduce false positives proceeds at the expense of increasing false negatives, which inhibit scientific growth in more fundamental ways. Instead, we advocate an open conversation about creative methodologies that can tackle fundamental issues of hypothesis generation and selection.

Summary

We have argued that overcoming the neglect of false negatives is crucial for several reasons. First, false negatives are superordinate to false positives because the failure to detect or substantiate an alternative hypothesis can render efforts to maximize the reliability of a focal hypothesis test moot. Second, false negatives are less likely to be corrected than published false positives, as only the latter suggest themselves for replication and critical assessment (cf. Denrell, 2005). Third, theoretical innovations arise when scientists overcome false negatives but they hardly ever arise from abandoning false positives (cf. Fleck, 1935/1979; Kuhn, 1962). Fourth, false negatives present a cognitively more demanding problem than false positives. Fifth, research strategies that aim at

overcoming false negatives can yield existence proofs and new discoveries, whereas strategies aiming to reduce false positives only yield ambiguous nonproofs (based on insufficient evidence). Last but not least, both statistical false negatives and theoretical false negatives are common, due to small, underpowered samples (Cohen, 1992), the failure to correct for sampling error and measurement error (Schmidt, 2010), and researchers' inattentional blindness to alternative hypotheses (Wason, 1960).

Publication and funding policy

Because most measures or strategies to decrease false positives can be expected to increase false negatives and vice versa, any informed debate about science policy must take the trade-off between both types of error into account. The view we have taken in this article is that false negatives and the context of discovery are superordinate to false positives and the context of justification and hypothesis testing. Notwithstanding the worthwhile aims of the call for more control of false positives, science would hardly prosper if unrealistically high thresholds inhibited the publication and dissemination of innovative ideas, discouraged (young) scientists from conducting bold research, protected the established mainstream from dissenters, and forced researchers to concentrate on the reliability of local research questions rather than engaging in open-minded validity tests of global research questions.

How is this quest for low-threshold, exploratory research compatible with the goal of quality control and validity as the ultimate criterion of science? What other tools can be used to separate the wheat from the chaff in empirical research, in addition to statistics, transparency of study conditions, and common access to data repositories?

Strong inference

We believe that a convincing answer to this final question was already articulated half a century ago in Platt's (1964) insightful article on "strong inference." The growth of science depends not so much on technical procedures of significance testing, but on clearly articulated theories and upfront debates leading to crucial tests of alternative hypotheses. Real progress can only be attained when clearly spelled-out theories enable and force researchers to predict what empirical results a theory excludes and what evidence might falsify a given theory or, preferably, allow for clear-cut decisions between two or more competing theories. Good science means devising contrastive theoretical predictions leading to enlightening studies of the *experimentum-crucis* type. Any theory that remains vague about what it excludes is of little value, according to this basic maxim.

Within such a transparent, explicit, and theoretically sound research environment, empirical results are unlikely to be arbitrary or haphazard, and it is unlikely that questionable data-handling procedures can turn worthless data into compelling results. Rather, theoretical constraints are strong enough that the burden is no longer on statistical procedures, which cannot answer theoretical questions anyway. Occasional α errors will be easily corrected in a culture that welcomes critical replications and encourages open debates (Spellman, 2012).

In psychological terms, we suggest that instead of taking a prevention-focus perspective (Higgins, 1997), imposing progressively prohibitive constraints on the dissemination of imperfect results, and increasing the surveillance of mistakes and bad habits, a promotion-focus perspective would provide a positive incentive to do sound and theory-driven research that survives the scrutiny of strong inference. If there is sufficient agreement among editors, reviewers, and funding agencies concerning strong inference as a criterion of good science, as opposed to prevention of bad science, false-positive errors would reduce to transitory mistakes. Because theoretically or practically important findings will (almost) always be replicated, false positives will soon be detected and discarded before they can cause irreversible mistakes in science and costs in society.

If the ultimate standard of excellence on which leading scientists' careers are built is not the publication of sexy findings of dubious replicability and unknown validity in journals with the highest citation rates, but innovative research based on precisely stated theories and empirical laws, then false positives will be nothing but embarrassing mistakes to be easily corrected in replication studies. We suspect that establishing a hall of fame for the best examples of good science will be a better strategy in the long run than propagating a methodology that minimizes false positives in significance testing while losing sight of the validity of the hypotheses being tested.

Acknowledgment

Jan De Houwer, Olivier Klein, Tobias Krüger, Reinhard Pekrun, and Momme von Sydow provided us with very thoughtful comments on a draft of this article.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

The research and scientific work underlying this article was supported by a Koselleck Grant of the Deutsche Forschungsgemeinschaft awarded to Klaus Fiedler (Fi 294/ 23-1).

Note

1. In a 2 × 2 design, for example, assuming an effect size of .25, decreasing sample size from 160 to 40 increases the β-error rate from 20% to 65%.

References

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678. doi:10.3758/s13428-011-0089-5

Bullock, J., Green, D., & Ha, S. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*, 550–558.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312. doi:10.1037/h0040950

Chamberlin, T. C. (1944). The method of multiple working hypotheses. *Scientific Monthly*, *59*, 357–362.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. doi:10.1037/h0045186

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

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*, 951–978.

Denrell, J., & Le Mens, G. (2012). Social judgments from adaptive samples. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 151–169). New York, NY: Psychology Press.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*, 395–416. doi:10.1037/0033-295X.85.5.395

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171. doi:10.1177/1745691611400237

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*, 1231–1236. doi:10.1016/j.jesp.2011.05.007

Fiedler, K., & Walther, E. (2004). *Stereotyping as inductive hypothesis testing*. New York, NY: Psychology Press.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom—A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes*, *88*(1), 527–561.

Fiedler, K., Walther, E., & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology*, *77*, 5–18. doi:10.1037/0022-3514.77.1.5

Fleck, L. (1979). Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv [Genesis and development of a scientific fact. Introduction to the theory of thought style and thought collective] (F. Bradley & T. J. Trenn, Trans.). Chicago, IL: Chicago University Press. (Original work published 1935)

Gollwitzer, P. M., Wicklund, R. A., & Hilton, J. L. (1982). Admission of failure and symbolic self-completion: Extending Lewinian theory. *Journal of Personality and Social Psychology*, *43*, 358–371. doi:10.1037/0022-3514.43.2.358

Greenberg, J., Solomon, S., & Pyszczynski, T. (1997). Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 61–139). San Diego, CA: Academic Press.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*, 1280–1300.

Huxley, A. (1963). *Literature and science*. New York, NY: Harper & Row.

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

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532.

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, *80*, 557–571. doi:10.1037/0022-3514.80.4.557

Kelly, G. A. (1955). *The psychology of personal constructs. Vol. 1: A theory of personality. Vol. 2: Clinical diagnosis and psychotherapy*. Oxford, England: W.W. Norton.

Klein, S. B. (2012). A role for self-referential processing in tasks requiring participants to imagine survival on the savannah. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1234–1242. doi:10.1037/a0027636

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118. doi:10.1037/0278-7393.6.2.107

Kroneisen, M., & Erdfelder, E. (2011). On the plasticity of the survival processing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1553–1562. doi:10.1037/a0024493

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

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Kuiper, N. A., & Rogers, T. B. (1979). Encoding of personal information: Self–other differences. *Journal of Personality and Social Psychology*, *37*, 499–514. doi:10.1037/0022-3514.37.4.499

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review Of General Psychology, 15*, 371–379. doi:10.1037/a0025172

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology*, *92*, 972–989.

Munro, G. D., & Stansbury, J. A. (2009). The dark side of self-affirmation: Confirmation bias and illusory correlation in response to threatening information. *Personality and Social Psychology Bulletin*, *35*, 1143–1153. doi:10.1177/0146167209337163

Nairne, J. S., Pandeirada, J. S., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, *19*, 176–180. doi:10.1111/j.1467-9280.2008.02064.x

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. doi:10.1037/1089-2680.2.2.175

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. (2011). Erroneous analyses of interactions in neuroscience: A problem of

significance. *Nature Neuroscience*, *14*, 1105–1107. doi:10.1038/nn.2886

Pfungst, O. O., Stumpf, C. C., & Rahn, C. L. (1911). *Clever Hans*. Oxford, England: Holt, Rinehart and Winston.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic books.

Reichenbach, H. (1952). *Experience and prediction*. Chicago, IL: University of Chicago Press (Original work published 1938).

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rosenthal, R. (1964). The effect of the experimenter on the results of psychological research. *Bulletin of the Maritime Psychological Association*, *13*(1), 1–39.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York, NY: Cambridge University Press.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242. doi:10.1177/1745691610369339

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. doi:10.1037/0033-2909.105.2.309

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

Snyder, M. (1992). Motivational foundations of behavioral confirmation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 67–114). San Diego, CA: Academic Press. doi:10.1016/S0065-2601(08)60282-8

Spellman, B. (2012). Data, data, everywhere . . . especially in my file drawer. *Perspectives on Psychological Science*, *7*, 58–59. doi:10.1177/1745691611432124

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decision. *Psychological Science in the Public Interest*, *1*, 1–26.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315. doi:10.1037/0033-295X.90.4.293

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law*, *11*(1), 3–41. doi:10.1037/1076-8971.11.1.3

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. doi:10.1037/a0022790

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140. doi:10.1080/17470216008416717

Wicklund, R. A., & Braun, O. L. (1987). Incompetence and the concern with human categories. *Journal of Personality and Social Psychology*, *53*, 373–382. doi:10.1037/0022-3514.53.2.373

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 294–298.

Zuckerman, M., Eghrari, H., & Lambrecht, M. R. (1986). Attributions as inferences and explanations: Conjunction effects. *Journal of Personality and Social Psychology*, *51*, 1144–1153. doi:10.1037/0022-3514.51.6.1144