AMERICAN PSYCHOLOGICAL ASSOCIATION

THEORETICAL & PHILOSOPHICAL PSYCHOLOGY

Journal of Theoretical and Philosophical Psychology

ISSN: 1068-8471

2021, Vol. 41, No. 4, 233–247
https://doi.org/10.1037/teo0000179

Are Psychologists Appraising Research Properly? Some Popperian Notes Regarding Replication Failures in Psychology

William O'Donohue

Department of Psychology, University of Nevada

Describes the implications of Popper's philosophy of science for appraising research and thus the replication crisis in psychology. A core problem may be a fundamental misunderstanding of an epistemology of scientific methods rather than more narrow, technical issues like statistical analyses. Popper argued against Platonic criteria of knowledge such as justification and truth and instead made the case for severe testing of theories of high empirical content that are designed to solve scientific problems. Popper also argued for an evolutionary epistemology in which surviving a (hopefully severe) falsification attempt is not seen as producing truth but as verisimilitude in which the theory is seen as having survived one among a very large number of possible tests. Puzzles regarding replication failures can then be seen as mistaken epistemic appraisals of research that do not properly take into account five interrelated dimensions: (a) the researcher's motivations (craving to be right vs. finding error as efficiently as possible), (b) neglecting a proper appraisal of the empirical content of the hypothesis under test, (c) neglecting an appraisal of the severity of the test, (d) neglecting to consider the ratio of tests survived versus all possible tests, and (e) neglecting an appraisal of the problem-solving status of the hypothesis.

Public Significance Statement

This article argues that psychological research may be based on a theory of knowledge or philosophy of science that is both a problem and responsible for some psychological research not to replicate. In addition, it proposes that a Popperian approach which is based on testing beliefs in an attempt to falsify these may be a remedy for this problem.

Keywords: popper, replication, falsifiability, research methods, philosophy of science

Scholars have recently claimed that there is a replication crisis in psychology (e.g., Laws, 2016). These concerns are not unique to psychology as Ioannidis (2005) argued in his now-classic article, "Why Most Published Research Findings Are False" published in a leading medical journal. The argument regarding psychology, particularly social psychology, states that there is a large and growing number of studies that have attempted to reproduce findings of several key psychological phenomena, but these have often failed to reproduce the pattern of data found in the original studies.

Although it is beyond the scope of this article to give an exhaustive listing of these replication failures, some illustrative findings that have failed to replicate might be useful to briefly review. One of the best-known instances is probably Bem (2011) precognition finding ("feeling the future") that subsequently failed to replicate (Aldhous, 2011). Schimmack (2020) has stated that Bem's study originated the replication crisis in social psychology. Shedding light on how a study could

This article was published Online First March 22, 2021.

William O'Donohue https://orcid.org/0000-0001-6679-6310

Correspondence concerning this article should be addressed to William O'Donohue, Department of Psychology, University of Nevada, Reno, NV, United States. Email: wto@unr.edu

produce such unusual and surprising effects, Simmons et al. (2011) explicitly used what has become to be known as questionable research practices (QRPs) and produced a finding of a "chronological rejuvenation effect" whereby subjects became 1.5 years younger by listening to the Beatles "When I am 64" instead of a control song. QRPs vary from continuing to run subjects until statistical significance is found, to reordering the importance of hypotheses to favor positive results, to failing to report outcomes that do not support the researcher's favored beliefs.

However, there is also a growing list of studies of more "standard" hypotheses that have also failed to replicate. For example, Darley and Gross (1983) originally reported that social class information biased subjects' interpersonal judgments. However, Baron et al. (1995) in two experiments utilizing samples that were much larger not only failed to replicate these findings but interestingly their findings were in the opposite direction. Importantly, Jussim et al. (2016) reported that since 1996, the original (Darley & Gross, 1983) study was cited 852 times, while the failed replications have been cited only 38 times (according to Google Scholar searches conducted on 9/11/15). This pattern of more attention being paid to the original but nonreplicated finding does not appear to be unusual. Jussim et al. also found other examples of failed replications not being as widely cited as the original studies. For example, Bargh et al. (1996) showed that unconsciously priming subjects with an "elderly stereotype" (unscrambling jumbled sentences that contained words like old, lonely, bingo, and wrinkle) subsequently made subjects walk more slowly. However, Doyen et al. (2012) failed to replicate this finding using more accurate measures of walking speed. During the period roughly of 2013–2015, the original Bargh et al. (1996) study has been cited 900 times while the Doyen et al. (2012) only 192. As a final example, a meta-analysis of 88 studies by Jost et al. (2003) found that political conservativism is a trait characterized by rigidity, dogmatism, prejudice, and fear, was not replicated by a larger better-controlled meta-analysis conducted by Van Hiel et al. (2010). However, again, during the period 2010–2015 the original study was cited 1,030 times while the latter by only 60. Jussim et al. (2016) have suggested that "This pattern of ignoring correctives likely leads social psychology to overstate the extent to which evidence supports the original study's conclusions . . . it behooves researchers to grapple with the full literature, not just the studies conducive to their preferred arguments" (p. 1364).

The interpretation of these replications failures has varied, from denying there is a replication problem, to suggest some wrongdoing on part of the original study authors, to wrongdoing on the part of those conducting the replication, to concerns that there has been a reliance on QRPs, to casting doubt on the notion of the scientific method in psychology (e.g., see Schimmack, 2020). The latter view was expressed by Francis (2012) who stated, "The scientific method is supposed to be able to reveal truths about the world, and the reliability of empirical findings is supposed to be the final arbiter of science; but this method does not seem to work in experimental psychology as it is currently practiced" (p. 3). However, probably the most common interpretation is a problematic reliance on QRPs in the original study. There are some data to support this interpretation. John et al. (2012) surveyed over 2,000 psychologists and found that psychologists admitted to a rather extensive use of QRPs such as failing to report all of the dependent measures for which they had collected data (78%), collecting additional data after checking to see whether preliminary results were statistically significant (72%), selectively reporting studies that supported their favored hypotheses (67%), claiming to have predicted what was actually an unexpected finding (54%), and failing to report all of the experimental conditions that they ran (42%).

On a more positive note there are some indications that there have been some recent changes in scientific practice that seem to be improvements. For example, Makel et al. (2012) found that the published replications rate after the year 2000 was 1.84 times higher than for the years 1950–1999. In addition, there have been several large-scale direct replication projects including: The Many Labs project (Klein et al., 2014); the Reproducibility Project of the Open Science Collaboration (2015); and the Pipeline Project (Schweinsberg et al., 2016). The Many Labs project involved 36 research groups across 12 countries that have replicated 13 psychological studies with over 6,000 participants with (77%) being successfully replicated. However, the Pipeline Project replication rate was much lower (60%)

while the Open Science Collaboration (36%) was lower still.

This article will provide a Popperian (1959) analysis of the problem of replication failures. One reason to examine Popper is that successful scientists in other disciplines, as well as psychology (e.g., Meehl, 1978), have attributed his philosophy of science as having a profound influence upon them and the success of their scientific efforts. For example, Mulkay and Gilbert (1981) called Popper a "philosopher of action" and have stated,

> 'I think Popper is incomparably the greatest philosopher of science that has ever been', writes Sir Peter Medawar winner of the Nobel Prize for medicine, and himself a experienced analyst of scientific thought and scientific practice. Medawar's judgement has been echoed by other eminent scientists. Sir Herman Bondi states that: 'There is no more to science than its method and there is no more to its method than what Popper has said'. Similarly John Eccles, another Nobel Prize winner, testifies to the impact of Popper's writings on his approach to research: 'my scientific life owes so much to my conversion in 1945 . . . to Popper's teachings on the conduct of investigations . . . I have endeavored to follow Popper in the formulation and in the investigation of fundamental problems in neurobiology' (p. 389).

Thus, Popper is somewhat unique in that it seems that more frequently scientists have attributed their success to his influence as compared to other philosophers of science. Second, Popper was one of the first of the modern philosophers of science to identify and combat the problem of confirmation bias in human judgment (although admittedly Popper was certainly heavily influenced by Bacon). To the extent that confirmation bias is still present in replication failures—perhaps by the use of QRPs in the initial study, then Popper's method of focusing on theories of high empirical content that are then severely tested seems a particularly useful antidote. Thus, Popper's analysis of the "craving to be right" might be most productive to understanding what is wrong with the use of QRPs (these allow confirmation bias to enter in psychological research and produce a result that is not robust, particularly if relatively hidden QRPs are not used in the replication attempt). This is not to say that Popper doesn't have critics—and we will deal with some of the most important below. However, it is important to note that analyses from the perspective of other philosophers of science of the problem of replication failures also might prove instructive.

Such an analysis suggests that problems associated with replication failures may not just be at the level of statistics or conventional conceptions of research design but follows Meehl (1978) when he stated: "the problem is epistemology, not statistics" (p. 393). Others have also found what they judge as an underappreciation of Popper in the design and appraisal of psychological research and have suggested remediation along Popperian lines. For example, Holtz and Monnerjahn (2017) found in an examination of popular social psychology introductory textbooks that Popper's philosophy of science has had "little to no traceable influence on the epistemology and practice of social psychology" (p. 348). They argued for increased attention to Popper's falsificationism. In another article, Holtz (2020) argued along Popperian lines that "the reasons for the present 'crisis' in psychology can be attributed in part to epistemological deficiencies—psychologists are too eager to find their theories corroborated by empirical evidence, they do not consider competing theories often enough, and they often do not pay enough attention to the inferences that can be drawn from not finding the expected results" (p. 1).

O'Donohue and Willis (2018) found similar results in an examination of a wider variety of introductory psychology textbooks. These authors found little agreement on how these textbooks defined science or the scientific method, as well as little mention of the key controversies found in the philosophy of science such as inductive versus deductive logic of research, or justificationism versus falsificationism. It ought not to be a loss that Popper's anti-inductive account runs counter to many popular conceptions of science. The inductivist uses empirical methods to collect (many) specific observations in an attempt derive (through an ampliative logical inference rule) a true and more general theory from these particular data; the Popperian falsificationist, on the other hand, aims at identifying an error in a theory by discovering faulty predictions as means of replacing more error-filled theories with better—less error full—theories. Thus, a growth of knowledge—if it is possible at all—is always relative; all the scientist can do is to attempt to put his or her theories to the test and to develop better theories. Thus it is fair to say that Popper is a staunch anti-inductivist, but also a fallibilist—if a theory passes an attempt at falsification attempt this is not an inductive confirmation of

the theory—but rather a survival of an attempt to falsify it—and the degree of corroboration (Popper's term) is a function of the severity of the test. Finally, Fidler et al. (2018) stated:

> We argue that although strict adherence to Popper's HD [hypothetico-deductive] model would not solve all the problems the 'replication crisis' has thrown at the scientific community, scientists could do worse than follow his advice about bold conjectures and risky tests in establishing the absence (or presence) of effects (p. 237).

It is certainly the case as with all philosophers of science that Popper's account has been criticized by a number of individuals in a number of ways (e.g., see Lakatos, 1980 below), particularly on the basis that he provides a normative account rather than a descriptive account (e.g., see Latour & Woolgar, 1986). However descriptive accounts themselves have come under criticism and it is both difficult to gain a representative sample of the work of scientists, and of particular importance a sample of successful problem-solving in science—particularly in psychology, and to agree on the elements responsible for that success. This article also suggests that Popper's philosophy of science might be useful for both understanding and remediating the replication crisis but goes beyond the previous work in that new implications are brought forth regarding several additional relevant Popperian dimensions as well as developing the implications of risky testing and the empirical content of a theory in more detail. In addition, the most general implication is that current research design in psychology is not being properly appraised and Popperian criteria for appraising research design are delineated and discussed.

Popper and Replications

Popper (1959) in his classic Logic of Scientific Discovery recognized the problem of replication failures:

> Every experimental physicist knows those surprising and inexplicable apparent 'effects' which in his laboratory can perhaps even be reproduced for some time, but which finally disappear without trace. Of course, no physicist would say in such a case that he had made a scientific discovery (though he might try to rearrange his experiments so as to make the effect reproducible). Indeed the scientifically significant physical effect may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed. No serious physicist would offer for publication, as a scientific discovery, any such 'occult effect,' as I propose to call it—one for whose reproduction he could give no instructions (pp. 23–24).

In addition, Popper (1959) was quite transparent when he declared "... a few stray basic statements contradicting a theory will hardly induce us to reject it as falsified. We shall take it as falsified only if we discover a reproducible effect which refutes the theory. In other words, we only accept the falsification if a low level empirical hypothesis which describes such an effect is proposed and corroborated" (p. 203; italics added). Although these statements are subject to multiple interpretations, Popper seems to recognize that scientists may themselves find an experimental effect that is not easily reproduced and should be cautious about considering this as a scientific finding; and in addition in pursuing falsifications of some claim—any finding proposed to falsify should itself be sufficiently robust to be reproducible.

Popper also recognized the general problem of confirmation bias in human cognition and was also one of the first to suggest that disciplines that some considered as scientific such as Marxist, Adlerian, and Freudian theories were actually pseudoscientific because all experience, come what may, was seen by their proponents as "confirmation" of the belief system. Popper (1962) stated:

> The most characteristic element in this situation seemed to me the incessant stream of confirmations, of observations which 'verified' the theories in question; and this point was constantly emphasize by their adherents. A Marxist could not open a newspaper without finding on every page confirming evidence for his interpretation of history; not only in the news, but also in its presentation—which revealed the class bias of the paper—and especially of course what the paper did not say. The Freudian analysts emphasized that their theories were constantly verified by their 'clinical observations.' As for Adler, I was much impressed by a personal experience. Once, in 1919, I reported to him a case which to me did not seem particularly Adlerian, but which he found no difficulty in analyzing in terms of his theory of inferiority feelings, Although he had not even seen the child. Slightly shocked, I asked him how he could be so sure. 'Because of my thousandfold experience,' he replied, whereupon I could not help saying: 'And with this new case, I suppose, your experience has become thousand-and-one-fold' (p. 35).

For the purposes of this article, the implications of Popper's philosophy of science will be categorized into four general domains: (a) proper and improper scientific motivation, (b) a rejection of

Platonic epistemology and a replacement with evolutionary epistemology, (c) a focus on appraising a claim being tested by both the empirical content of the claim as well as the severity of the test, and (d) the problem-solving function of science.

For Popper the Proper Scientific Motivation Is Not the Craving to Be Right but Rather the Desire to Efficiently Identify and Root Out Error

For Popper the "craving to be right" *will undermine the scientific process; it must be replaced by the craving to efficiently identify and eliminate error*. For Popper science is supposed to be an antidote for the human tendencies such as the "craving to be right" or what might now be better known as confirmation bias (Nickerson, 1998). Popper stated, "With the idol of certainty (including that of degrees of imperfect certainty or probability) there falls one of the defences of obscurantism which bar the way of scientific advance . . . . The wrong view of science betrays itself in the craving to be right; for it is not his possession of knowledge, of irrefutable truth, that makes the man of science, but his persistent and recklessly critical quest for truth" (Popper, 1959, p. 281).

However, the replication crisis may reveal that Popper was right in his psychology of improper science—that the "craving to be right" can still motivate researchers doing what they and others may (mistakenly) take to be proper science (such as Popper found with Adler, Freud, and Marx) but actually so warp the quality of these efforts so that the research done is actually unscientific. Admittedly, Popper criticized psychologism—in which an understanding of science is explicated in motivational versus logical terms—however, he still argued that psychological tendencies can lead the would-be knower astray. Thus, part of the reason for a study not to be replicated is that it was conducted not as an attempt to efficiently find an error in the favored beliefs but as an attempt to demonstrate the correctness of the researcher's favored beliefs. Thus, beliefs favored by a scientist that are actually false can "survive" the scientist's problematic test. The investigator, being (improperly) motivated to demonstrate that he or she is right might rely on QPRs to produce data that are consistent with their favored beliefs. However, Popper would argue that in these cases the "test" conducted was not actually a test at all, i.e., the belief under test had no chance of being seen as false by the scientist and thus no chance to be reported as false to subsequent research consumers. Subsequently, this result then fails to replicate when another scientist who does not have the same allegiance to the belief conducts testing with a different motivation (and perhaps relies on fewer QRPs).

Moreover, there is some experimental evidence to support the notion that humans are not particularly good at understanding how to conduct falsificatory tests. For example, in Wason (1966) selection task, participants are given a conditional rule in the form "If P then Q" as well as a set of cards. Each of four cards depicts an exemplar that may be consistent or inconsistent with the provided rule: For example, one side of the card depicts whether the instance has the property P and the other side depicts whether it has the property Q. Participants can only see one side of each card, and the side facing the participants' displays either P, non-P, Q, or non-Q. The participants' task, then, is to decide what card or cards need to be turned over in order to determine whether the given rule is true or false. Participants usually have no difficulty deciding to turn over the card that displays the P property in order to see whether there is indeed a Q on the other side. However, importantly for a Popperian analysis, participants seldom (i.e., generally less than 10%, Klauer et al., 2007) choose the non-Q property card in order to determine whether the false consequent is not paired with the true antecedent (i.e., P). Instead, individuals tend to choose the card displaying the Q property (i.e., the instance that allows them to "confirm" but not falsify the rule), or else they choose the P property card alone (Wason, 1966).

There is also some evidence that becoming a scientist does not automatically correct these deficits. Mahoney and Kimper (1976) found that the vast majority of scientists drawn from a national sample showed a strong preference for confirmatory experiments. Importantly, over half of the scientists did not even recognize the valid logical inference rule used in falsification (i.e., modus tollens; If P then Q; Not Q; Therefore Not P) as a valid reasoning form. In another study, Mahoney and DeMonbreun (1977) examined the reasoning skills of 30 scientists compared to those of 15 "relatively uneducated" Protestant

ministers. Where there were performance differences, these tended to favor the ministers. Confirmatory bias was prevalent in both groups, but the ministers used disconfirmatory logic almost twice as often as the scientists did.

Mahoney (1977) also examined how journal reviewers rated studies that were identical in every way except whether the results confirmed or disconfirmed the original hypothesis. He found that studies that reported results consistent with the original hypothesis were rated more highly than the same study (i.e., identical methodology) that reported results that disconfirmed the original hypothesis. Mahoney concluded: "Despite this clear mandate from logic, however, our research programs and publications policies continue in their dogmatically confirmatory tradition. They offer ample testimony to Bacon's (1621/1960) astute observation that "the human intellect . . . is more moved and excited by affirmatives than by negatives" (p. 173).

Popper's Rejected Conventional Platonic Epistemology and Instead Advocated Evolutionary Epistemology

Although the underlying epistemology of conventional research methods used by psychologists is inchoate, Popper's logic of research rejects the Platonic notion that knowledge is justified, true, belief.

i. A positive research result does not produce truth. Part of the problem of understanding replication failures may be due to an underappreciation of the implications of Popper's rejection of the Platonic notion that knowledge is justified by true belief. This is important as the Platonic conception leads to views that are implicated in the replication crisis. For example, psychologists seem to interpret data analyses with certain "positive" outcomes (e.g., a rejection of the null hypothesis) as functioning to epistemically justify a hypothesis and thus this hypothesis can be then be considered true. Popper rejected this view. Popper (1959, p. 10) stated, "I never assume that by force of 'verified' conclusions, theories can be established as 'true,' or even as merely 'probable'." In addition, this is true even after passing multiple tests. Popper stated, "in an infinite universe . . . the probability of any (non-tautological) universal law will be zero" (Popper, 1959, p. 375; emphasis in original).

Popper advanced instead of an evolutionary epistemology. For Popper, the method of trial and error learning is "fundamentally the same whether it is practiced by lower or higher animals, by chimpanzees or by men of science" (Popper, 1972, p. 216). Popper (1999) thus considered science a naturalized "biological phenomenon" (p. 5); a means by which the human species adapts itself to the environment and its challenges. Scientific knowledge only differs from other knowledge in the methods by which errors are systematically criticized and rectified. Accordingly, for Popper the "difference between the amoeba and Einstein" is that "the amoeba dislikes to err while (Einstein) consciously searches for his errors in the hope of learning by their discovery and elimination" (Popper, 1972, p. 70).

From the perspective of evolutionary epistemology, the growth of scientific knowledge is thus characterized by "the repeated overthrow of scientific theories and their replacements by better and more satisfactory ones" (Popper, 1962, p. 215). The function of science is "not to save the lives of untenable systems but, on the contrary, to select the one which is by comparison the fittest, by exposing them all to the fiercest struggle for survival" (p. 42). Importantly, the "fittest" theories help the human species adapt to its current environment. Conversely, insofar as unsuccessful species become extinct, so too do untenable scientific theories (e.g., Aristotelian-Ptolemaic formulations of nature and the universe).

Thus another possibility suggested by Popper is that part of the replication problem is the standard interpretation that the original research study justifies claiming that the hypothesis under test is true and thus it becomes all the more surprising that some claim interpreted to be both justified and true—turns out not to be so when it subsequently fails to replicate. However, Popper argued that a test result does not establish the truth of the claim or hypothesis but at best its verisimilitude—its truth likeness (Popper, 1959). For Popper, a test never establishes the truth of the claim but at best it supports a much more modest conclusion that thus far the claim functions as if it is true. Perhaps this is a somewhat subtle distinction but it is a lower expectation, and this more modest expectation would result in less surprise when a subsequent result is inconsistent with the original result.

ii. Again, inconsistent with the Platonic conception of knowledge, for Popper the initial

positive test does not mean that the claim is justified (or warranted, or proven, or probable) but rather corroborated. Thus, Popper warns that passing one or even several tests does not make the claim under test "justified." For Popper there is no justification but rather something much weaker, that he called "corroboration," that simply means that the claim has survived an attempt (or attempts) at falsification (in Popper's later evolutionary epistemology—it has survived an attempt to kill it). Popper (1959) stated, "So long as a theory withstands detailed and severe tests . . . we may say that it has 'proved its mettle' or that it is 'corroborated' by past experience" (p. 33). Thus, a corroborated result—one that has survived testing thus far—sets lower expectations for continued survival than the Platonic conception of a claim being justified by a past test or even tests.

iii. Relatedly, Popper claimed that the degree of corroboration of any result—even one that has survived many tests is near zero because it is always the case that the ratio of all conducted tests to all possible tests is near zero. Thus, another non-Popperian, Platonic position is to focus on the prior tests as "supporting" or "justifying" the hypothesis tested rather than the Popperian position on focusing on the ratio of tests survived versus all possible tests. Thus, the view that some hypothesis has been supported by x studies may seem impressive the larger x is, but for Popper x is always very small in relation to all possible tests of the claim. The eventual falsification of Newton's theory by experiments derived from Einstein's theory is a case that illustrates that even though a theory has survived many tests it should not be regarded as the final truth. The Chambless Report (Chambless et al., 1996) suggests two positive randomly controlled trials for a psychotherapy to be considered empirically supported, this number is fairly trivial for Popper as he considering 2 to be the numerator and some very large number of possible tests to be in the denominator. This is part of the well-known underdetermination thesis (Quine, 1990): That experimental data always underdetermine the truth of the claim. Thus, in the replication crisis the original result that has survived one test (or even a handful)—but does not survive a subsequent test—is not unexpected in the Popperian calculus of testing. Its surprise only comes when one ignores the very large number of all possible tests it has not been subjected to.

Psychologists Neglect Both the Popperian Dimension of the Empirical Content of a Hypothesis or Theory as Well as the Severity of the Test

According to Popper, theories of low empirical content are less important in science but rarely are the theory or hypothesis under test in psychology appraised for their empirical content. The empirical content of a hypothesis or theory can be assessed prior to any testing of the hypothesis or theory. An informal understanding can be gained regarding the empirical content of a claim by considering that empirical content reflects how much information the claim conveys about the empirical world. Analytic claims such as "All bachelors are unmarried" provides no information about the world, but instead just presents information about the language. The proposition is just a linguistic shortcut, that is, the predicate unpacks what already is in the subject ("All unmarried male adults are unmarried")—one need not actually examine the world to see whether this claim is true or false. On the other hand the empirical claim, "On average, men are taller than women." makes a claim about the world, that is, it rules out two empirical states of affairs, that is, that men and women have an equal average height and that women have an average height that is greater than men's. However, notice that the claim "Men are on average 2.5 in taller than women." contains much more information (i.e., has greater empirical content) as it not only rules out the two cases just mentioned, but this more precise claim also rules out any mean difference other than 2.5 in.—and because the mean can possibly take on many values—it rules out quite a lot. In general, the broader the subject of a sentence and the more precise the predicate, the greater the empirical content of the claim, and thus the more ways the claim is informative and the greater its ability to potentially be falsified (by observing a state of affairs it rules out, e.g., the average height difference is 2.7 in.).

Popper (1959) stated, "The empirical content of a statement increases with its degree of falsifiability: The more a statement forbids, the more it says about the world of experience" (p. 103). According to Popper, the impressive aspect of Einstein's theory is that it is inconsistent with certain possible results of observation. It made

risky predictions: It states that certain states of affairs cannot happen.

Two criteria can be used to evaluate the empirical content of a theory: (a) its level of universality (e.g., "All men . . . ." is more universal than "North American men . . . ") and (b) its degree of precision (2.5 in. is more precise than the qualitative "somewhat taller"). Thus the dimension of universality specifies how many empirical states of affairs the claim can be applied to—with more of these states of affairs representing greater empirical content. The second dimension of precision refers to the specificity of the prediction, that is, how many subclasses of realizations it allows—in this case, it is falsified if any value rather than 2.5 in. is found. This is important to the experimenter as it is the experimenter's task to see if one of the states of affairs that the theory rules out, actually obtains and as such the experimenter needs to understand the potential set of falsifying observations prior to designing the experiment. Thus, a theory with greater empirical content tells more about the world, and as such also has a larger set of potential falsifiers. Therefore, in a study it could be useful for authors to clearly explicate the set of potential falsifiers of the theory or hypothesis under test. This more readily allows an appraisal of how well the original study sought to determine if any of these actually obtain. For Popper, fewer failures to replicate would be found if the original study would be a more efficient search for one of these states of affairs. As Magee (1974) asserted:

> It is not truisms which science unveils. Rather, it is part of the greatness and the beauty of science that we can learn, through our own critical investigations that the world is utterly different from what we ever imagined—until our imagination was fired by the refutations of our earlier theories (p. 37).

However, Pettigrew (1991) has argued that to date hypotheses or theories of high empirical content are not generally put forth by psychologists. Pettigrew (1991) argued that the theories of social psychology are rather "timid" and not "bold" in the sense that they offer "rich and falsifiable content" (p. 23); instead, according to Pettigrew the hypotheses and theories are "patched up" (p. 23) to withstand refutation attempts and criticism (see discussion of the Duhem–Quine thesis below). He urged social psychologists to "go beyond these formulations and pursue Popper's contention that we will learn far more from having our bold theories falsified than we will from failing to falsify narrow conceptions" (p. 23). However, it seems that Pettigrew is correct in that rarely, if ever, in psychology is a theory evaluated by the magnitude of its empirical content; rarely if ever are two competing theories judged by this dimension.

Fidler et al., (2018, p. 239) have also argued that the empirical content of hypotheses in null hypothesis testing is particularly problematic: "Nil null hypotheses are not bold conjectures: In psychology, nil null hypotheses constitute the majority of hypotheses that are tested using NHST [null hypothesis significance testing] (Bakker et al., 2012). The statistical null model is almost certainly wrong to some degree, and NHST will almost certainly provide 'statistically significant' results given a large enough sample size (Meehl, 1967). In a quantitative science, null hypotheses would take a range of values, not simply zero." The reader is referred to Trafimow (2019) who has argued that there are so many assumptions in the statistical model and provides a useful taxonomy of these. Trafimow argued that even if one accepted that the null/nil hypothesis had a reasonable probability of being true, or even if the null hypothesis specified a range of points rather than a single point, the statistical model nevertheless still cannot be shown to be true given these myriad assumptions.

For Popper the most important property of the appraisal of the verisimilitude of a claim is provided through an appraisal of the severity of the test—if the test was less severe it is less likely to root out error—and hence it ought to be less surprising to have a less reliable result. It is important to understand and appraise a test's severity—although Popper failed to specify how to accomplish this in any detail. O'Donohue (2013) suggested the general notion of the severity of a test can be captured in the following example. Suppose one wants to test the claim, "My religious leader does not swear." A less severe test would be sampling the religious leader's sermons, parochial school talks, and conversations at church socials; while a more severe test would be sampling utterances when he or she hits a thumb with a hammer, when a car cuts off him or her in traffic, or when he or she is having a temper tantrum. In this case, it would be much more efficient to examine the latter situations than the former if one is seeking to falsify the proposition that the religious leader never swears—one

might find out in a day what the sample of sermons and school talks would never reveal.

Thus, a severe test is one that tests the least plausible claims of a theory. In a severe test, the good scientist deliberately stacks the deck against the hypothesis. Scientists use background information to choose a test that he or she reasons would more likely to yield a refutation. Basically the researcher needs to ask, "If I want to find out most efficiently if this hypothesis is false, what is the set of potential falsifiers and what study is most likely to efficiently find one of these, if the claim is indeed false?"

Mayo (2018) has done some important technical work on Popper's notion of the severity of a test and has advanced a severity principle:

> *Severity Principle (SP):* Data x (produced by process G) provides a good indication or evidence for hypothesis H if and only if x results from a test procedure T which, taken as a whole, constitutes H having passed a severe test—that is, a procedure which would have, at least with very high probability, uncovered the falsity of, or discrepancies from H, and yet no such error is detected (p. 25).

Mayo also has argued that the reasoning behind a severe test corresponds to a variation of an "argument from error": That is, "There is evidence that an error is absent when a procedure of inquiry with a high probability of detecting the error's presence nevertheless regularly yields results in accord with no error" (p. 25). Although Mayo (2018) attempts to use this severity principle in the context of null hypothesis testing, the use of it here is just to add some depth and detail to Popper's concept of the severity of a test.

Thus, in understanding replication failures there needs to be an appraisal of the severity of the test that the claim has initially survived. A study that is not a severe test is by definition more prone to fail to detect the error of the claim and thus the erroneous belief may be more likely to fail to replicate in a subsequent test that is more severe (as in an indirect replication). Although not explicitly considered by Popper all QRPs (Chambers et al., 2015) would function to decrease the severity of the test as these either reduce the set of potential falsifiers (e.g., by changing the hypothesis after data are collected so that the hypothesis no longer rules out the pattern of data actually found) or by ignoring actual falsificatory data (e.g., selectively reporting only positive results). Perhaps it would be useful to include a section in journal articles covering the following three aspects of the logic of the research: (a) explicitly evaluating the empirical content of the hypothesis under test, (b) explicitly describing the set of empirical states of affairs that would falsify the hypothesis, and (c) arguing how the research design is an efficient method of detecting whether or not these falsifying empirical states of affairs actually obtain.

There also have been important arguments that null hypothesis testing is not risky testing. Fidler et al. (2018, p. 239) have stated, "NHT [null hypothesis testing] is typically practiced is neither bold nor risky. Substantive hypotheses are not subjected to risky tests: When nil null hypotheses are tested, underdeveloped substantive hypotheses are represented only by the alternative hypothesis (i.e., 'not zero'). A failure to find statistically significant results in typical NHST practice does not allow one to reject the alternative hypothesis (Greenland, 2012), which means that even the underspecified substantive hypothesis is not exposed to any real risk." Lakens (2017) has also argued that equivalence tests may be superior to NHST in providing more stringent testing based on the specific claims of the theory.

However, in falsificatory testing scientists must properly address the Duhem–Quine (Quine, 1990) thesis. That is, there is an important and, unfortunately, the inevitable role of auxiliary hypotheses in research, and improperly handing these auxiliary hypotheses can direct the logical arrows of modus tollens away from the hypothesis under test and thus eliminate the riskiness of the test and allow false hypotheses to stand (and fail to replicate in subsequent research). The Duhem–Quine thesis states that in a research study it is actually not possible to isolate and test a single hypothesis because it is always a (large) conjunction of propositions that is under test. The French physicist Pierre Duhem and the analytic philosopher Willard Quine &, Ullian, 1978 both pointed out that it is not the theory alone that is involved in entailing the observation statement but rather the theory and a host of additional hypotheses. Thus, the Duhem–Quine thesis suggests that the actual logic of research is not:

1. If Theory then Observation
2. Not Observation
3. Therefore, Not Theory

But rather:

1. If Theory and AuxHyp1 and AuxHyp2 and ....AuxHypN, then Observation
2. Not Observation
3. Therefore, Not Theory or Not AuxHyp1 or Not AuxHyp2 or .... or Not Aux HypN

For example, if the theory under test is that "Men are taller than women." auxiliary hypotheses would include claims like, "The measuring tape being used is a faithful to the standard measure of inches and feet."; "The research assistants are accurately using the tape and accurately recording values."; "We are accurately categorizing males and females."; "The sample size is sufficiently representative."; "Our computations were accurate."; and so on. Thus, the Duhem–Quine thesis demonstrates that researchers who wish to save their favored belief from falsification are logically free to point the arrows of modus tollens away from the hypothesis under test and toward one or more of the auxiliary hypotheses involved in the research. This has been regarded as an ad hoc strategy unless special conditions to be met. This has led Quine and Ullian (1978 , p. 43) to state that "Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system." Moreover, Quine &, Ullian, 1978 asserted that:

> The totality of our so-called knowledge or beliefs, from most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field. Truth values have to be redistributed over some of our statements. Reevaluation of some statements entails reevaluation of others, because of their logical interconnections—the logical laws being in turn simply certain further statements of the system, certain further elements of the field. Having reevaluated one statement we must reevaluate some others, which may be statements logically connected with the first or may be the statements of logical connections themselves. But the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience (pp. 42–43).

This deflecting the logical force of modus tollens away from the hypothesis under test also can be done when the proponent of the original unreplicated study argues that the replication failure was due to some inevitable differences between the two studies, for example, to differences in the sample of subjects across the two studies.

Many philosophers of science following Popper have struggled to identify proper ways of handling the Duhem–Quine problem. One of the most popular is that of Lakatos (1980) who suggested that the quality of a research endeavor needs to be evaluated by examining how scientists handle the Duhem–Quine thesis. Lakatos argued that a series of research studies needs to be examined in order to capture the treatment of auxiliary hypotheses. He called this series a scientific research program. Lakatos argued that scientific research programs contain what he called "a hard-core" that are fundamental beliefs that a scientist actually never attempts to falsify. For example, Lakatos would argue that for behavior analysts the phenomenon of positive reinforcement is never actually under test in any research project. If in a study in behavior analysis, some stimulus thought to be a reinforcer is presented contingently upon the emission of some behavior but that behavior fails to increase in rate, this prediction failure would be blamed on some auxiliary hypothesis such as, "The contingent stimulus is a reinforcer for that behavior." The failure of the contingent stimulus to increase the rate of the behavior would never be seen as falsifying the law of effect—because for Lakatos the law of effect is in the hard-core of the behavior analytic research program.

Lakatos argued that a research program constitutes good science if it is progressive, and bad science if it is degenerating. According to Lakatos, a research program is progressive if it must meet two conditions: (a) it must be theoretically progressive. That is, each new theory in the sequence must have excess empirical content over its predecessor; it must predict novel and hitherto unexpected empirical states of affairs, and (b) it must be empirically progressive, that is, some of that novel content actually has to be corroborated, that is, some of the newly predicted "facts" must actually turn out to be true (i.e., resist falsificatory attempts). A research program is degenerating if the successive theories do not deliver novel predictions or if the novel predictions that these deliver turn out to be false.

Thus research can be appraised by how the researcher addressed the Duhem–Quine thesis in

order to see if the test was actually risky. In a prediction failure, was the favored belief protected from the arrows of modus tollens by blaming the prediction failure on some auxiliary hypotheses? If this was in fact done, how good is the argument for doing this? Over time is the research program progressive (i.e., is its empirical content increasing and is some of this new content corroborated)? Finally, to what extent are auxiliary hypotheses being used illegitimately to reject the findings of the replication failure?

Popper Suggested Appraising a Theory's Problem-Solving Effectiveness

Popper emphasized science as problem-solving and the hypothesis and the findings need to be judged by their problem-solving effectiveness. Popper suggested that science begins with problems which can come from a wide variety of sources and proceeds according to this schema:

$$P1 \rightarrow TS \rightarrow EE \rightarrow P2 \rightarrow TS2 \rightarrow EE2 \ldots$$

where P is a problem; TS is a tentative solution; and EE is an error elimination attempt. There are many different kinds of problems in which the scientist may be interested. These can range from questions concerning particular matters of fact (e.g., "What is the surface temperature of Mars?"), to more general questions of fact (e.g., "What is the incidence of child sexual abuse in Canada?"), to questions of cause ("What causes an individual to sexually abuse a child?"), to deeper questions of cause ("Why do males sexually abuse much more frequently than females?"). Koertge suggested that problems arise for a variety of reasons,

> Scientific problems arise when our expectations are violated, when what we consider to be regularities call for a deeper explanation, when two previously disparate fields look as if they could be unified, or when a good scientific theory clashes with our familiar metaphysical framework (p. 347).

According to Popper a claim that has high empirical content and that survived a severe test and that functions like a problem solution would like to be more robust than a stray finding that is not understood in the context of a problem solution. For example, a finding that some therapy for severe depression results in statistically significant improvements (e.g., results in a 2-point reduction on the Beck Depression Inventory) as compared to a no-treatment control is not a problem solution. However, the large impact of certain antibiotics on strep throat, for example, curing 98% of individuals is much more like a problem solution—and thus more likely to be replicable. Popper would want both of these findings analyzed in terms of the extent to which these serve as problem solutions so that all would be clear on this. Both of these cases reported as producing "statistically significant" or even "clinically significant" results fail to clearly do this. Popper's basic notion is that a problem solution is much more likely to "carve nature at its joints" and thus it would reliably produces a change sufficiently large to solve the problem. Findings of smaller and more fragmented findings are more likely to be due to natural variation in the underlying phenomena or measurement error in the outcome measure which can produce a small 3-point decrement in the BDI. Rarely, are findings in psychology evaluated with respect to the extent to which these actually represent a problem solution. For example, O'Donohue and Moore (2007) found that in randomly controlled studies of childhood obesity, evidence-based programs generally only produced a 5 lb decrement during the first year. However, this benchmark was never clearly reported in these studies but needed to be computed from the data contained in the articles (when these were provided). Thus, in a replication failure the extent to which the original finding was a problem-solving solution rather than a small partial change ought to be clearly assessed.

Conclusions

The underlying epistemology of conventional research methods in psychology is unclear. However, there is little evidence that psychologists have relied on a Popperian epistemology in their research. This neglect may be part of the reason for the replication crisis. It may also explain why so few theories are considered to be shown to be false in psychology—ebbs and flows of theory popularity seem to influenced by a myriad of other factors than that these have been falsified (Meehl, 1978). In the Popperian model, research would be motivated by a sincere attempt to efficiently find error rather than based on the craving to be right. However, this attitude is not typically taught in research methods classes in psychology nor used in the appraisal of research. Instead, in clinical psychology, for

example, proponents of a favored therapy use QRPs to manufacture a series of positive randomly controlled trials (see, e.g., O'Donohue et al., 2016, for an example of this with Acceptance and Commitment Therapy). Adversarial research designs also might be useful toward moving toward increasing the likelihood that the research design, implementation, and interpretation is more consistent with this motivation.

Perhaps this also points to a pedagogical problem in the field: Research methods may be taught and generally conceptualized in a manner in which both their epistemological basis and related metascientific basis are neglected. Basic questions like, "If I implement design such and such, why does this produce knowledge?"; "What are the most efficient methods for uncovering knowledge in science, and why is this the case?" would be key to cover in research design courses. That is, the field following Meehl (1967) should have a better linkage of research design and epistemology, especially with regard to the question of how knowledge grows.

The empirical content of the theory (and hence its improbability) also is rarely appraised in psychology. If psychologists used Popperian severe testing, then it is more likely that any error in the original belief system would be uncovered, leaving less room for false beliefs to be uncovered by a subsequent replication failure. But, again, currently rarely do psychologists appraise the severity of the test or explicitly design tests to be severe. It is important to recognize that the use of QRPs would also serve (perhaps in a hidden manner) to decrease the severity of the test as these moves are made to either hide falsifications (e.g., the file drawer) or to prevent falsifying data to emerge (e.g., reanalyzing data until a statistically significant result is found). Thus, the entire open science movement designed to decrease the use of QRPs ought to have a salutary effect of increasing the severity of tests. Relatedly there has been little to no attention to the proper handling of the Duhem–Quine thesis and improperly handling this can lead to less risky tests in which initial problematic results are allowed to stand as well as replication failures to be minimized. In addition, original findings are not typically viewed as Popper would have us do as having survived only one among many, many possible tests but instead as passing a sufficient testing to providing enough evidence to justify the belief that the hypothesis under test is indeed true. A Popperian analysis would suggest a much more cautious epistemic position. In addition, the problem-solving effectiveness of the belief is not appraised and as such another important appraisal dimension is neglected. Hypotheses need to be evaluated on the extent to which they solve the problems these are designed to solve.

All this has implications both for the scientific training of psychologists and the appraisal of research programs by journal editors, grant reviewers, dissertation committees, and consumers. Instead of a narrower training in standard research methods and statistics, psychologists may profit from a deeper and broader training in the epistemology and the philosophy of science. Researchers need to be trained to be better able to authentically implement the proper motivation for conducting research (efficiently identifying error vs. proving themselves right) as well as to develop theories and hypotheses of greater empirical content. Severe testing needs to be taught alongside more standard ways of analyzing research designs such as internal and external validity. This all will likelihood necessitate a move away from null hypothesis testing (Fidler et al., 2018).

Finally, philosophers of science have a long tradition of trying to understand science from both what is known as an "internal" perspective—focusing on the logic of research as well as an "external" perspective—focusing on human and historical factors influencing the behavior of the scientist. Popper focused on cognitive factors such as the craving to be right (i.e., confirmation bias) as a distorting influence but certainly subsequent to Popper others have pointed out additional incentives that may distort science in a way that can produce results that are not replicable. For example, a blog (corelab.blog; March 5, 2020) contrasted what they call a "scientist as logician" perspective versus a "scientist as human" perspective. In the scientist as logician perspective the following is emphasized: (a) "Scientists are primarily truth-seekers, (b) Scientists rely on logic to develop the most efficient way of discovering truth, and (c) If a reformer uses logic to identify flaws in a scientist's current truth-seeking process, then, as long as the logic is sound, that scientist will change his or her scientific practices". In contrast, the "scientist as human" model, the following beliefs are emphasized: (a) "Humans, including scientists, have a wide variety of goals (including discovering the truth, being accurate, but also other goals such as demonstrating that he or she

is right, career advancement, fame, money, and so on), (b) Humans, including psychological scientists, are also embedded in complex social systems including political, economic, and professional ones, (c) Humans, including scientists, are sensitive the imperatives and incentives of these systems, (d) Reformers must attend to human goals as well as social, political, economic, and professional imperatives that influence the scientist to create lasting changes in human behavior as well as the social and political imperatives that might prevent scientists from engaging in a certain behavior, especially behavior consistent with the scientist as logician perspective. The reformer then works to align those imperatives with the desired behaviors".

A recent concrete step to instantiate this may be an increased use of "adversarial collaborations" (Kahneman & Klein, 2009) that can allow closer inspection and thus detection and eliminating either cognitive or human biases in the research project. This step has the potential to increase the severity of testing. As another possibility, Munafò et al. (2017) have suggested that researchers use statisticians who are blind to which data points represent the experimental condition and which represent the control conditions as a method of protecting against confirmation bias. Thus, the overall goal of these reforms would be to bring in new domains for potential criticism as perverse incentive systems in science especially those unduly favoring confirmations can also be critiqued and reformed. This is a useful distinction and analysis of the problem and both perspectives need to be considered to effectively improve the behavior of psychological scientists.

References

Aldhous, P. (2011). *Journal rejects studies contradicting precognition*. https://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition/

Bacon, F. (1960). *Novum organum*. Bobbs-Merrill. (Originally published 1621).

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. https://doi.org/10.1177/1745691612459060

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244.

Baron, R. M., Albright, L., & Malloy, T. E. (1995). Effects of behavioral and social class information on social judgment. *Personality and Social Psychology Bulletin*, *21*(4), 308–315.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. https://doi.org/10.1037/a0021524

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., & Pope, K. S. (1996). An update on empirically validated therapies. *Clinical Psychologist*, *49*(2), 5–18.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind but whose mind? *PLOS ONE*, *7*, Article e29081. https://doi.org/10.1371/journal.pone.0029081

Fidler, F., Thorn, F. S., Barnett, A., Kambouuris, S., & Kruger, A. (2018). The epistemic importance of establishing the absence of an effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 1–17.

Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology*, *22*, 364–368. https://doi.org/10.1016/j.annepidem.2012.02.007

Holtz, P. (2020). Two questions to foster critical thinking in the field of psychology. *Metapsychology*, *4*, 1–13.

Holtz, P., & Monnerjahn, M. (2017). Falsificationism is not just 'potential' falsifiability, but requires 'actual' falsification: Social psychology, critical rationalism, and progress in science. *Journal for the Theory of Social Behaviour*, *47*, 348–362.

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

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

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375.

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology*, *66*, 116–133.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526.

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 680–703.

Klein, R. A., Ratliff, K., Vianello, M., Adams, A. B., Jr., Bahník, S., Bernstein, N. B., & Cemalcilar, Z. (2014). Investigating variation in replicability. A "Many Labs" replication project. *Social Psychology*, *45*, 142–152.

Lakatos, I. (1980). *Falsification and the methodology of scientific research programs*. Cambridge University Press.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The [social] construction of scientific facts* (2nd ed.). Princeton University Press.

Laws, K. R. (2016). Psychology, replication and beyond. *BMC Psychology*, *4*, Article 30.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542.

Magee, B. (1974). *Popper*. Frank Cass.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175.

Mahoney, M. J., & DeMonbreun, B. G. (1977). Confirmatory bias in scientists and non-scientists. *Cognitive Therapy and Research*, *1*, 161–175.

Mahoney, M. J., & Kimper, T. P. (1976). From ethics to logic: A survey of scientists. In M. J. Mahoney (Ed.), *Scientist as subject*. Ballinger.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115. https://doi.org/10.1086/288135

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*, 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Mulkay, M., & Gilbert, G. N. (1981). Putting philosophy to work: Karl Popper's influence on scientific practice. *Philosophy of the Social Sciences*, *11*, 389–407.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J. & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature: Human Behaviour*, *1*, Article 0021.

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

O'Donohue, W. (2013). *Clinical psychology and the philosophy of science*. Springer.

O'Donohue, W., & Moore, B. (2007). Introduction. In W. O'Donohue, B. Moore & B. Scott (Eds.), *Handbook of adolescent and pediatric obesity treatment*. Routledge.

O'Donohue, W., Snipes, C., & Soto, C. (2016). A case study of overselling psychotherapy: An ACT intervention for diabetes management. *Journal of Contemporary Psychotherapy*, *46*(1), 1–25.

O'Donohue, W., & Willis, B. (2018). Problematic images of science in undergraduate psychology textbooks: How well is science understood and depicted? *Archives of Scientific Psychology*, *6*(1), 51–62.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac471.

Pettigrew, T. F. (1991). Toward unity and bold theory: Popperian suggestions for two persistent problems of social psychology. In C. W. Stephan, W. G. Stephan, & T. Pettigrew (Eds.), *The future of social psychology* (pp. 13–27). Springer.

Popper, K. R. (1959). *The logic of scientific discovery*. Routledge.

Popper, K. R. (1962). *The open society and its enemies* (5th ed., Vols. 1-2). Routledge.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford University Press.

Popper, K. R. (1999). *All life is problem solving*. Routledge.

Quine, W. V. O. (1990). Three indeterminacies. In R. B. Barrett & R. F. Gibson (Eds.), *Perspectives on Quine* (pp. 1–16). Blackwell.

Quine, W. V. O., & Ullian, J. (1978). *The web of belief*. Random House.

Schimmack, U. (2020) *A meta-psychological perspective on a decade of replication failures in social psychology*. Unpublished manuscript.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., & Srinivasan, M. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67.

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.

Trafimow, D. (2019). A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571–583. https://doi.org/10.1080/13645579.2019.1610592

Van Hiel, A., Onraet, E., & De Pauw, S. (2010). The relationship between social–cultural attitudes and behavioral measures of cognitive style: A meta-analytic integration of studies. *Journal of Personality*, *78*(6), 1765–1800.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (Vol. 1, pp. 106–137). Penguin Books.

Received June 12, 2020
Revision received December 9, 2020
Accepted December 11, 2020 ■