Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of *Psi* as a Case Study of Deficiencies in Modal Research Practice

Etienne P. LeBel

The University of Western Ontario

Kurt R. Peters

Norwich University

Correspondence concerning this article should be addressed to:

Etienne LeBel
Department of Psychology
The University of Western Ontario
Social Science Centre
London, Ontario N6A 5C2
Canada

Email: elebel@uwo.ca
Phone: (519) 661-2111 Ext. 80048

Abstract

In the following methodological commentary, we use Bem's (2011) recent article reporting experimental evidence for *psi* as a case study for discussing important deficiencies in modal research practice in empirical psychology. We focus on (a) over-emphasis on conceptual rather than close replication, (b) insufficient attention to verifying the soundness of measurement and experimental procedures, and (c) flawed implementation of null hypothesis significance testing. We argue that these deficiencies contribute to weak method-relevant beliefs, which in conjunction with overly strong theory-relevant beliefs lead to a systemic and pernicious bias in the interpretation of data that favors a researcher's theory. Ultimately, this interpretation bias increases the risk of drawing incorrect conclusions about human psychology. Our analysis points to concrete recommendations for improving research practice in empirical psychology. We recommend (a) a stronger emphasis on close replication, (b) routinely verifying the integrity of measurement instruments and experimental procedures, and (c) using stronger, more diagnostic forms of null hypothesis testing.

[156 words]

Key words: psi; close replication; NHST; file drawer problem; modal research practice.

Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of *Psi* as a Case Study of the Deficiencies in Modal Research Practice

> "At the heart of science is an essential tension between two seemingly contradictory attitudes—an openness to new ideas, no matter how bizarre or counterintuitive they may be, and the most ruthless skeptical scrutiny of all ideas, old and new. This is how deep truths are winnowed from deep nonsense."
>
> —Carl Sagan

In a recent issue of *Journal of Personality and Social Psychology*, Bem (2011) reported a series of nine experiments that are claimed to provide evidence for the existence of *psi*, specifically the anomalous retroactive influence of future events on an individual's current behavior. Bem's strategy for generating this experimental evidence was to reverse the causal direction of four well established psychological effects (i.e., priming, habituation, recall, and approach/avoidance). For instance, in a memory recall study, Bem found that participants were better at recalling rehearsed versus non-rehearsed words even though the words were rehearsed *after* the memory test had been completed. The goal of our commentary is to use Bem's article as a case study for discussing important deficiencies in modal research practice (MRP; Cook & Groom, 2004)—that is, the accepted methodology empirical psychologists most commonly employ in their research—and to suggest how these practices might be improved.

Bem (2011) deserves praise for his commitment to experimental rigor and the clarity with which he reports procedures and analyses, which generally exceed the standards of MRP in empirical psychology. That being said, it is precisely because Bem's report is of objectively high quality that it is diagnostic of potential problems with MRP. By employing accepted standards for experimental, analytic, and data reporting practices, yet arriving at a fantastic conclusion, Bem has put empirical psychologists in a difficult position: forced to consider either revising beliefs about the fundamental nature of time and causality or revising beliefs about the soundness

of MRP. In this commentary, we will explore the possibility that deficiencies in MRP can indeed provide an alternative explanation for the publication of Bem's article. In particular, we will focus on three methodological issues in Bem's work, each of which reflects a general deficiency in MRP: (a) over-emphasis on conceptual replication, (b) insufficient attention to verifying the integrity of measurement instruments and experimental procedures, and (c) problems with the way null hypothesis significance testing (NHST) is implemented. Taken singly, these deficiencies may appear relatively innocuous, but collectively they add up to a pernicious *interpretation bias* that skews the reporting of data in empirical psychology. We contend that it is this systemic bias in MRP, rather than any one crucial methodological flaw, that accounts for the publication of Bem's article in a top empirical journal.

Based on these considerations, we believe that the most valuable contribution of Bem's (2011) article is that, by revealing how general features of MRP bias the interpretation of data, it can promote needed discussion regarding improvements to research practice in empirical psychology. In fact, given that the methodological problems in Bem's article reflect quite general deficiencies in MRP, it follows that our criticisms and recommendations for improved practice apply directly to *all* research conducted within this tradition. Thus, throughout our commentary, it is important to keep in mind that Bem's article is exceptional only in terms of its findings; in contrast, its methodology and reporting practices adhere closely to (or even exceed) the accepted standards of MRP. The present commentary adopts Bem's article as a case study for this discussion simply because it makes the tension between confidence in methods versus results unusually obvious and difficult to ignore.

The Interpretation Bias

Our general argument is rooted in the fact that because empirical data underdetermine theory choice (Duhem, 1954; Quine, 1953), alternative explanations of data are always possible, both when the data statistically support the researcher's hypothesis and when they fail to do so. Deficiencies in MRP, however, lead to entrenched biases in the interpretation of data in both of these cases, which together constitute what we will call the *interpretation bias*: a bias toward interpretations of data that favor a researcher's theory—both when the null hypothesis is statistically rejected and when it is not. As realized in MRP, this bias entails that regardless of how data turn out, the theory whose predictions are being tested is artificially buffered from falsification (see Fanelli, 2010). The ultimate consequence is an increased risk of reporting false positives and disregarding true negatives, and so drawing incorrect conclusions about human psychology.[1]

When understood in terms of these converging pressures, that aspect of MRP known as the "file-drawer problem" is seen to be much more dangerous than is commonly acknowledged (Rosenthal, 1979). It is important to note, however, that this problem and the interpretation bias underlying it in no way depend upon unscrupulous motives. Due to the weakness of the knowledge system in empirical psychology (discussed in detail in the next section; see also Meehl, 1978), MRP leads even the most well intentioned researcher down the garden path to biased interpretations of data. Personal integrity, though necessary for sound science, is nevertheless insufficient because the theory-favorable bias in the interpretation of data in MRP is *systemic*. At worst, researchers can be accused of engaging in a process of motivated reasoning

[1] Indeed, our diagnosis of this interpretation bias is consistent with recent press coverage of difficulties with scientific replication and the "decline effect" in particular, whereby well-established effects subsequently shrink in size and become difficult to replicate (e.g., Lehrer, 2010; Zimmer, 2011). In line with Fanelli (2010), we suggest that this bias is particularly pernicious in empirical psychology due to the relative weakness of method-relevant beliefs in the field, which increases the ambiguity of data.

when it comes to the interpretation of data (Greenwald, Gonzalez, Harris, & Guthrie, 1996), but this process is strongly encouraged (if not required) by MRP. Thus, the ultimate target of our criticism is not any individual researcher but, more generally, modal research practice itself.

At the outset, we also want to make clear that our goal is not to apply a hyper-critical standard in the evaluation of *psi* research. Rather, we are interested in Bem's (2011) article precisely for what it tells us about accepted research standards in empirical psychology more broadly. As already mentioned, the criticisms we discuss are relevant to any research conducted within the tradition of MRP, regardless of whether the theories being tested are psychological or parapsychological. Ultimately, clarifying the ways in which theory-choice in MRP is biased points to concrete steps that can be taken to promote the high level of methodological rigor required for a cumulative science of psychology. These strategies include improvements to how empirical findings are replicated, how the integrity of measurement instruments and experimental procedures is verified, and how empirical tests of psychological theories are formulated.

Conservatism in Theory Choice

To understand how the interpretation bias arises from the weakness of the knowledge-system in empirical psychology, it will help to briefly discuss the issue of theory choice in science that lies at the heart of our argument. The knowledge-system that constitutes a science such as psychology can be roughly divided into two types of belief: *theory-relevant beliefs*, which concern the theoretical mechanisms that produce behavior, and *method-relevant beliefs*, which concern the procedures through which data are produced, measured, and analyzed.[2] In any

[2] To prevent any misconception, this distinction is employed here only for heuristic purposes and is not intended to invoke the positivist distinction between theory and data. Rather, theory-relevant beliefs and method-relevant beliefs are equally *beliefs*, differing only in their content, and are in principle equally susceptible to revision. Consequently,

empirical test of a hypothesis, interpretation of the resulting data depends on both theory-relevant and method-relevant beliefs, since both types of belief are required to bring the hypothesis to empirical test. Consequently, the resulting data can always be interpreted as theory-relevant, telling us something about the theoretical mechanisms underlying behavior, or as method-relevant, telling us something about the procedures employed to test the theoretical hypothesis.

So much is well known (if not well heeded) philosophy of science (see Duhem, 1954; Quine, 1953). Our argument, however, is not simply that data underdetermine theory choice, but that weaknesses in the current knowledge-system of empirical psychology *bias* the resulting choice of interpretation in favor of the researcher's theory. In particular, deficiencies in MRP systematically bias (a) the interpretation of confirmatory data as theory-relevant and (b) the interpretation of disconfirmatory data as method-relevant, with the result that the researcher's hypothesis is artificially buffered from falsification. The interpretation of data, however, should hinge not on what the pertinent beliefs are about (i.e., theoretical mechanisms versus empirical methodology) but rather on the *centrality* of those beliefs. The centrality of a belief reflects its position within a knowledge-system: Central beliefs are those upon which many other beliefs depend (e.g., the belief that temporal precedence is a property of causation), whereas peripheral beliefs are those with few dependent beliefs (e.g., the validity of a novel psychological measurement instrument). The rejection of central beliefs to account for observed data thus entails a major restructuring of the overall knowledge-system, whereas the rejection of peripheral beliefs entails little or no restructuring.

Quine and Ullian (1978) referred to the use of belief-centrality as a criterion for theory choice as *conservatism*: choosing the theoretical explanation consistent with the data that

---

the distinction is not a sharp one, but is nevertheless useful for describing the composition of a scientific knowledge-system.

requires the least amount of restructuring of the existing knowledge-system. Generally, conservatism in theory choice is a virtue, as it reduces ambiguity in the interpretation of data. For example, when method-relevant beliefs are relatively central, a methodology is considered *rigorous*. The value of methodological rigor is precisely that, by leveraging conservatism, it becomes more difficult to blame negative results on flawed methodology; this constrains the field of alternative explanations and so makes empirical tests more diagnostic. Conversely, when method-relevant beliefs are peripheral and easily rejected, empirical tests become more ambiguous.

Theory-relevant beliefs, in contrast, should not be so central that they approach the status of logical necessity. Rather, a theory's strength should be measured by the extent to which it is falsifiable, as judged by its fecundity for deriving falsifiable predictions (Popper, 1963). Theories that are too central risk becoming logical assumptions that are near-impossible to dislodge with empirical test. Thus, it is critical that a hypothesis under test be described in a way that makes it empirically falsifiable and not logically necessary.

Unfortunately, the knowledge-system in empirical psychology is such that conservatism becomes a vice rather than a virtue in theory choice. On the one hand, method-relevant beliefs are too peripheral, making them easy to reject. This increases the ambiguity of negative results, which contributes directly to the file drawer problem. On the other hand, theory-relevant beliefs often appear too central, making them difficult to reject. This leads to a process of confirmatory hypothesis testing, exacerbating the file drawer problem. Below we will address three specific ways in which method-relevant beliefs in empirical psychology are too weak (i.e., peripheral), taking Bem's (2011) article as a case study to illustrate each point, and then briefly address difficulties with the logical strength of theory-relevant beliefs. In conjunction, these deficiencies

in MRP contribute to a pernicious interpretation bias that increases the risk of drawing incorrect conclusions from evidence—despite the fact that this evidence is produced using accepted standards for research in empirical psychology. In our opinion, it is this systemic bias, rather than any single methodological flaw, that provides an alternative explanation for the publication of Bem's article.

Deficiencies in Modal Research Practice

Over-emphasis on Conceptual Replication

Bem (2011) reports nine experiments, each of which would be considered conceptual replications; in other words, none of the nine experiments were replicated exactly (i.e., without intentionally introducing procedural differences). The exclusive focus on conceptual replication is in keeping with the ethos of "continuous theoretical advancement" that is a hallmark of MRP. An over-emphasis on conceptual replication at the expense of close replication, however, weakens method-relevant beliefs in the knowledge-system of empirical psychology, with the result that reports consisting entirely of conceptual replications may be less rigorous than those including a judicious number of close replications.

Typically in MRP, a statistically significant result is followed by a conceptual replication in the understandable interest of extending the underlying theory. The problem with this practice, of course, is that when the conceptual replication fails, it remains unclear whether the negative result was due to the falsity of the underlying theory or to methodological flaws introduced by changes in the conceptual replication. Given the original, statistically significant finding, however, the natural preference is to choose the latter interpretation (see Meehl, 1967, p. 114) and to proceed with another, slightly different, conceptual replication. This process can be

repeated a number of times until a second statistically significant finding is achieved, and in such cases each of the "methodologically flawed" conceptual replications ends up in the file drawer.

What starts out as an attempt to extend an original finding can thus end up compromising confidence both in that finding and its eventual extension. The danger arises because conceptual replication allows the researcher too much latitude in the interpretation of negative results. In particular, the choice of which studies count as replications is made post-hoc, and these choices are inevitably influenced by the interpretation bias: An extension that fails to reject the null hypothesis is not counted as a replication *precisely because it did not replicate the original finding*, and therefore the altered methodology must be to blame. The consequence is that a successful extension becomes a conceptual replication, whereas a failed extension becomes a methodologically flawed pilot study (Miller, 2009)—and it is tacitly understood that failed pilot studies belong in the file-drawer.

The emphasis on conceptual replication in MRP might be defended by objecting that, due to the context-sensitivity of psychological processes, it can be tricky to get a procedure "just right." We agree that new areas of research, including Bem's (2011) research on *psi*, often require "extensive pilot testing" (p. 47) to configure the experimental procedures and measurement instruments; however, this fact remains weak warrant for the inflation of Type I error that the process of conceptual replication in MRP often entails. A second objection could be that although our argument applies in the case of failed replications, in many high impact papers, and in Bem's article in particular, the reporting of several successful conceptual replications can actually be seen as more compelling than several successful exact replications, because the results were duplicated using slightly different procedures or measures. Yet this is true *if and only if* the successful replications were not achieved at the expense of many failed

"pilot studies." For most empirical psychology articles, including Bem's article, this possibility cannot be confidently ruled out given the systemically biased interpretation of negative results in empirical psychology, as described above.[3]

Integrity of Measurement Instruments and Experimental Procedures

Nowhere in Bem's (2011) article is an attempt made to verify the integrity of measurement instruments and experimental procedures. To begin with, no effort is made to verify that the measurement instruments used to assess the primary dependent variable (e.g., hit rates indicative of precognition) are operating correctly; for example, no reliability estimates are reported for the dependent variable or for the individual difference measure of sensation seeking. As is typical of MRP, neither of these measures are previously validated, but are rather designed ad hoc for the purposes of Bem's studies (e.g., the ad hoc 2-item measure of sensation seeking). In addition, no effort is made to verify that the various experimental procedures used in the nine studies are operating correctly, other than ensuring that known effects can be replicated (e.g., as with the standard priming effects in Experiments 3 and 4). The failure to verify the integrity of measurement instruments and experimental procedures directly weakens method-relevant beliefs, and thus increases ambiguity in the interpretation of negative (and even positive) results.

Admittedly, determining whether a manipulation or measurement procedure is operating correctly raises difficult issues in empirical psychology (Borsboom, Mellenbergh, & van Heerden, 2004; Runkel, 2007), many of which require treatment independently of, and prior to, the use of such procedures in tests of substantive psychological hypotheses. Partly because of these challenges, little effort is put into independently validating and calibrating methodological

[3] The possibility that successful conceptual replications were achieved at the expense of many failed "pilot studies" also cannot typically be ruled out given the publication practice of current psychology journals, which does not require authors to report results of *all* attempts to produce the effects reported in submitted articles. It is conceivable, however, that such a system could be implemented in the future.

procedures in MRP outside of the main theory-testing experiments. Instead, experiments are required to verify procedures *and* test psychological theories simultaneously. The result is that it becomes easy to attribute negative results to methodological flaws, and hence relegate them to the file drawer.

Although pilot studies confirming the operation of construct manipulations are sometimes reported in multi-experiment articles, such verification studies are not consistently performed given that they are not required for publication. And even when manipulation checks are reported, it can often be difficult to determine whether the manipulation had its intended effect on participants due to the disconnect between a researcher's "operational definition" of a construct and a participant's subjective interpretation of a stimulus (Runkel, 2007). This difficulty reflects the more general problem of construct validity in psychology, which arises because the context-sensitivity of psychological processes makes it very difficult to know when a manipulation or measurement is valid (Borsboom et al., 2004; Michell, 1997; Peters, 2011).

The integrity of measurement procedures is also often difficult to substantiate. For instance, reliability estimates for test scores are frequently not reported (Gawronski, Deutsch, & Banse, 2011; Kashy, Donnellan, Ackerman, & Russell., 2009; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). Moreover, due to the small cell sizes typically used in experimental designs (Maxwell, 2004), it is often impossible to determine accurate reliability estimates of test scores within experimental conditions (LeBel & Paunonen, 2011). And even when reliability can be accurately estimated, this methodological check is, strictly speaking, only the tip of the iceberg in determining whether observed scores primarily reflect the construct of interest rather than some other construct (Messick, 1989). Taken together, the inconsistent, informal, and arduous

nature of verifying the integrity of manipulation and measurement procedures leaves method-relevant beliefs much weaker than required for a rigorous empirical science.

Problems with NHST

As is typical in MRP, Bem (2011) treats null hypothesis significance tests as the sole criterion for determining theory choice within experiments. Exclusive reliance on the number .05 is problematic, however, both because (a) the standard null hypothesis of no difference will almost always be false and because (b) it divorces theory choice from the context of the broader scientific knowledge-system, encouraging myopic interpretations of data that can lead to bizarre conclusions about what has been empirically demonstrated. Even though it might be argued that the use of a null hypothesis of no difference *is* theoretically appropriate in Bem's tests for precognition, the fact remains that NHST, as implemented by Bem and in MRP generally, is biased *against* this null hypothesis (Wagenmakers, 2007). Thus, while it is well known that negative ("null") results are ambiguous and difficult to interpret, exclusive reliance on NHST makes positive results equally ambiguous, since they can be explained by flaws in the way NHST is implemented rather than by a more theoretically interesting mechanism (Meehl, 1967). In this way, exclusive reliance on NHST increases the ambiguity of theory choice and undermines the rigor of empirical psychology (Meehl, 1978; see also Wilkinson & the Task Force on Statistical Inference, 1999).

The first problem in this regard is that, in MRP, the null hypothesis is most often formulated as a "nil hypothesis" (Cohen, 1994; Tukey, 1991), which claims that the means of different populations are identical. This is a weak hypothesis because it is almost by definition false: Differences between different populations are inevitable, even if they only reflect ambient noise or "crud" (Lykken, 1968; Meehl, 1990). The statistical rejection of the nil hypothesis is

therefore contingent only upon a sample size sufficient to make the difference between means *statistically* significant (Kirk, 1996). In the context Bem's (2011) experiments, for example, we contend that the reliance on one-sided, one sample *t*-tests with sample sizes of $N = 100$ (and even $N = 150$ and $N = 200$) is sufficient to exploit the nil hypothesis (e.g., consider Experiment 2's hit rate of 51.7%, which achieved statistical significance with a sample size of $N = 150$).

Simply put, the nil hypothesis is a straw man—a bit of statistical fluff with no theoretical substance—and because the nil hypothesis is not theory-driven, it is hard to argue that its rejection implies anything whatsoever about the choice of an alternative hypothesis. The rejection of the nil is therefore not equivalent to the rejection of a *theoretically appropriate* null hypothesis, and assuming that it is leads to the inflation of Type I error. Specifically, a Type I error rate of .05 derived from the nil hypothesis will in most empirical instances be an underestimate, because a true nil difference in the population is extremely unlikely; in contrast, a non-nil difference is more likely to occur. Interpreting the rejection of the nil hypothesis as support for the researcher's own theory therefore runs a higher-than-.05 risk of being a false positive (Kline, 2004).

A second problem with NHST is that treating statistical significance as the sole criterion of theory choice when interpreting new data, as is typically the case in MRP, ignores all other evidence relevant to the interpretation of those data. Empirical tests are not conducted in a theoretical vacuum, and existing evidence for or against a hypothesis should be factored into the interpretation of new data to supplement NHST. Furthermore, NHST on its own does not tell us what we want to know (i.e., the updated probability of the null given the new data) but something much less informative (i.e., the probability of the data given that the null is true; Cohen, 1994; Dawes, 1988). Basing theory choice on null hypothesis significance tests thus

detaches theories from the broader knowledge-system of empirical psychology.[4] Combined with the bias against the nil hypothesis in NHST, this myopic view of data strengthens the interpretation bias, making unlikely theories such as *psi* appear more probable than they otherwise would. In the long run, this over-reliance on NHST threatens the cumulation of evidence and the coherence of the knowledge-system in empirical psychology (e.g., Bakan, 1996; Meehl, 1978; Kirk, 1996; Rozeboom, 1997; Thompson, 1996). As Kirk (1996) noted, "Our science has paid a high price for its ritualistic adherence to NHST" (p. 756).

The Logical Strength of Theory

Weak, peripheral method-relevant beliefs make it easy to discount negative results. The motivation to do so is amplified, however, by the apparent centrality of theory-relevant beliefs in psychology: The more it appears that a theoretical explanation *has* to be the case, the more likely it is that disconfirming data will be attributed to methodological flaws. Indeed, a longstanding criticism of psychological hypotheses is that they frequently approach the status of logical necessity (Gergen, 1982; McGuire, 1973; Wallach & Wallach, 1994). Given that participants act rationally and understand the operative contingencies in a situation, the psychological hypothesis explaining their behavior becomes almost logically necessary; as McGuire (1973, p. 449) put it, "Experiments on such hypotheses naturally turn out to be more like demonstrations than tests."

Although this criticism does not apply to all psychological hypotheses, the damage has been done by the aura of logical necessity that psychological hypotheses *as such* have acquired. This aura of necessity makes psychological hypotheses inherently appear central to the overall knowledge-system, regardless of what they actually claim—indeed, this is made clear by the fact that even Bem's (2011) parapsychological hypothesis is able to take advantage of the weak

[4]It is worth mentioning that any statistical technique (i.e., NHST, Bayesian approaches, or otherwise) used in isolation and divorced from substantive theoretical considerations will be insufficient for determining theory choice.

method-relevant beliefs in empirical psychology, as described above. Consequently, the interpretation of negative results is by default biased toward favoring the researcher's theory, since rejecting it would presumably require more extensive revisions to the knowledge-system (e.g., requiring us to assume that people do not in general act rationally) than would rejecting beliefs about the integrity of particular methodological procedures.

Summary

The result of the combination of peripheral method-relevant beliefs and central theory-relevant beliefs is that conservatism in MRP becomes an unconditional bias toward interpretations of data that favor the researcher's theory: It is easy to reject weakly held method-relevant beliefs when results disconfirm a strongly held theory; on the other hand, the ambiguities of weak method-relevant beliefs are discounted or ignored when results confirm the theory. Yet conservatism should only bias theory choice toward interpretations of data that minimize revision of the knowledge-system, regardless of whether a particular interpretation favors method-relevant or theory-relevant beliefs. The extent to which this aspect of the interpretation bias is entrenched in MRP, and the extent to which conservatism is no longer sensitive to belief centrality *tout court*, is made obvious by the publication of Bem's (2011) article: Even when the primary theoretical beliefs being tested are extremely peripheral (as parapsychological beliefs surely are within a naturalistic knowledge system), MRP may still bias interpretations of data in ways that ultimately favor the theory.

Strategies for Improving Modal Research Practice

By focusing on this systemic interpretation bias—which we believe accounts for the publication of Bem's (2011) evidence for *psi* in a respected journal—our analysis points directly to strategies for improving research practice in psychology. The overarching recommendation is

that methodology must be made more rigorous by strengthening method-relevant beliefs, in order to constrain the field of alternative explanations available for a psychological finding. This is true both when data statistically support a researcher's theory and when they do not: By making MRP more rigorous, the ambiguity of theory choice is reduced and empirical tests become more diagnostic. A complementary recommendation is that the logical status of theory-relevant beliefs must be weakened. We will suggest three concrete strategies for strengthening method-relevant beliefs and also provide preliminary recommendations for making theory-relevant beliefs in psychology easier to reject.

Recommendations for Strengthening Method-Relevant Beliefs

Stronger emphasis on close replication. First, MRP would benefit greatly from a stronger emphasis on close relative to conceptual replication, which echoes recent recommendations in other fields of inquiry (e.g., Moonesinghe et al., 2007). Across all scientific disciplines, close replication is the gold standard for corroborating the discovery of an empirical phenomenon (Lindsay & Ehrenberg, 1993; Sohn, 1998; Guttman, 1977; Fisher, 1934; Falk, 1998), and the importance of this point for psychology has been noted many times by methodologists and statisticians (Cohen, 1994; Cumming, 2008; Greenwald et al., 1996; Rosenthal, 1993; Tukey, 1967; Thompson, 1992; Falk & Greenbaum, 1993). Indeed, the fundamental scientific axiom of repeatability requires that what has occurred once under specific conditions will occur again under those same conditions (Dennis, 1926). Even the inventor of statistical significance tests, Ronald A. Fisher (1934), strongly emphasized the need for close replication to determine whether an observed effect is real or simply due to sampling error: Confidence in whether our results are real, Fisher stressed, can only be achieved via "agreement

between *parallel* experiments" and our confidence should increase after each replication (Fisher, 1925, p. 111, emphasis added; see also Greenwald et al., 1996).

It is critical to realize that the type of replication these methodologists and statisticians have in mind is *close* replication rather than conceptual replication. Particularly in the early stages of research, close replications are necessary to ensure that an effect is real and hence can be reliably reproduced under the exact same procedural conditions. As Mulkay and Gilbert (1986) state: "We can agree about results because we have learned that experiments carried out under precisely the same conditions do actually lead to the same results" (p. 21). This point is further strengthened by considering an important statistical assumption underlying the logic of NHST: *p*-values calculated within the NHST framework depend on values derived from sampling distributions of many repetitions of the *same* experiment (Krushke, 2010). Hence, it follows that close replications are more diagnostic regarding the veracity of an experimental result (compared to conceptual replications) because the probability of making a Type I error systematically decreases with each successive close replication (Miller, 2009), whereas with each conceptual replication the Type I error rate is reset to .05 (see also Amir & Sharon, 1991).

Furthermore, close replications are crucial because a failed close replication is the most diagnostic test of whether an observed effect is real or not given that no differences between the original study and the replicating study were intentionally introduced. It follows that confidence in a negative result increases directly the closer the design of the replication study is to that of the original study (Lindsay & Ehrenberg, 1993). In the case of a close replication, we cannot easily blame a negative result on methodological variation, because in a close replication methodological differences are not deliberately introduced into the replication.[5] Once successful

[5] Of course, it is possible that there may be a change in the methodological integrity of the measurement instruments or procedures in a close replication, which is precisely why verification of their integrity needs to be routine.

close replications have been achieved in a new area of research, however, the value of further close replications diminishes and the value of conceptual replications increases dramatically (Collins, 1984). We contend that this important balance between close and conceptual replication has been almost completely overlooked in MRP; consequently, close replications require much stronger emphasis in everyday research practice (see also Rosenthal, 1991, and Hendrick, 1991, for other approaches to replication).

Verify integrity of methodological procedures. Second, to make method-relevant beliefs stronger and more difficult to reject, it is critical that verifying the integrity of empirical instruments and procedures become a routine component of psychological research. In this spirit, it should be standard practice to verify the integrity of manipulation procedures and measurement instruments outside of the main theory-testing experiments. Respecting this distinction requires that pilot studies be explicitly designed to fine-tune manipulations or calibrate measurement instruments; consequently, pilot studies should be identified as such *a priori*, before results are known. Maintaining a clear distinction between pilot studies designed to verify the integrity of instruments and procedures and primary studies designed to test theories will do much to diminish the influence of the interpretation bias on the reporting of results.

Beyond placing more emphasis on verifying the integrity of empirical procedures in dedicated pilot studies, it should also be standard practice to routinely check the internal consistency of the scores of any measurement instruments that are used (Kashy et al., 2009; LeBel & Paunonen, 2011) and to confirm measurement invariance of instruments across conditions (DeShon, 2004). It should also be standard practice to employ objective markers of instruction comprehension (e.g., Oppenheimer, Meyvisb, & Davidenkoc, 2009) and participant non-compliance (e.g., by recording time spent on instruction screens). Again, these checks and

balances need to be applied in a systematic, a priori manner to strengthen methodological beliefs and reduce the ambiguity of negative results.

A related strategy is to develop a principled account of the context-sensitivity of psychological processes, so that the problem of construct validity in psychology—that is, the question of *what* is being manipulated or measured—can be made manageable (Borsboom et al., 2004). Toward this end, it may help to develop an empirically supported account of how the context-sensitivity of mental processes varies under different operating conditions (Peters, 2011). Such an account would help to guide the selection of instruments and procedures appropriate for testing a given hypothesis, and would also constrain the theoretical interpretation of data according to the conditions under which they were observed.

Use stronger forms of NHST. Finally, confidence in a researcher's theory will be increased to the extent that a rejected null hypothesis implies a theoretically substantive alternative. Minimally, null hypotheses should not be formulated in terms of a "nil" hypothesis of identical populations, given that such a hypothesis is at an inherent empirical disadvantage. It has long been argued that this weak form of NHST does not provide strong tests of theoretical hypotheses (Cohen, 1994; Kirk, 1996; Meehl, 1967; 1978; 1990; see also Mulaik, Raju, & Harshman, 1997). Hence, methodologists have called for a stronger form of NHST (e.g., Popper, 1959), which has been employed in astronomy and physics for centuries. In its strong form, NHST requires that the null hypothesis be a theoretically derived point value of the focal variable, which the researcher then attempts to reject upon observation of the data (Meehl, 1967). In physics, this involves comparing a theoretically predicted value x_o with the observed mean $\bar{x}_o$, and asking whether the predicted value falls within the band of probable error of the empirically

observed mean (Meehl, 1967).[6] More broadly, significance tests should be treated as just one criterion informing theory choice, in addition to relevant background knowledge and considerations of belief centrality. Bayesian analytic techniques, which explicitly seek to incorporate base-rate information into hypothesis testing, may help push the interpretation of psychological data in this direction (Krushke, 2010; Wagenmakers, 2007; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

Recommendations for Weakening Theory-relevant Beliefs

A final, complementary recommendation for improving MRP is to weaken the aura of necessity that attaches to psychological hypotheses. Considered individually, not all psychological hypotheses appear logically necessary, but insufficient attention has been paid to identifying the criterion that distinguishes between falsifiable and non-falsifiable psychological hypotheses, particularly within the cognitive paradigm (see Gergen, 1982; McGuire, 1973; Wallach & Wallach, 1994). This question requires deeper consideration, but an initial heuristic might involve distinguishing between cognitive hypotheses that depend on the conventional meaning of words—which amount to descriptions of rational behavior—versus those that depend on the subjective, context-sensitive meaning of a stimulus. Whereas hypotheses of the first type will necessarily be confirmed given that (a) experimental contingencies are clearly communicated to participants and (b) participants are motivated to achieve desired goals, hypotheses of the second type may depend on non-normative, contingent processes, particularly those that operate automatically or below linguistic awareness. Critically, hypotheses of the

[6] A concrete example, summarized by Mulaik, Raju, and Harshman (1997), may help to illustrate this point. Early in the 20th century, Newton's theory of gravity predicted that gravitation would deflect light from a star passing near the edge of the sun by one-half the amount predicted by Einstein's theory of relativity (0".87 r_0/r vs. 1".75 r_0/r, where r_0 = the radius of the sun and r = the closest distance of the star's light to the center of the sun). Data from two independent observation sites during a total eclipse of the sun confirmed that Einstein's predicted value fell within the band of probable error of the observed value for both sites whereas Newton's predicted value fell outside the band, hence supporting Einstein's theory over Newton's theory.

second type are easier to reject, since their not being the case does not imply that people do not act rationally.

Regardless of how this criterion is drawn, the important point is that making the disconfirmation of a psychological hypothesis more plausible will reduce the bias toward methodological interpretations of negative results. At minimum, then, care needs to be taken that hypotheses under test are stated such that their *not* being the case is possible, so that their truth is contingent rather than necessary. When the researcher's hypothesis is plausibly falsifiable and the null hypothesis is plausibly confirmable, statistical tests pitting these two hypotheses against each other will be much more informative for theory choice. Thus, making methodological beliefs more central and theoretical beliefs less central has the salutary effect of making empirical data more diagnostic, reducing the potential for the interpretation bias to produce false positives and suppress true negatives.

Conclusion

Modal research practice in empirical psychology is systemically biased toward interpretations of data that favor the researcher's theory. This interpretation bias arises because method-relevant beliefs are too peripheral and theory-relevant beliefs are too central in the knowledge-system of empirical psychology. The three methodological problems in Bem's (2011) article reviewed above, which reflect general deficiencies in MRP, contribute directly to weak method-relevant beliefs: (a) the over-emphasis on conceptual relative to close replication (i.e., close replications are virtually never reported in published articles), (b) the failure to verify the integrity of measurement instruments and experimental procedures (e.g., reliability estimates are often not reported for experimental dependent variable measures; studies often conflate the verification of instruments and procedures with the testing of substantive theories), and (c)

flawed implementation of NHST (i.e., testing against an inadequate null hypothesis; focusing exclusively on significance tests for determining theory choice).[7] Ultimately, these deficiencies lead to an interpretation bias that increases the risk of reporting false positives and disregarding true negatives, and ultimately of drawing incorrect conclusions about human psychology. Because this bias is systemic, it is subtle and typically goes unacknowledged in day-to-day research practice and during the peer-review process; indeed, this bias is particularly troublesome because many of the practices that contribute to it are invisible to the peer-review process. Nevertheless, it is a pernicious bias that skews the interpretation of evidence relevant to explanatory theories in psychology.

Although this systemic bias can account for the publication of Bem's (2011) evidence for *psi*, acknowledging the existence of this bias seriously undermines confidence in the rigor of MRP in empirical psychology. The objection might be made that the prima facie choice presented here—between accepting the reality of *psi* versus accepting that MRP can lead to unwarranted conclusions—is a false dichotomy; instead, we should wait and see if Bem's results replicate before drawing any conclusions. This objection, however, only delays the inevitable. If Bem's data do replicate, we are still faced with the same choice: explain them by appeal to parapsychological theories (assuming naturalistic theoretical explanations are not forthcoming) or as due to deficiencies in MRP. On the other hand, if Bem's data fail to replicate, we must nevertheless account for the fact that the data were collected and published in a top journal while respecting (and even exceeding) the standards of methodological rigor in MRP.

Thus, at a higher level, the publication of Bem's (2011) article forces empirical psychologists to choose between two interpretations of anomalous data collected using accepted

[7] It is of course possible that there are other deficiencies, beyond the three discussed in this commentary, that also contribute to the interpretation bias in MRP.

standards of methodological rigor: one that requires dramatic revision of the knowledge-system of natural science with respect to beliefs about time and causality, and another that requires revision of beliefs about the methodological rigor of empirical psychology. It is much more in keeping with the virtue of conservatism to favor the methodological interpretation of Bem's evidence over the theoretical interpretation he prefers, for the straightforward reason that the latter upends theoretical commitments at the core of scientific knowledge, whereas the former occasions the revision only of relatively peripheral beliefs—specifically, beliefs about the rigor of MRP in empirical psychology.

In light of these costs, the choice between these interpretations is made even starker when we consider what each buys us. With Bem's (2011) preferred parapsychological interpretation, the increase in confidence given prior beliefs about the probability of *psi* is miniscule, but comes at a massive cost—that is, the cost of revising beliefs about time and causality at the core of the naturalistic knowledge-system. In contrast, a methodological interpretation provides no gain, but relatively minimal cost. Indeed—before objecting that these costs are far from minimal—psychologists might consider how myopic it would be to favor the theoretical interpretation over the methodological interpretation as a means of defending MRP, given that the former requires revision of almost everyone's beliefs about the world, whereas the latter requires belief revision only within the subpopulation of empirical psychologists.

Moreover, given that acknowledging deficiencies in MRP points directly to strategies for its improvement, it can easily be argued that a methodological interpretation of Bem's (2011) results is a net gain for empirical psychology. Specifically, to strengthen method-relevant beliefs we recommend that researchers (a) pay greater attention to the balance between close versus conceptual replication, (b) routinely verify the integrity of measurement instruments and

experimental procedures, and (c) use stronger, more diagnostic forms of NHST. With regard to weakening the logical status of theory-relevant beliefs, we suggest that this longstanding criticism of psychological explanation (e.g., Gergen, 1982; McGuire, 1973; Wallach & Wallach, 1994) receive more attention in the field as a whole. An explicit description of the criterion distinguishing falsifiable from non-falsifiable psychological hypotheses will go a long way toward dispelling the aura of necessity that currently attaches to such hypotheses, making them easier to reject in the face of disconfirmatory data.

In keeping with the words of Carl Sagan that opened our commentary, we are not arguing that *psi* phenomena are impossible. Impossibility implies certainty, and science is in the business of calibrating confidence, not establishing truth. Rather, the point is that good philosophy of science, and indeed "good sense" (Duhem, 1954), suggest that the response to Bem's (2011) data ought to be conservative. Although this choice spares the great majority of scientific knowledge from revision, it casts severe doubt on beliefs about the rigor of modal research practice in empirical psychology. Bem's greatest contribution may thus be the unintended one of putting psychologists in a position where they are unable to ignore longstanding criticisms of their research practice—or else be forced to accept a fantastic theory about the workings of the natural world. In any event, a choice must be made, and it should be clearly recognized that the continuation of the status quo represents an implicit choice endorsing Bem's parapsychological conclusions; the only alternative would be to conclude that psychologists are comfortable continuing to conduct research with an unsound methodology.

References

Amir, Y., & Sharon, I. (1991). Replication Research: A 'Must' for the scientific advancement of psychology. In J. W. Neuliep (Ed.), *Replication Research in the Social Sciences*. Newbury Park, CA: SAGE Publications, Inc.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 1-29.

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.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061-1071.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Collins, H. M. (1984). Discussion: When do scientists prefer to vary their experiments. *Studies in the History and Philosophy of Science*, *15*, 169-174.

Cook, T. D., & Groom, C. (2004). The methodological assumptions of social psychology: The mutual dependence of substantive theory and method choice. In C. Sansone, C. C. Morf & A. T. Panter (Eds.), *The sage handbook of methods in social psychology* (pp. 19-44). Thousand Oaks, CA: Sage.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286-300.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

Dennis, W. (1926). The experimental methods of psychology. In C. Murchison (Ed.), *Psychologies of 1925.* Worcester: Clark University Press.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137-149.

Duhem, P. (1954). *The aim and structure of physical theory*. Trans. P. P. Wiener. Princeton, NJ: Princeton University Press.

Falk, R. (1998). Replication – A step in the right direction. *Theory & Psychology*, *8*, 313-321.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology*, *5*, 75-98.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE 5(4): e10068*.

Fisher, R. A. (1926). *Statistical methods for research workers* (1st ed.). London: Oliver & Boyd.

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). London: Oliver & Boyd.

Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. C. Klauer, C. Stahl, & A. Voss (Eds.), *Cognitive methods in social psychology* (pp. 78-123). New York: Guilford Press.

Gergen, K. J. (1982). *Toward transformation in social knowledge*. New York: Springer-Verlag.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect size and p values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175-183.

Guttman, L. (1997). What is not what in statistics. *The Statistician*, *26*, 81-107.

Hendrick, C. (1991). Replications, strict replications, and conceptual replications: Are they important? In J. W. Neuliep (Ed.), *Replication Research in the Social Sciences*. Newbury Park, CA: SAGE Publications, Inc.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131-1142.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Kline, R. B. (2004). *Beyond significance testing*. American Psychological Association, Washington, DC.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658-676.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: Impact of unreliability on the replicability of experimental findings involving implicit measures. *Personality and Social Psychology Bulletin*, *37*, 570-583.

Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker.* Retrieved July 20, 2011 from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician*, *47*, 217-228.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-159.

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

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103-115.

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.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195-244.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology*, *88*, 355-383.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*, 617-640.

Moonesinghe, R., Khoury, M. J., & Janssens, C. J. W. (2007). Most published research findings are false – But a little replication goes a long way. *PLoS Medecine*, *4*, 218-221.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Erlbaum.

Mulkay, M., & Gilbert, G. N. (1986). Replication and mere replication. *Philosophy of the Social Sciences*, *16*, 21-37.

Oppenheimer, D. M., Meyvisb, T., & Davidenkoc, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867-872 .

Peters, K. R. (2011). Cronbach's challenge: Putting psychological explanation in context. Unpublished manuscript, The University of Western Ontario, London, Canada.

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

Popper, K. R. (1963). *Conjectures and refutations*. London: Routledge and Kegan Paul.

Quine, W. V. O. (1953). Two dogmas of empiricism. In *From a logical point of view* (2nd ed., pp. 20-46). Cambridge: Harvard University Press.

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (2nd ed.). New York: Random House.

Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication Research in the Social Sciences*. Newbury Park, CA: SAGE Publications, Inc.

Rosenthal, R. (1993). Cumulating evidence. In (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519-559). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-391). Hillsdale, NJ: Erlbaum.

Runkel, P. J. (2007). *Casting nets and testing specimens: Two grand methods of psychology*. Hayward, CA: Living Control Systems Publisher.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, *8*, 291-311.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, *70*, 434-438.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83-91.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100-116.

Vacha-Haase, T., Ness, C., Nilsson, J.,& Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education*, *67*, 335-341.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. 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.

Wallach, L., & Wallach, M. A. (1994). Gergen versus the mainstream: Are hypotheses in social psychology subject to empirical test? *Journal of Personality and Social Psychology*, *67*, 233-242.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Zimmer, C. (2011, June 25). It's science, but not necessarily right. *The New York Times*. Retrieved July 20, 2011, from http://www.nytimes.com/2011/06/26/opinion/sunday/26ideas.html

Authors Note

Both authors contributed equally to the manuscript. We would like to thank Paul Conway, Chester Kam, and Yang Ye for their valuable comments on a previous version of this article.