Behavioral and Brain Sciences (forthcoming)

This Target Article has been accepted for publication and is currently under commentary. This article has not yet been copyedited and proofread. The article may be cited using its doi (About doi), but it must be made clear that it is not the final version.

MAKING REPLICATION MAINSTREAM

Rolf A. Zwaan, Alexander Etz, Richard E. Lucas, & M. Brent Donnellan

Erasmus University Rotterdam, University of California, Irvine, Michigan State University, Texas A&M University

Word count: 12,363

Corresponding author:

Rolf A. Zwaan
Department of Psychology
Erasmus University Rotterdam
zwaan@fsw.eur.nl

Abstract

Many philosophers of science and methodologists have argued that the ability to repeat studies and obtain similar results is an essential component of science. A finding is elevated from single observation to scientific evidence when the procedures that were used to obtain it can be reproduced and the finding itself can be replicated. Recent replication attempts show that some high profile results---most notably in psychology, but in many other disciplines as well---cannot be replicated consistently. These replication attempts have generated a considerable amount of controversy and the issue of whether direct replications have value has, in particular, proven to be contentious. However, much of this discussion has occurred in published commentaries and social media outlets, resulting in a fragmented discourse. To address the need for an integrative summary, we review various types of replication studies

and then discuss the most commonly voiced concerns about direct replication. We provide detailed responses to these concerns and consider different statistical ways to evaluate replications. We conclude there are no theoretical or statistical obstacles to making direct replication a routine aspect of psychological science.

Keywords: [Reproducibility, Replication, Psychological Research, Research Programs]

> "The proof established by the test must have a specific form, namely, repeatability. The issue of the experiment must be a statement of the hypothesis, the conditions of test, and the results, in such form that another experimenter, from the description alone, may be able to repeat the experiment. Nothing is accepted as proof, in psychology or in any other science, which does not conform to this requirement." (Dunlap, 1926).

The ability to systematically replicate research findings is a fundamental feature of the scientific process. Indeed, the idea that observations can be recreated and verified by independent sources is usually seen as a bright line of demarcation that separates science from non-science (Dunlap, 1926). A defining feature of science is that researchers do not merely accept claims without being able to critically evaluate the evidence for them (e.g., Lupia & Elman, 2014). Independent replication of research findings is an essential step in this evaluation process, and thus, replication studies should play a central role in science and in efforts to improve scientific practices.

This perspective on replication is succinctly encapsulated in the opening quote from Knight Dunlap. The value of replication as a normal feature of psychology, however, has proven surprisingly controversial in recent years. Debates exist over terminology used to describe replication studies, the statistical evaluation of replication attempts, the informational value of different types of replication studies, the interpretation of replication results, and the relative importance of within-lab versus independent replication attempts. Some of the most active discussions surrounding these issues have occurred in the context of specific replication attempts, and the exchanges often appear in relatively informal outlets such as blog posts and on social media. The objective of the current review is to advance our view of the value of replications and to synthesize many of the recent discussions about replication to provide a foundation for future replication efforts. Ultimately, we hope that this discussion will make

replication studies a more regular and integral part of research, a shift that could potentially increase confidence in the veracity of findings. Although debate about replication have recently occurred in the context of a recent "crisis of confidence" in psychology (Pashler & Wagenmakers, 2012), we aim to make this discussion broadly applicable to other disciplines that struggle with similar issues.

Definitions and Background

Replication is viewed by many as essential to scientific discovery. Popper (1959/2002) noted that an "effect" that has been found once but cannot be reproduced does not qualify as a scientific discovery; it is merely "chimeric." In fact, he notes, "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" (pp. 23-24) In a similar vein, Dunlap, (1926, p. 346) stated: "[P]roof is not begun until the conditions of the experiment, as well as the results, are so accurately described that another person, from the description alone, can repeat the experiment."

There are two important aspects to these insights that inform scientific thinking. First, a finding needs to be repeatable to count as a scientific discovery. Second, research needs to be reported in such a manner that others can reproduce the procedures. Thus, a scientific discovery requires both a consistent effect and a comprehensive description of the procedure used to produce that result in the first place. Neither of these points means that all replication attempts should be expected to succeed (i.e., a single failed replication does not necessarily mean that the original effect is a false positive) or that there are no specific skills required to conduct replications. Effects in psychology are often probabilistic and expertise is required to understand and follow comprehensive descriptions of procedures. Nonetheless, replicability is,

in principle, an essential criterion for the effect to be accepted as part of the scientific literature (Dunlap, 1926; Hüffmeier, Mazei, & Schultze 2016; Lebel, Berger, Campbell, & Loving, 2017 Lykken, 1968) and replication studies therefore evaluate the robustness of scientific findings (Schmidt, 2009).

Replications also play an important role in the falsification of hypotheses. If a finding that was initially presented as support for a theory cannot be reliably reproduced using the comprehensive set of instructions for duplicating the original procedure, then the specific prediction that motivated the original research question has been falsified (Popper, 1959/2002), at least in a narrow sense.This does not necessarily lead to a wholesale falsification of the theory from which that prediction was derived (Lakatos, 1970; Meehl, 1990). Under Lakatos' notion of *sophisticated falsificationism,* an auxiliary hypothesis can be formulated, which enables the expanded theory to accommodate the troublesome result. If more falsifications arise, however, and even more auxiliary hypotheses must be formulated to account for the unsupported predictions, problems begin to accrue for a theory. This *strategic retreat* (Meehl, 1990) can cause a research program to become *degenerative*:

> "As more and more ad hockery piles up in the program, the psychological threshold (which will show individual differences from one scientist to another) for grave scepticism as to the hard core will be increasingly often passed, inducing an increasing number of able intellects to become suspicious about the hard core and to start thinking about a radically new theory." (Meehl, 1990: 112)

If, on the other hand, the auxiliary hypotheses are empirically successful, the program acquires greater explanatory power and is deemed *progressive.* Thus, replications are an instrument for distinguishing progressive from degenerative research programs.

Issues with Replicability

Concerns about the replicability of scientific findings have arisen in a number of fields, including psychology (Open Science Collaboration, 2015), genetics (NCI-NHGRI working group on replication in association studies, 2007; Hewitt, 2012), cancer research (Errington, Iorns, Gunn, Tan, Lomax, & Nosek, 2014), neuroscience (Button et al., 2013), medicine (Ioannidis, 2005), and economics (Camerer et al., 2016). Thus, although vigorous debates about these issues have occurred within psychology (hence our focus), concerns about the replicability of findings exist in many disciplines. Perhaps disciplines that have not struggled with this issue (at least minimally) have simply not yet systematically examined the replicability of their findings. Indeed, a good portion of psychology likely had ignored this question before the recent crisis of confidence.

Problems with replicability can emerge for a variety of reasons. For example, *publication bias*, the process by which research findings are selected based on the extent to which they provide support for a hypothesis (as opposed to failing to find support), can on its own lead to high rates of false positives (Greenwald, 1975; Ioannides, 2005; Kühberger, Fritz, & Scherndl, 2014; Smart, 1964; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Yet there are additional forces and practices that can increase the rates of false positives. For instance, there is a growing body of meta-scientific research showing the effects of excessive *researcher degrees of freedom* (John, Lowenstein, & Prelac, 2012; Simmons, Nelson, & Simonsohn, 2011) or latitude in the way research is conducted, analyzed, and reported. If researchers experience pressure to publish statistically significant findings, then the the existence of researcher degrees of freedom allows investigators to try multiple analytic options until they find a combination that provides a significant result. Importantly, confirmation bias alone can convince investigators that the procedures that led to this significant result were the “best” or “most justifiable” approach in the first place. Thus, capitalizing on researcher degrees of freedom need not feel like an intentional decision to try multiple options until a set of procedures “works” (Gelman & Loken,

2014). It can seem like a reasonable approach for extracting the most information from a dataset that was difficult to collect.

The research practices that allow for this flexibility vary in terms of their severity and in the amount of consensus that exists about their permissibility (John, Lowenstein, & Prelec, 2012). For example, researchers have sometimes omitted failed experiments that do not support the focal hypothesis, and there are disagreements about the severity and acceptability of this practice. Researchers also form hypotheses after having examined the data, a practice called HARKing (Hypothesizing After the Results are Known; Kerr, 1998). When HARKing is undisclosed to readers of a paper, it might strike some researchers as deceptive. However, this strategy was once presented as the hallmark of sophisticated psychological writing (Bem, 2004): "If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything—interesting. No, this is not immoral."

Researcher degrees of freedom and publication bias that favors statistically significant results have produced overestimations of effect sizes in the literature, given that the studies with nonsignificant effects and smaller effect sizes have been relegated to the file drawer (Rosenthal, 1979). If a replication study is carried out in a field characterized by such practices, then it is likely to obtain a smaller effect size, often so small as to not be distinguishable from zero when using sample sizes typical of the literature. Replication thus has an important role in providing more accurate estimates of effect sizes. Even if all questionable research practices were eliminated, replication would remain essential to science because effect sizes are not only impacted by questionable research practices but also by sampling error. Sometimes researchers obtain a statistically significant result purely by chance; such a fluke does not reflect a real discovery. Effect size estimates can be inflated by sampling error alone. Thus, at a fairly

abstract level, there are good reasons why replication is necessary in science. Nevertheless, there are ongoing debates about nearly all aspects of replications, from terminology to purpose to their inherent value.

Types of Replication Studies

Replication studies serve multiple purposes and these objectives dictate how a replication study is designed and interpreted. Schmidt (2009) identified five functions: (1) to address sampling error (i.e., false-positive detection); (2) to control for artifacts; (3) to address researcher fraud; (4) to test generalizations to different populations; and (5) to test the same hypothesis of a previous study using a different procedure. A single replication study cannot simultaneously fulfill all five of these functions.

Given these different functions, a number of typologies have been offered for classifying replication studies (e.g., Lykken, 1968; Schmidt, 2009; Hüffmeier, Mazei, & Schultze, 2016; Lebel et al., 2017; Schmidt & Oh, in 2016). For example, Hüffmeier et al. (2016) provide a five-category typology whereas Schmidt and Oh (2016) delineate three types and Lebel et al. (2017) provide a replication taxonomy. Drawing on Schmidt (2009) and others (e.g., Crandall & Sherman, 2016; Makel et al., 2012), we focus on a distinction between direct and conceptual replication studies in this article, as this distinction has proven most controversial.

A number of definitions have been offered for direct and conceptual replications. A workable definition of direct replication is a study that attempts to recreate the critical elements (e.g., samples, procedures, and measures) of an original study where those elements are understood according to "a theoretical commitment based on the current understanding of the phenomenon under study, reflecting current beliefs about what is needed to produce a finding" (Nosek & Errington, 2017). Under this definition, a direct replication does not have to duplicate all aspects of an original study. Rather it must only duplicate those elements that are believed

necessary for producing the original effect. For example, if there is no theoretical reason to assume that an effect that was produced with a sample of college students in Michigan will not produce a similar effect in Florida, or in the UK, or Japan, for that matter, then a replication carried out with these samples would be considered direct.

If, however, there are theoretical reasons to assume a difference between samples, for example a hypothesis about the moderating effects of geographical or cultural differences, then the replication attempt would not be considered direct. It would be considered conceptual because the experiment is designed to test whether an effect extends to a different population given theoretical reasons to assume it will be either significantly weaker or stronger in different groups.

In some cases, a direct replication is necessarily different from the original experiment, although these difference are usually superficial. For example, it may be necessary to adapt stimulus materials for historical reasons. Current events questions asked in the 1980s (e.g., “Who is the President of the United States?”, “What is the capital of Germany?”, “What is the currency used in Italy?”) all have different answers in 2017. A direct replication of study about the impact of distraction on tests of current events would use updated questions, assuming there are no theoretical reasons to expect a different performance on the part of participants in this kind of study.

As noted, a conceptual replication is study where there are changes to the original procedures that might make a difference with regard to the observed effect size. Conceptual replications span a range from having one theoretically meaningful change with regard to the original experiment (e.g., a different dependent measure) to having multiple changes (Lebel et al., 2017). On this view, the notion “conceptual replication” is a bit of a misnomer. What such a study in effect does, is test an extension of the theory to a new context (as there are different auxiliary hypotheses involved in the operationalization of the key variables). It might therefore be more informative to speak of “extensions” rather than of “conceptual replications.”

Nonetheless, to connect to the previous literature we will retain the use of the term “conceptual replication” in this article.

Conceptual replications do not serve the same purposes as direct replications. Therefore, we encourage researchers to adopt different terminology when describing conceptual replications in the future. This will yield a clearer distinction between studies that use the same procedures as the original studies as opposed to studies that use different procedures. The goal of a direct replication is to determine whether a specific way of testing a theoretical idea will produce a similar result in a subsequent attempt. The objective of a conceptual replication is broader—the point is to test the same theoretical idea in a novel way. Conceptual replications evaluate the robustness of a theoretical claim to alternative research designs, operational definitions, and samples. Direct replications are useful for reducing false positives (i.e., claims that a specific effect exists when it was originally a chance occurrence or fluke), whereas conceptual replications provide information about the generalizability of inferences across different ways of operationally defining the constructs and across different populations. Using *alternative test* in place of the term conceptual replication might help clear up confusion in the literature that occurs when researchers disagree as to whether or not an effect has been replicated.

A few additional clarifications about our definitions are warranted. The term *exact replication* is occasionally used as a synonym of direct replication. The chief objection to the use of “exact” is that it implies a level of precision that is impossible to achieve in psychology. In psychological experiments it is impossible to use the same subjects in a replication and expect them to be in exactly the mental state as they were in the first experiment. For one, the mere fact of having participated creates awareness and the possibility of internal changes in participants, although some cognitive-psychological findings prove remarkably robust, even when nonnaïve participants are used (Zwaan et al., 2017). In addition, as we’ve noted before, historical changes over time may lead to differences in expected results. For reasons such as

this, Schmidt (2009, p. 92) noted that in the social sciences “There is no such thing as an exact replication.” The defining aspect of direct replication is the attempt to recreate the essential elements of the original study rather than all of the elements.

Controversy over the use of the term “exact replication” reflects the reality that debates exist about when a replication study deviates from the procedures of a previous study so much that it becomes a conceptual replication. Although this may seem like a semantic issue, it is critically important for the appropriate interpretation of a failure to replicate. If a direct replication fails to obtain the same result as the original study, researchers may question whether the initial result was a false positive (and this will be especially true after multiple failed direct replications) or whether there is a misunderstanding about the understanding of the essential features required to produce an effect. This will likely prompt a more critical evaluation of the similarities between the original study and the replication.

Any evaluation of the degree of similarity between an original study and a replication might seem to have subjective elements. However, Nosek and Errington’s (2017) notion of “theoretical commitment” helps solve this problem, as researchers should be able to agree on what the critical elements of an experiment are to produce an effect. Nevertheless, evaluations of whether a study is a direct or conceptual might sometimes change, as more evidence about the nature of the underlying phenomenon is obtained. For example, researchers may conduct a simple study that they believe should emerge in any sample of U.S. college students. An independent researcher may then attempt a direct replication of the original effect in a sample of students from a different university. If that second researcher fails to replicate the original result, the direct nature of the replication may reduce confidence in the effect. However, the failed replication may lead the original researcher to formulate new hypotheses about why the specific university population might matter (e.g., regional differences in psychological characteristics might attenuate effects). In other words, the understanding of which factors matter with regard to producing an effect has changed. Subsequent studies that show that the effect reliably

emerges at some universities but not at others would change this characteristic of the study from an inconsequential one to a consequential one, and thus, studies that used different populations would, from that point forward, be considered conceptual replications. At the theoretical level, the successful auxiliary hypothesis has enhanced the explanatory power of the theory.

We emphasize that there is no reason to accept all *post hoc* discussions of potential moderators as compelling reasons to disqualify a study from being considered a direct replication. Instead, evidence that what was initially considered to be an inconsequential factor (e.g., region of a country) has reliable effects on the results is required. Researchers who conduct original studies can facilitate replications and reduce disagreement about hidden moderator explanations by following Lykken's (1968, p. 155) admonishment that they should "accept more responsibility for specifying what they believe to be the minimum essential conditions and control for producing their results" (p. 155). Disagreements are likely to be minimized if original authors spend some time articulating theoretically-grounded boundary conditions for particular findings (Simons, Shoda, Lindsay, in press).

A controversial issue surrounding the definition of direct and conceptual replications concerns who actually plans and conducts the replication research; some distinguish replications conducted by the original authors from those conducted by an independent group (e.g., Hüffmeier et al., 2016). The rationale for these distinctions often rest on concerns about expertise or unidentified moderators that may vary across research laboratories.The basic idea is that some people have the expertise to carry out the replication (usually the original authors) whereas others do not have such skills (usually researchers who fail to replicate an effect). Likewise, original authors are often working in similar settings to the original study so many potential moderators are held constant. Dunlap (1926, p. 346) puts the issue in these terms:

"The importance of repetition as a part of proof is, then,

due to the necessity, in general, of certifying that the descriptions

> of conditions and results are accurate to the requisite degree. When another experimenter, setting up the conditions from the description of the first experimenter, obtains results which he describes in the same way as that in which the first experimenter describes his, the presumption of accuracy is enormously increased. Repetition of the experiment by the same experimenter does not have as great demonstrative value because of the possibility that the experimenter in the second experiment may not be actually following his own description, but may be following his first procedure, and therefore may vary from the description in the same way."

Our definitions do not make such distinctions, as they do not directly address scientifically relevant features of the research. As we explain in more detail below, original researchers can address this issue by clearly defining the procedures of a study and identifying the special skills required to duplicate the procedures.

In summary, we use *direct replication* to refer to studies intended to evaluate the ability of a particular method to produce the same results upon repetition and *conceptual replication* to refer to studies designed to test the same theoretical idea using an intentionally different method than previous studies. In the next sections, we address the frequent concerns that have been raised about direct replications.

Concerns About Replication

The interpretation of specific replication studies has produced considerable disagreement and controversy. For instance, consider the 2015 paper by the Open Science

Collaboration, which presented the results of a large-scale attempt to replicate approximately 100 studies from top journals in psychology (Open Science Collaboration, 2015). A headline finding from this project was that only 36% of the attempted replications were "successful," in the sense that a significant effect in the same direction as the original was found. The publication of this report was met with a wide range of responses, including some focused on the fidelity of the replication studies, the criteria used to determine whether a replication was successful, the value of such a large-scale investment of resources, and so on (e.g., Anderson et al., 2016; Etz & Vandekerckhove, 2016; Gilbert, King, Pettigrew, 2016; Kunert, 2016; Maxwell, Lau, & Howard, 2015; Morey & Lakens, 2016; van Aert & van Assen, 2017; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). These responses (which continue to be published at the time of our writing this paper) focus on a broad range of challenges and objections to the value of replication studies as a whole. We now consider many of the most frequent concerns that are raised about replications.

Concern I: Context Is Too Variable

Perhaps the most commonly voiced concern about direct replications is that the conditions under which an effect was initially observed may no longer hold when a replication attempt is performed (Barsalou, 2016; Cesario, 2014; Coyne, 2016). This ever-present possibility of a change in context, it is argued, renders failures to replicate uninformative, especially early on in the life cycle of a finding. The factors that contribute to the ability to independently reproduce an effect may be historical and/or geographical in nature (Cesario, 2014) or be the result of unknown conditions, including such seemingly irrelevant features as the lighting in the lab or whether or not the experimenter has a beard (Coyne, 2016). As Cesario puts it: "replication failures at this stage will necessarily be ambiguous because we cannot be

sure that features that appear incidental to the researcher are not actually integral to obtaining the original effect."

Barsalou (2016) offers the most elaborate theoretical account of contextual variability, focusing on an area where context effects may be particularly salient: studies in the area of social priming research. Priming research, in general, focuses on the extent to which exposure to a specific stimulus can affect memory for, perception of, or behavior in response to a subsequently experienced stimulus. Social priming research, in particular, focuses on a wide variety of mundane and subtle social stimuli that can affect respondents in sometimes powerful ways. Traditionally, social psychological research on social priming has emphasized the surprising ways that exposure to seemingly inconsequential environmental cues can lead to substantial changes in behavior. A quintessential example is the notion that presenting participants with images of money will increase (or prime) certain kinds of political views given associations between the two in the minds of participants (see e.g., Rohrer, Pashler, & Harris, 2015).

Central to Barsalou's account of contextual variability is the notion of situated conceptualization: people perceive and interpret situations that are experienced and store them as multi-modal (e.g., visual, auditory, olfactory) mental representations in long-term memory. If a type of situation occurs repeatedly, a category of exemplars of this type is formed. The more features a conceptualization in long-term memory shares with a newly experienced situation, the more likely that conceptualization is to become activated. Once activated, it will generate pattern-completion inferences that will sometimes match and sometimes mismatch features of the current situation.

For example, the (repeated) experience of visiting a coffeehouse leads to the formation of a situated conceptualization of experience. When a new coffeehouse is visited, this conceptualization is likely to become activated. Once activated, this conceptualization will allow the person to generate predictions about what to expect during a visit to a different coffeehouse,

e.g., the smell and taste of coffee, the sight of people working on laptops, the murmur of conversations, the noise of espresso machines, and the cerebral atmosphere. Inference generation is an involuntary mechanism. Some of the predictions will hold in the new environment whereas others will not (e.g., some patrons brought their children rather than laptops).

This configuration gives rise to two mechanisms. First, any feature (e.g., the smell of coffee) of a new situation can activate a situated conceptualization. Second, any element of a situated conceptualization can be inferred as a pattern-completion inference. How do these ideas relate to the reproducibility of social-priming and other context-sensitive experiments? Barsalou argues (p. 9) that "simple direct pathways from primes to primed response rarely, if ever, exist. Instead, these pathways often appear to be modulated by a host of situational variables," given that an activated situated conceptualization colors the perception of and action in the current situation.

According to Barsalou, three factors are necessary to obtain robust priming effects in social psychology: (1) participants need to have had similar situational experiences with the prime and primed response so that they have situated conceptualizations of them in memory. (2) There should be a strong overlap between the situated conceptualizations in memory and the current experimental situation. (3) The prime should not be part of other situated conceptualizations that lead to other, better matching responses. Barsalou argues that, given that people have diverse situational experiences, often not all three of these conditions are met, which will then result in diverse responses to primes. For this reason, Barsalou proposes to abandon the notion of social priming to focus on specific mechanisms.

In short, the contextual argument posits that direct replications of priming effects in social psychology (as well as a host of other effects) will not be scientifically useful or successful because the intricate network of factors contributing to certain effects is largely unknown and that many of these factors are often exquisitely specific to a particular population with shared

experiences. Proponents suggest it is too difficult to specify all of these contextual factors; and even if they could be articulated, it is extremely difficult for independent investigators to recreate these conditions with precision. As a result, it is never possible to determine whether a "failed" replication is due to the fact that the original demonstration was a false-positive, or to whether the context has changed sufficiently to wipe out that effect.

Response

Changes in context can and should be considered as a possible explanation for why a replication study failed to obtain the same results as in the original. There are very few effects in psychology where context could never matter; and indeed, if context is taken to include scientific expertise, then there are few effects in science where such factors would never play a role in the outcome. In addition, as noted above, it is impossible to conduct exact replications; some contextual features—even if very minor—will always vary from one study to the next. So even the most fervent advocate of direct replications would not deny that context matters in psychological research.

Nevertheless, the *post hoc* reliance on context sensitivity as an explanation for all failed replication attempts is problematic for science. A tacit assumption behind the contextual sensitivity argument is that the original study is a flawless, expertly performed piece of research that reported a true effect. The onus is then on the replicator to create an exact copy of the original context to produce the same exact result (i.e., the replicator must conduct an exact replication). The fact that contextual factors inevitably vary from study to study means that *post hoc*, context-based explanations are always possible to generate, regardless of the theory being tested, the quality of the original study, or the expertise and effort made by researchers to conduct a high-fidelity replication of an original effect. Accordingly, the reliance on context-sensitivity as a *post hoc* explanation, without a commitment to collect new empirical evidence that tests this new idea, renders the original theory unfalsifiable. Such reasoning is

representative of a degenerative research program: the auxiliary hypotheses that are put forth do not enhance the theory's explanatory power (Lakatos, 1970).

An uncritical acceptance of *post hoc* context-based explanations of failed replications ignores the possibility that false positives (even those based solely on sampling error) ever exist and seems to irrationally privilege the chronological order of studies over the objective characteristics of those studies when evaluating claims about quality and scientific rigor. It is possible to simultaneously acknowledge the importance of context and to take seriously the informational value of well-run replication studies. For instance, according to the definitions provided above, direct replications are designed to duplicate the *critical features* of a study, while inevitably allowing for inconsequential features to vary somewhat. If there are contextual factors that could play an important role in the ability to find the effect (such as the specific population that was sampled, the specific time of year in which the study was run, or even the specific time period in which the result was obtained), it would be reasonable to expect authors to specify that these variables are critical for producing the effect in the original report as part of the detailed description of the procedures of that study. For instance, in justifying the specific methodological choices for a given study, authors could approach this justification by considering how they would create a template for producing (and reproducing) the original effect.

Alternatively, if a failed replication brings to light some factor that could potentially affect the result and that differed between the original study and the replication, conducting further investigations into the impact that this factor has on the result is a reasonable scientific endeavor. In short, the *post hoc* consideration of differences in features should lead to new testable hypotheses rather than blanket dismissals of the replication result. In terms of Lakatos' (1970) sophisticated falsificationism, the original theory was unable to explain the new nonsignificant finding and an auxiliary hypothesis (or hypotheses) has to be invoked to accommodate the new finding. The auxiliary hypothesis then predicts the original finding when

the experiment has contextual feature O (from the original study) but not when it has contextual feature R (from the replication). If this auxiliary hypothesis is supported, the augmented theory is not falsified. If the auxiliary hypothesis is not supported, perhaps a new auxiliary hypothesis can be generated. If this hypothesis is also not supported, the research program might run the risk of becoming degenerative, falling into a fruitless cycle of constantly invoking auxiliary hypotheses that fail to garner support.

It is sometimes argued that a detailed description for replicating an original result might be impossible in some domains (such claims have been made about areas such as social psychology or infant research; Barsalou, 2016; Coyne, 2016), where the combination of contextual factors and expertise that is needed to produce a specific effect is complex, and perhaps even unknowable. If these sorts of claims are true, however, then this would raise serious doubts about the validity, informational value, and contribution to a cumulative body of knowledge of the original study. There are at least two reasons why such arguments are scientifically untenable.

First, if the precise combination of factors that led to a scientific result is unknowable even to the original author, then it is not clear how the original authors could have successfully predicted their effect to emerge in the first place. For instance, imagine that a hypothetical priming effect in social psychology can only emerge when: (a) a sample of college students has a specific average level of political conservatism, (b) the experiment took place at a particular time in the semester, (c) the experiment was conducted at a particular time of the day, (d) the experimenter who first meets the participant dresses in a lab coat to emphasize the serious, scientific nature of the study, and (e) experimental stimuli are presented on a computer as opposed to on paper. Let us also stipulate that the original theory does not clearly predict that any of these factors should matter for the effect to emerge, so the original authors did not explicitly consider the specific sample they recruited (they were recruited from the population that was available), the time of year when the study was run (they began data collection when

IRB approval was obtained), the time of day when the study was conducted (the decision resulted from research assistant availability), the dress of the experimenter (the lab coat might have been standard procedure from other, unrelated studies), or the method of administration (computers may have simply been chosen to ensure blindness to condition).

If a replication was conducted by a separate group of researchers, some of these idiosyncratic, seemingly irrelevant factors would change, resulting in the failure to find an effect. What cannot be explained, however, is how the original authors happened upon the exact set of conditions that led to the predicted result in the first place, in light of the impoverished nature of the underlying theory. It is no more likely for an original author to hit upon the exact combination of factors that "work" than it is for a replicator. Thus, the idea that certain phenomena are so susceptible to subtle contextual factors that no replication should be expected to succeed would also raise serious questions about how an original researcher could have predicted the outcome of an original study in light of all of the complexity.

A second reason why strong forms of the context sensitivity argument are scientifically problematic is that such an argument would prevent the accumulation of knowledge within a domain of study. *A priori* predictions are made precisely because the original researchers believe that they have enough knowledge about a phenomenon to be able to predict when and how that phenomenon will occur. If researchers do not know enough about a phenomenon to predict when it will and when it will not be replicated, it is not possible for subsequent research to build on this individual finding. If findings are so tenuous that replication results cannot be taken for granted, it is difficult if not impossible for new knowledge to build on the solid ground of previous work. Moreover, there is little reason to expect that findings that emerge from a non-cumulative perspective will have practical relevance given that results are highly contingent upon a complex mosaic of factors that will be present in a limited set of circumstances. Such a research program can be characterized as degenerative (Lakatos, 1970). It would be gravely mistaken to speculate about the applied value of such a research program in published papers.

An inability to specify the conditions needed to produce an effect is a serious impediment to scientific progress. The ability to specify a clear set of procedures that reliably elicit a predicted effect allows for independent verification and provides the foundation for practical applications and studies that extend the original result. For a discovery to be counted as scientific, it should be accompanied by a description of the procedure that lead to the discovery so that others can replicate it. Several authors have lamented the lack of procedural specificity in many psychology articles. They call for more detailed descriptions of experiments, such that the conditions under which an effect is expected to replicate are specified (Fabrigar & Wegener, 2016; Simons, Shoda, & Lindsay, in press). Likewise, it should be possible to specify the skills needed to conduct a particular study to produce a particular effect. It might be impossible to pre-specify all such conditions and required experimenter skills, but in cases where a replication attempt fails to obtain the original result, claims of context effects or limited skills of the experimenter should be proposed as testable hypotheses that can be followed up with future work. Until future studies can be conducted to test hidden moderator arguments, researchers should strive to ignore the chronological order of the original and replication studies when evaluating their belief in a phenomenon and rely more on the relatively quality of the two studies such as sample size and the existence of pre-registered analytic plans to constrain analytic flexibility.

On balance, contextual variability is not a serious problem for replication research. It is only a problem when the context is not sufficiently specified in the original findings so that the source of the reported effects cannot be identified. Only once the context is sufficiently specified are both direct replication and actual investigation of contextual variability possible. Preregistered multi-lab replication reports thus far have not provided strong evidence for variability across labs (Alogna et al., 2014; Eerland et al. 2016; Hagger et al., 2016; Wagenmakers et al., 2016).

Two strategies for solving the concerns outlined in this section are to (a) raise standards in reporting of experimental detail, such that original papers contain *replication recipes* (Brandt, et al., 2014; Dunlap, 1926; Popper, 1959/2002) and (b) find ways to encourage original authors to identify potential boundary conditions and caveats in the original paper (i.e., statements about the limits of generalization; Simons et al., in press).

Concern II: The Theoretical Value of Direct Replications is Limited

Several arguments against replication converge on a general claim that direct replications are unnecessary because they either have limited informational value (at best) or are misleading (at worse). Crandall and Sherman (2016, p. 95) argue that direct replications only help to "uphold or upend specific findings" which, in their view, makes direct replications uninformative and uninteresting from a theoretical perspective. For instance, difficulties reproducing a specific effect can only suggest a problem with a specific method used to test a theoretical idea. Likewise, a successful direct replication has little implication for theory because "[a] finding may be eminently replicable and yet constitute a poor test of a theory" (Stroebe & Strack, 2014). If the dependent measures of an original study are poorly chosen, a finding might replicate consistently, yet its replicability is problematic because it reinforces the wrong interpretation (Rotello, Heit, & Dubé, 2015). The concern is that the direct replications provided a false sense of certainty about the robustness of the underlying idea.

The utility of direct replications has also been challenged in fields that might be characterized by capitalizing on correlations between conceptually overlapping variables such as studies investigating depressive symptoms and self-reported negative affectivity (Coyne, 2016): "This entire literature has been characterized as a "big mush." Do we really need attempts to replicate these studies to demonstrate that they lack value? We could do with much

less of this research." Moreover, just as original studies can be unreliable, so can replications, which means that one can be skeptical about the value of any individual replication study (McElreath & Smaldino, 2015).

Response

One part of this concern reflects the fact that neither failed nor successful direct replication studies make novel contributions to theory. This argument rests on the idea that studies that intentionally test mediators, moderators, and boundary conditions all provide different bricks in a wall of evidence, whereas direct replications can only address specific bricks in that wall (Spellman, 2015). For many researchers, work that does not directly advance theory is not worth doing, especially when it is possible to simultaneously address concerns about reliability and validity with new conceptual replications that are designed to replicate and extend prior work. Part of this argument is that successful conceptual replications will only occur when the prior research identified a real effect. There is an implicit assumption that it is impossible to create a "wall" of empirical findings that support an underlying theory if most of the specific bricks in that wall were not already solid.

Unfortunately, there is increasing evidence that this seemingly reasonable assumption about the totality of evidence that emerges from a series of conceptual replications is wrong. The combined effects of researcher degrees of freedom, chance findings from small sample studies, and the existence of publication bias mean that it is possible to assemble a seemingly solid set of studies that appear to support an underlying theory, even though no single study from that set could survive a direct replication attempt. There are now a number of widely studied theories and effects that have been supported by dozens, if not hundreds of conceptual replications, that also appear to collapse when meta-analyses that are sensitive to publication bias are reported or systematic replications of critical findings are conducted (Cheung, Campbell, et al., 2016; Hagger, Chatzisarantis, et al., 2016; Shanks, Vadillo, et al., 2015;

Wagenmakers, Beek, et al., 2016). Those who argue that a large set of successful conceptual replications would not be possible in the absence of a real effects assume that publication bias and questionable research practices are not powerful enough to create a wall full of defective bricks. However, this is an empirical question that can be best answered with direct replications of foundational bricks in theoretical walls.

Moreover, in a direct replication of earlier work, the question of whether a particular method is an appropriate test of a hypothesis was previously answered in the affirmative. After all, the original study was published because its authors and the reviewers and editors who evaluated it endorsed the method as a reasonable test of the underlying theory. It is therefore not consistent to claim, after the fact, that the results should not be interpreted because the manipulation was not valid or the outcome variable was inappropriate.

It is important to contrast this strength of direct replication with the ambiguity that comes with failed conceptual replications. It is always possible to attribute a failed conceptual replication to the changes in procedures that were made. In other words, conceptual replications (at least those that are not preregistered) are biased against the null hypothesis (Pashler & Harris, 2012) because researchers might be tempted to discard an experiment that does not produce the expected effect on the basis that it was not a good operationalization of the hypothesis after all. Direct replications do not have this interpretational ambiguity.

Direct replications are not only important with regard to earlier work. They are also necessary if researchers want to further explore a finding that emerged in exploratory research, for example in a pilot study. In this case, the approach would normally be to make explicit the procedure that is likely to (re)produce the finding observed during the exploratory phase, preregister that procedure, and then run the experiment. In such cases one would not necessarily assume that the initial procedure was an appropriate test.

The argument that conceptual replications effectively serve the same purpose as direct replications but with additional benefits, is sometimes accompanied by the argument that a field

that is focused on direct replications simply cannot progress because it would make no new discoveries. There are two issues here. First, the strong form of the claim that direct replications make no new discoveries holds, if and only if, the original finding was a true positive. The repeated demonstration that a theoretically predicted effect is *not* empirically supported adds knowledge to the field; it is a discovery. It is only in hindsight that one can claim that direct replications fail to add knowledge. Likewise, research that leads to the identification of moderators and boundary conditions adds knowledge. Moreover, such a strong claim may not withstand critical scrutiny because even in the cases of a successful replication, there is additional knowledge gained by learning that a finding is replicable.

Second, it is not clear what the benefits of conceptual replications are without direct replication. A conceptual replication would have to be replicated directly before it could count as a scientific finding (see our Introduction). No one would argue that all the collective resources of a field should be spent determining whether past findings survive replication attempts. Instead, devoting some time to direct replications is an important goal for the field especially with concerns of the winner's curse (Button et al., 2013) and the effects of researcher degrees of freedom and publication bias. As mentioned earlier, direct and conceptual replications serve different purposes. Direct replications assess the robustness of a finding when using a specific set of procedures, whereas conceptual replications assess the validity of a construct or underlying theory. It only makes sense to first assess the reliability of a specific finding obtained with a particular method before venturing out into what might turn out to be a dead-end street by using a different method to test the same theoretical claim.

We noted earlier, but like to reiterate here, that direct replications play an important role at the theoretical level. An unsuccessful replication might prompt researchers to form an auxiliary hypothesis that explains the discrepancy between the results of the original study and those of the replication. After all, the direct replication was based on a theoretical understanding of the elements of the original experiment that were thought critical for producing the effect.

Apparently, this understanding was incomplete or incorrect. If the auxiliary hypothesis is supported, the theory is strengthened. If it is not supported, the theory is weakened. Either way, the direct replication has had an impact on the theory.

It is important to note that there are other procedures that other procedures can be used to accomplish at least some of the aims of direct replications. For example, preregistration can reduce or prevent researcher degrees of freedom, which can reduce the rates of false positives introduced into a literature. In preregistration, a researcher details the study design and analysis plan on a website, for example on the Open Science Framework or on Aspredicted.org, before the data are collected (Chambers, 2017, pp. 174-196). In addition, committing to *public* preregistration can at least help to reduce publication bias, as the number of failed attempts to test a hypothesis using a specific paradigm can be tracked. Replications are but one tool in the methodological toolbox. They may be especially important for evaluating important research from the past, before preregistration was normative; but the use of preregistration and especially registered reports may reduce the informational yield of direct replication as research practices evolve (but we doubt that such practices will ever eliminate the need for direct replications).

The above discussion focused primarily on the relative value of direct versus conceptual replications. However, another part of the concern is that direct replications might be problematic when the original study that is being replicated is itself not valid or theoretically important. This is a red herring. It goes without saying that scientific judgment should be used to assess the validity and importance of a study before deciding whether it is worth replicating, and many replicable effects provide only weak contributions (if any) to theory. To be sure, one can argue whether the resources that have been spent on massive-scale systematic replication attempts would have been better spent targeting a different set of studies (or doing original research; Finkel, Eastwick, & Reis, 2015). However, at least in psychology, at this moment of reflection on the practices in the field, explicit tests of the replicability of individual findings—regardless of how the specific findings that are replicated are chosen—have important

informational and rhetorical value that go beyond the impact that arguments about researcher degrees of freedom or publication bias can make. Moreover, there will probably be a fair bit of disagreement among researchers as to when an original study is theoretically important as opposed to silly or trivial.

Concern III: Direct Replications Are Not Feasible in Certain Domains

It is sometimes argued that conducting replication studies may not be desirable—or even possible—merely due to practical concerns. For example, replications may not be feasible in certain domains, such as large-scale observational and clinical-epidemiological studies (Coyne, 2016). Alternatively, certain studies may capitalize on extremely rare events like the occurrence of a natural disaster or an astronomical event, and replicating studies that test the effects of these events is simply impossible. Thus, if the ability to replicate a finding is taken as an essential criterion by which we judge whether a finding or program of research is "scientific," then the application of this criterion would exclude a great deal of research from consideration. This might create a caste system whereby some topics are privileged as more scientific and rigorous than others.

A related concern is that replication studies are more feasible and thus more common in areas where studies are easier to conduct (e.g., studies that use college student participants to advance knowledge in cognitive and social psychology). This means that those researchers working in the easy-to-replicate domains are more subject to the reputational concerns that may arise when their studies fail to replicate (see below for an explicit discussion of these reputational issues). More importantly, if studies that vary in difficulty also vary in rates of replicability (e.g., if studies that were easier to conduct had lower rates of replication than

studies that required more resources), then systematic efforts to investigate the replicability of findings in the field would lead to biased estimates of those rates.

Response

There are practical limitations that impact all studies including direct replications. For some specific studies---and maybe even for entire research areas---replication studies may be difficult or impossible. This may prevent direct replication studies from becoming a commonplace component of the research process in those domains. However, concerns about feasibility are orthogonal to the overarching value of direct replications for advancing scientific knowledge. The fact that replication studies are not always possible does not undermine their value when they can be conducted.

It is also important to note that even for those studies where the research community would agree that replication would be difficult or impossible, the initial concerns that motivate a focus on direct replication studies (such as researcher degrees of freedom and publication bias) still hold. Thus, researchers who work in areas where replication is difficult should be especially alert to such concerns and make concerted efforts to avoid the problems that result. Large scale developmental studies that follow participants for 30 and 40 years are one example as is research with difficult-to-study populations such as infants, prisoners, or individuals with clinical disorders. Researchers in such areas would benefit from preregistering their hypotheses, designs, and analysis plans, to protect themselves from concerns about researcher degrees of freedom and the use of questionable research practices. They can also blind the analysis, or set aside a certain proportion of the data for a confirmatory test. At the very least, discussion sections from papers that describe these results can be be appropriately calibrated to the strength of the evidence.

A related, but distinct concern is that because replication is easier in some domains than others, any costs of doing replication studies will disproportionately be borne by researchers in

those areas. For instance, if there are reputational costs to having one's work subject to replication attempts, then those who conduct easy-to-replicate research will be most affected. Alternatively, if a subfield of research includes easy-to-replicate studies and more difficult studies to conduct, and if the easy-to-replicate studies are of lower quality (and hence, less likely to replicate), then one may get a biased view of the quality of work in that area, when only attempted replications of easy studies are conducted.

Although occasional failures to replicate should not have any bearing on scientific reputation (an issue we return to in more detail below), the very fact that someone conducts research that is easy to replicate in the first place provides a simple solution to this potential problem. If a study is so easy to conduct that it is likely to attract replication attempts by outside researchers, then it would be worthwhile for the original author to invest some time in conducting within-lab, pre-registered, direct replications as part of the original publication. In many cases, high-profile direct replications have focused on single studies (that were often conducted with relatively small samples) that had not previously been subjected to direct replication attempts. If these replication studies are preregistered and conducted with large samples, a subsequent failure to find an effect can lead to strong concerns about the reliability of the original finding. If, however, the original finding already had a pre-registered, high-powered direct replication included as a part of the original publication, then the effect of the new failed replication on people's beliefs is lessened. Thus, concerns about "easy" studies being the target for replication attempts cut both ways—the ease with which these studies can be conducted should allow original authors to provide even stronger evidence in their initial demonstrations.

In regard to the concern that easy-to-replicate studies are not a representative sample of the studies in a field (and thus, attempts to replicate them may provide a misleading picture of the replication rate for that field), it should be noted that most replication studies are not conducted with the goal providing a precise estimate of the replication rate within a field.

Instead, the goal of many such studies is to test the robustness of a particular effect. In recent history, more systematic attempts to replicate large sets of studies have been conducted. Even in these studies, however, a primary aim is to evaluate whether the methodological practices that are in current use can result in the publication of studies that have low likelihood of replicating. One clear interpretation of the various systematic efforts that have been conducted so far is that this outcome is certainly possible. The fact that the studies selected for inclusion are not representative means that we cannot draw conclusions about the average replication rate, but the inclusion of seemingly many unreplicable studies in the published literature is still cause for concern.

It is evident that pragmatic concerns and availability of resources must be considered when evaluating the potential for replication studies. However, one might anticipate that to the extent that direct replication becomes a more routine aspect of psychological science, more resources will be available to conduct such studies. If the field demands evidence of replicability, then researchers will invest resources in conducting direct replications of studies. Ideally, as scientific norms change, even funders would be more willing to support research that tackles the challenges that have been identified, including research on replication attempts. For example, in 2016, the Netherlands Organisation for Scientific Research (NWO) launched a program to fund replication studies. As this change occurs, it may be possible to conduct replications with challenging designs such as longitudinal studies, studies based on specialized populations and harder to sample populations.

Concern IV: Replications are a Distraction

Many of the challenges addressed thus far come from the view that there is, in fact, not really a replicability problem in psychology or in science more broadly. A fourth concern, conversely, emanates from the view that the problems that exist in the field may be so severe

that systematic attempts to replicate studies that currently exist will be a waste of time and may even distract from bigger problems that are facing psychology (Coyne, 2016). For instance, Schmidt and Oh (2016) noted that "[o]ur position is that the current obsession with replication is a red herring, distracting attention from the real threats to the validity of cumulative knowledge in the behavioral sciences."

A related argument is that the primary problem in the accumulation of scientific knowledge is the existence of publication bias. According to this view, failed replications—whether direct or conceptual—do exist but are not making it into the literature. Once the systematic omission of these studies is addressed, meta-analyses will no longer be compromised and will then provide an efficient means to identifying the most reliable findings in the field. Similar arguments can be made about any additional strategy for improving psychological science, including an increased emphasis on preregistration or the reduction of questionable research practices. Again, the idea here is that even if replication studies tell us something useful, there are more efficient strategies for improving the field that have fewer negative consequences.

Response

As mentioned earlier, replication studies are one strategy among a broader a set of strategies that can be implemented simultaneously to improve the field. However, direct replication attempts have some unique benefits that should earn them a central role in future attempts at building a cumulative psychological science.

First, there is a certain rhetorical value to a replication study, whether failed or successful. The idea of replication is simple: if a finding is robust, independent groups of scientists should be able to obtain it. This idea is taught in most introductory classes in psychology and is foundational in science more broadly. And not surprisingly, when large sets of important studies—including studies whose results had previously been assumed to be robust—

fail to replicate, people outside of the field take notice. For those who believe that existing methodological practices could be much better, demonstrating these concerns through systematic replication attempts provides a compelling illustration. Such efforts have been a major motivation for change and impetus for the increase in resources that have been targeted towards improving the field.

It is clear that thus far, failures to replicate past research findings have received the most attention. However, large-scale successful replications also have rhetorical power, showing that the field is capable of producing robust findings on which future work can build (e.g., Alogna et al., 2014; Zwaan et al., 2017), and such results will likely become more common in the near future. Some have raised concern that with increased attention to replication studies, only failures to replicate are surprising and newsworthy enough to warrant publication, a phenomenon that would provide a misleading picture of the replicability of results in the field. However, the use of registered replication reports furthermore assuages the concern that only negative replications are incentivized. These reports are provisionally accepted for publication before any data are collected, and thus, any bias for or against successful replications is eliminated.

A second component of the argument that replication studies are a waste of time is the assumption that agreement exists that most research in the field is of poor quality, and thus not worth replicating. This assumption is not warranted. Instead, systematic attempts at replication—at least in the short term—are a way of testing whether the field is doing well or not. Indeed, a broader point is that there is debate about the extent of the problems that face psychology or other fields that have struggled with concerns about replicability such as the impact of publication bias (and what to do about it; e.g., Cook et al., 1993; Ferguson & Brannick, 2012; Ferguson & Heeney, 2012; Franco, Malhotra, Simonovits, 2014; Kühberger, Fritz, & Scherndl, 2014; Rothstein & Bushman, 2012). or the prevalence and severity of questionable research practices (Fiedler & Schwarz, 2015; John, Lowenstein, & Prelac, 2012; Simmons,

Nelson, & Simonton, 2011). Replication studies, in concert with alternative approaches to improve methodological practices allow for empirical tests of their impact. If preregistration truly does improve the quality of psychological research, then preregistered studies should be more replicable. If methods for detecting publication bias work, then replication attempts of effects that publication-bias-sensitive meta-analyses suggest are robust should be more successful than attempts of effects that seem to stem from a biased literature. In short, replication studies provide a simple, easily understandable metric by which we can evaluate the extent of the problem and the degree to which various solutions really work.

Coyne (2016) and others correctly argue that it would be wasteful to perform direct replications of research with highly obvious flaws. However, it is unclear how easy it is to judge the obviousness of flaws in the absence of evidence about replicability. Moreover, the claim that replications distract from bigger problems is perhaps based on the misconception that replication is being proposed as a panacea for all of the problems facing psychological science, which we have addressed earlier. It is just one element of the toolbox of methodological reform.

Concern V: Replications Affect Reputations

Debates about the value of replication studies often focus on the scientific value of replication. However, some debates concern the reputational effects of replication studies. These extra-scientific issues are relevant, both for those whose work is replicated and those who are doing the replications.

Replication studies—and especially failed replications—may have reputational costs for the authors of the original studies. At first blush, this may seem surprising. Presumably, researchers are evaluated positively for their ability to come up with strong and novel tests of an existing theory. Studies that have been selected for publication are those that gatekeepers have agreed provide important test of a valuable theory. If, in a later study, the specific result appears

to be unreplicable, this does not necessarily have any bearing on the competence of the original author, who should still be given credit for identifying an interesting question and for developing a reasonable test of the underlying theory in an ideal system. Likewise, fluke findings can happen to anyone.

In practice, however, the scientific process does not always proceed in along this idealized trajectory. Authors of failed replications might face questions of competency and may feel victimized. At least in some fields, authors are less likely to be rewarded for an especially well-designed experiment that tests an existing theory than for a novel theoretical insight that happens to be demonstrated through a particular study. As a result, researchers may feel a sense of ownership over specific research findings, which can mean that failures to replicate can feel like a personal attack, one that can have implications for evaluations of their competence.

Moreover, in a climate where questionable research practices and fraud occasionally contaminate discussions about replication, a failure to replicate can sometimes be interpreted as an accusation of fraud. This contamination is probably an unfortunate accident of history. Concerns about questionable research practices, which gained attention as a response to the evidence for Extrasensory Perception put forth by Bem (2011), coincided with the uncovering of evidence of widespread fraud by the social psychologist Diederik Stapel (Levelt Committee, Noort Committee, Drent Committee, 2012). This underscores the importance of separating discussions of fraud and discussions of best research practices. Conflating the two generates harm and reactance.

Replications also create reputational concerns for the replicators who deserve credit for a thorough effort of assessing the robustness of the original finding (in an ideal world). Again, however, reality can be different from the ideal. To publish original research, one must be creative and daring, whereas such characteristics are not necessarily required of those conducting replication studies. Indeed, Baumeister (2016) has gone so far as to argue that the

replication crisis has created “a career niche for bad experimenters.” Another reputational concern results from the fact that several of the most highly visible replication projects to date have involved relatively large groups of researchers. How does one determine the contributions of and assign credit to authors of a multi-authored replication article (Coyne, 2016)? This problem occurs, for example, when promotion/tenure decisions have to be made.

Response

An increased emphasis on replication studies will lead to new issues regarding reputational concerns. Any form of criticism can sting, and failed replication attempts may feel like a personal criticism, despite the best intentions of those conducting and interpreting these replication attempts. This should be taken seriously. Replicators should go out of their way to describe their results carefully, objectively, and without exaggeration about the implications for the original work. In addition, those whose studies are the focus of replication attempts should give replicators the benefit of the doubt when considering the contribution of the replication study and the replicators’ motivations.

It can be useful for both replicators and original authors to have contact. In some cases, an *adversarial collaboration* (Hofstee, 1984; Kahneman, 2003) may be attempted. An adversarial collaboration is a cooperative research effort that is undertaken by two (groups of) investigators who hold different views on a particular empirical question (e.g., Matzke et al., 2015; Mellers, Hertwig, & Kahneman, 2001). However, contact is often not essential. As noted in the opening sections, if a comprehensive description of the procedures exists, there is little need for contact between replicators and original authors. Some recommendations for collaboration might reinforce the misconception that the original author somehow owns a particular finding as opposed to the finding existing independently of the author as part of the scientific record.

There is some preliminary empirical evidence that failed replications may not exact a reputational toll on authors of the original findings. Fetterman and Sassenberg (2015) surveyed published scientists on how they view researchers whose findings fail to replicate and found that reputational costs are at least overestimated (also see Ebersol, Axt, & Nosek, 2016). As replication attempts become more normative, concerns about reputational costs will lessen. After all, it is likely that all active researchers have published at least some false positives over the course of their career, which means that all researchers should expect some of their work not to replicate. As more replications are conducted, the experience of having a study fail to replicate will become more normative and hopefully less unpleasant.

Many of the reputational costs for those who conduct replications are quite similar to issues that already exist in the field regarding the evaluation of contributions for authorship. Researchers already participate in a wide variety of projects that vary in their novelty and the extent to which the projects are seen as ground-breaking versus incremental. Although many replication studies tend towards the incremental, they can be ground-breaking and novel (such as the systematic attempts to replicate large sets of studies; e.g., Klein et al., 2014, Open Science Collaboration; 2015 Schweisberg et al., 2016). In addition, researchers often already collaborate on large-scale projects with many co-authors, and allocating credit is something that colleagues and promotion committees struggle with quite regularly. Thus, in terms of credit, being involved in replication studies does not differ much from the status quo. This does not mean, however, that researchers should be encouraged to make a career out of conducting replications (and we are unaware of anyone who has given such advice or actually tried this strategy). Conducting replications is a service to the field, but promotion and tenure committees likely will continue to be looking for originality and creativity. Given the current incentive structure in science, some sage advice for early career researchers is to conduct replications with the goal of building on a finding or as only one small part of a portfolio of meaningful research activity.

Concern VI: There is no Standard Method to Evaluate Replication Results

A question that often comes up in practice concerns the interpretation of replication results. Two researchers can look at the same replication study and come to completely different conclusions about whether the original effect was successfully duplicated. This is not entirely unexpected given the importance of judgment in the scientific process (e.g., Cohen, 1990), but nonetheless it can be unnerving to some. For example, the Open Science Collaboration (2015) used a variety of statistical methods to evaluate replication success for the Reproducibility Project: Psychology: (1) Did the focal statistical test produced a statistically significant *p*-value using a predetermined alpha level (typically .05) in the same direction as the original study? (2) Did the point estimate from the original study fall within the 95% confidence interval from the replication study? (3) Does combining the information from original and replication studies produce a significant result? These different metrics can lead to different conclusions, and it is not clear which, if any, one should focus on. This challenge raises an important issue: what is the point of running replication studies at all if the field cannot agree on which ones are successful?

Response

There are always multiple ways to approach a statistical analysis for a given data set (Silberzahn et al., 2017), and the analysis of replications is no different. There is, however, a growing consensus regarding which analyses are the most likely to give reasonable answers to the question of whether a replication study provides results consistent with those from an original study. These analyses include both frequentist estimation and Bayesian hypothesis testing. These different methods may not always agree when they are applied to a particular

case, but often they do (see Etz, 2015; Simonsohn, 2016). Given the multiple options available, investigators should consider multiple approaches and also consider pre-registering analytic plans and committing to how evidence will be interpreted before analyzing the data. Inferences that are robust across approaches are more likely to be more scientifically defensible. Two approaches are especially promising.

One approach is the "small telescopes" approach (Simonsohn, 2015) which focuses on interpreting confidence intervals from the replication study. The idea is to consider what effect size the original study would have 33% power to detect and then use this value as a benchmark for the replication study. If the 90% confidence interval from the replication study excludes this value then we say the original study could not have meaningfully examined this effect. Note that this does not license concluding that the first study was a false positive; as noted by Simonsohn (2015), the focus of this approach shifts attention to the design of the original study instead of just the bottom line result.

A second approach is the "replication Bayes factor" approach (Ly, Etz, Marsman, Wagenmakers, 2017; Verhagen & Wagenmakers, 2014; Wagenmakers, Verhagen, Ly, 2016). The Bayes factor is a number that represents the amount by which the new data (i.e., the results of the direct replication) shift the balance of evidence between two hypotheses, and the extent of the shift depends on how accurately the competing hypotheses predict the observed data (Etz & Wagenmakers, 2017; Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016; Wrinch & Jeffreys, 1921). In the case of a replication study, the researcher compares statistical hypotheses that map to (1) a hypothetical optimistic theoretical proponent of the original effect, and (2) a hypothetical skeptic who thinks the original effect does not exist (i.e., any observed difference from zero is due only to sampling error). The optimist's theoretical position is embodied by the posterior distribution for the effect from the original study, and the skeptic's theoretical position is embodied by the typical null hypothesis that there is no effect. Each of these hypotheses makes different predictions about the results researchers expect to find in a

replication attempt, and the replication Bayes factor can be used to compare the accuracy of these predictions. The skeptic's hypothesis predicts that the replication effect size will be close to zero, whereas the proponent's hypothesis predicts the replication effect size will be away from zero (because the posterior from the original study will typically be centered on non-zero effects) and closer to the result found in the original study. This formulation connects the original and replication results in a way that respects the fact that the two sets of results are linked by a common substantive theory, and in this approach a replication is deemed "successful" if the proponent's hypothesis is convincingly supported by the replication Bayes factor and a "failure" if the skeptic's hypothesis is supported.

Summary and Conclusions

Repeatability is an essential component of science. A finding is arguably not scientifically meaningful until it can be replicated with the same procedures that produced it in the first place. Direct replication is the mechanism by which repeatability is assessed and a tool for for distinguishing progressive from degenerative research program. Recent direct replications from many different fields suggest that the replicability of many scientific findings is not as high as many believe it should be. This has led some to speak of a "replication crisis" in psychology and other fields. This concern is shared by a broad community of scientists. A recent *Nature* survey showed that 52% of scientists across the sciences believe their field has a significant crisis, while an additional 38% believe there is a slight crisis (Baker, 2016). According to the survey, 70% of researchers have tried and failed to reproduce another scientist's findings.

Although the idea that a finding should be able to be replicated is a foundational principle of the scientific method, putting this principle into practice can be controversial. Beyond debates about the definition of replication, many concerns have been raised about (1) when replication studies should be expected to fail, (2) what informational value they provide in a field that hopes

to pursue novel findings that push theory forward, (3) the fairness and reputational consequences of the replication studies that are conducted, and (4) the difficulty in deciding when a replication has succeeded of failed. We have reviewed the major concerns about direct replication and we have addressed them. Replication cannot solve all the field's problems, but when used in concert with other approaches to improving psychological science they help clarify which findings the field should have confidence in as we move forward. Thus, there are no substantive and methodological arguments against direct replication. In fact, replication is an important instrument for theory building. It should therefore be made a mainstream component of psychological research.

Author Note

Alexander Etz was supported by grant #1534472 from NSF's Methods, Measurements, and Statistics panel, as well as the National Science Foundation Graduate Research Fellowship Program (#DGE1321846). We thank the Society for the Improvement of Psychological Science (SIPS) as this paper grew out of discussions at the inaugural meeting. We also thank Dan Gilbert, Brian Nosek, Bobbie Spellman, E.J. Wagenmakers, and two anonymous reviewers for helpful feedback on a previous version. Brent Donnellan is now at Michigan State University.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*, 556–578.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... & Della Penna, N. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037-1037.

Baker, M. (2016). Is there a reproducibility crisis? *Nature, 533,* 452-454.

Barsalou, L.W. (2016). Situated conceptualization offers a theoretical account of social priming. *Current Opinion in Psychology, 12*, 6-11.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. doi:10.1016/j.jesp.2016.02.003

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic: A career guide* (2nd ed., pp. 185–219). Washington, DC: American Psychological Association.

Bem, D.J. (2011). Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect. J*ournal of Personality and Social Psychology. 100*, 407–25.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433-1436.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40-48.

Chambers, C. (2017). *The 7 deadly sins of psychology: a manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., ... & Carcedo, R. J. (2016). Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750-764.

Cohen, J. (1990). Things I have learned (so far). *American psychologist*, *45*(12), 1304.

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., ... & Oxman, A. D. (1993). Should unpublished data be included in meta-analyses?: Current convictions and controversies. *JAMA*, *269*(21), 2749-2753.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology, 4*, 28. doi:10.1186/s40359-016-0134-3.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. DOI: 10.1016/j.jesp.2015.10.002.

Dunlap, K. (1926). The experimental methods of psychology. In: C. Murchison (Ed.) *Psychologies of 1925* (pp. 331-353). Worchester: Clark University Press.

Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' reputations are based on getting it right, not being right. *PLoS biology*, *14*(5), e1002460.

Eerland, A., Sherrill, A.M., Magliano, J.P., Zwaan, R.A., Arnal, J.D., Aucoin, P., … Prenoveau, J.M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*, 158-171.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *Elife*, *3*, e04333.

Etz, A. (2015, August 30). The Bayesian Reproducibility Project[weblog post]. Retrieved 23 August 2017. https://web.archive.org/web/20160407113631/http://alexanderetz.com:80/2015/08/30/the-bayesian-reproducibility-project/

Etz A., & Vandekerckhove, J. (2016) A Bayesian Perspective on the Reproducibility Project: Psychology. *PLoS ONE 11(2)*: e0149794. doi:10.1371/journal.pone.0149794

Etz, A., & Wagenmakers, E. J. (2017). JBS Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science, 32*(2), 313-329.

Fabrigar, L.R., & Wegener, D.T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66*, 68-80.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*, 120-128.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7(6)*, 555-561.

Fetterman, A. K., & Sassenberg, K. (2015). The Reputational Consequences of Failed Replications and Wrongness Admission among Scientists. *PloS one*, *10*(12), e0143723.

Fiedler, K., Schwarz N. (2015). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7, 45-52. DOI: 10.1177/1948550615612150.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275-297. DOI: 10.1037/pspi0000007.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460.

Gilbert, D. T., King, G., Pettigrew, S, Wilson, T. D. (2016). Science 351, 1037.

Greenwald, A. G. (1975). Significance, nonsignificance, and interpretation of an ESP experiment. *Journal of Experimental Social Psychology, 11*, 180-191

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573.

Hewitt, J. K. (2012). Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior genetics*, *42*(1), 1-2.

Hofstee, W. K. B. (1984). Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychologica, 56*, 93–109

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81-92.

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124.

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press, Oxford, UK.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 0956797611430953.

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist, 58*, 723.

Kerr, N.L. (1998). HARKing: Hypothesizing After the Results Are Known. *Personality and Social Psychology Review, 2*, 196–217. doi:10.1207/s15327957pspr0203_4.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, J., Reginald B., Bahnik, S., Bernstein, M. J., . . . Nosek, B. A.(2014). Investigating variation in replicability: A “manylabs” replication project. So*cial Psychology, 45*, 142–152.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS one*, *9*(9), e105825.

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic bulletin & review*, *23*(5), 1631-1638.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In: Lakatos, I., & Musgrave, A., (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability Is Not Optional. *Journal of Personality and Social Psychology, 113*, 254-261.

Levelt Committee, Noort Committee, & Drent Committee, (2012). Flawed science: the fraudulent research practices of social psychologist Diederik Stapel. (2012, November 28). Retrieved August 21, 2017, from www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf

Lupia, A., & Elman, C. (2014). Openness in political science: Data access and research transparency. *PS - Political Science and Politics, 47*, 19-42

Ly, A., Etz, A., Marsman, M., Wagenmakers, E. J. (2017). Replication Bayes factors from Evidence Updating. *PsyArXiv preprints.* Retrieved from https://psyarxiv.com/u8m2s/

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19-32.

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

Makel, M. C., Plucker, J.A., & Hegarty, B. (2012). Replications In Psychology Research. *Perspectives on Psychological Science* 7, 537-542.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. Journal of Experimental Psychology. General, 144(1), e1-e15. DOI: 10.1037/xge0000038

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*, 487–498. http://dx.doi.org/10 .1037/a0039400

Meehl, Paul E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141. doi:10.1207/s15327965pli0102_1

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration.Psychological Science, 12, 269–275.

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. Retrieved from https://raw.githubusercontent.com/richarddmorey/psychology_resolution/master/paper/response.pdf

NCI-NHGRI working group on replication in association studies (2007) Replicating genotype-phenotype associations. *Nature, 447*, 655–660.

Nosek, B.A., & Errington, T.M. (2017). Making sense of replications. *eLife 6*:e23383.Pashler, H. & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531-53.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H. & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

Pashler, H. & Wagenmakers, E.J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science, 7*, 528-530.

Popper, K.R. (1959). *The Logic of Scientific Discovery*, translation of *Logik der Forschung*, Oxford: Routledge.

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views?. *Journal of Experimental Psychology: General*, *144*(4), e73.

Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results, *Psychological Bulletin, 86*, 838-641.

Rotello, C., Heit, E., & Dube, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22*, 944-954.

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: comment on Ferguson and Brannick (2012). Psychological Methods, *17*, 129-136.

Schmidt, S. (2009) Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100.

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4*(1), 32-37. http://dx.doi.org/10.1037/arc0000029.

Schweinsberg, M. et al. (2016). The pipeline project: pre-publication independent replications of a single laboratory's research pipeline. journal of experimental social psychology, 66, 55–67.

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., ... & Puhlmann, L. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives?.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. C., ... & Carlsson, R. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *PsyArXiv Preprint.*

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. http://dx.doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (in press). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science.*

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science, 26*, 559-569.

Simonsohn, U. (2016, March 3). [47] Evaluating Replications: 40% Full ≠ 60% Empty.

Retrieved 23 August 2017.

https://web.archive.org/web/20170709184952/http://datacolada.org/47

Smaldino P.E. & McElreath R. (2016). The natural selection of bad science. *Royal Society Open Science, 3,* 160384.

Smart, R. G. (1964), The Importance of negative results in psychological research. *Canadian Psychologist, 5*, 225-232.

Spellman, B.A. (2015). A Short (Personal) Future History of Revolution 2.0. P*erspectives on Psychological Science, 10,* 886-899.

Sterling T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*, 108–112.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*, 59–71. doi:10.1177/1745691613514450.

van Aert, R. C., & van Assen, M. A. (2017). Bayesian evaluation of effect size after replicating an original study. *PloS one*, *12*(4), e0175302.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 201521897.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143*, 1457-1475.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., . . . Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*, 917–928.

Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods, 48*, 413-426.

Wrinch, D., & Jeffreys, H. (1921). XLII. On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *42*(249), 369-390.

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2017). Participant nonnaiveté and the reproducibility of cognitive psychology. Psychonomic Bulletin & Review. doi:10.3758/s13423-017-1348-y