BEHAVIORAL AND BRAIN SCIENCES (2018), Page 1 of 61
doi:10.1017/S0140525X17001972, e120

Making replication mainstream

Rolf A. Zwaan
Department of Psychology, Education, and Child Sciences, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands
zwaan@essb.eur.nl
https://www.eur.nl/essb/people/rolf-zwaan

Alexander Etz
Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100.
etz.alexander@gmail.com
https://alexanderetz.com/

Richard E. Lucas
Department of Psychology, Michigan State University, East Lansing, MI 48824
lucasri@msu.edu
https://www.msu.edu/user/lucasri/

M. Brent Donnellan[1]
Department of Psychology, Texas A&M University, College Station, TX 77843
donnel59@msu.edu
https://psychology.msu.edu/people/faculty/donnel59

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: psychological research; replication; reproducibility; research programs

1. Introduction

> 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

 0140-525X/18

studies a more regular and integral part of research, a shift that could potentially increase confidence in the veracity of findings. Although debates 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.

2. 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 no specific skills are 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 et al. 2016; Lebel et al. 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 1990b). 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 1990b) 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 1990b, p.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.

3. Issues with replicability

Concerns about the replicability of scientific findings have arisen in a number of fields, including psychology (Open Science Collaboration 2015), genetics (National Cancer Institute-National Human Genome Research Institute Working Group on Replication in Association Studies [NCI-NHGRI] 2007; Hewitt 2012), cancer research (Errington et al. 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 et al. 2014; Smart, 1964; Sterling 1959; Sterling et al. 1995). Yet there are additional forces and practices that can increase the rates of false positives. For example, there is a growing body of meta-scientific research

ROLF ZWAAN, Professor of Psychology and Chair of the Brain and Cognition area at Erasmus University Rotterdam, has (co)authored over 150 publications in the areas of language processing, memory, and cognition.

ALEXANDER ETZ, Ph.D. student at the University of California, Irvine, is the author of ten publications in the areas of psychology and statistics, including works published in *Psychonomic Bulletin & Review* and *Statistical Science*, and is a recipient of the National Science Foundation Graduate Research Fellowship.

M. BRENT DONNELLAN, Professor of Psychology at Michigan State University, is the author or coauthor of over 175 publications in the areas of personality psychology, developmental psychology, and personality assessment.

RICHARD E. LUCAS, MSU Foundation Professor in the Department of Psychology at Michigan State University, is the author of over 100 publications in the area of personality and subjective well-being. He has received awards such as the Carol and Ed Diener Award for Mid-Career Contribution to Personality Psychology.

https://doi.org/10.1017/S0140525X17001972

showing the effects of excessive *researcher degrees of freedom* (John et al. 2012; Simmons et al. 2011) or latitude in the way research is conducted, analyzed, and reported. If researchers experience pressure to publish statistically significant findings, then 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 data set 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 on their permissibility (John et al. 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 2003):

> 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 affected not only 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.

4. 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 functions.

Given these different functions, a number of typologies have been offered for classifying replication studies (e.g., Hüffmeier et al. 2016; Lebel et al. 2017; Lykken 1968; Schmidt 2009; Schmidt & Oh 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, because 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 United Kingdom 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 differences 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 a 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 does, in effect, is test an extension of the theory to a new context (because 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 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 and 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 the exact mental state 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 such reasons, 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 critical 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, because 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 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 et al. 2017).

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 rests 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 settings similar to those of 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

https://doi.org/10.1017/S0140525X17001972

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, because they do not directly address scientifically relevant features of the research. As we explain in more detail in section 5.2.1., 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.

5. Concerns about replication

The interpretation of specific replication studies has produced considerable disagreement and controversy. For example, 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 et al. 2016; Kunert 2016; Maxwell et al. 2015; Morey & Lakens, 2016; van Aert & van Assen, 2017; Van Bavel et al. 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.

5.1. 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 may 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 et al. 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 multimodal (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, for example, 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 whether the context has changed sufficiently to wipe out that effect.

5.1.1. 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 of 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 more than the objective characteristics of those studies when evaluating claims about quality and scientific rigor. It is possible simultaneously to 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 example, 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 and 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 example, imagine that a hypothetical priming effect in social psychology can emerge only when: (1) a sample of

https://doi.org/10.1017/S0140525X17001972

college students has a specific average level of political conservatism, (2) the experiment took place at a particular time in the semester, (3) the experiment was conducted at a particular time of the day, (4) the experimenter who first meets the participant dresses in a lab coat to emphasize the serious, scientific nature of the study, and (5) 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 recruited (they were recruited from the population that was available), the time of year when the study was run (they began data collection when approval was obtained from the institutional review board), 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 would 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 noncumulative 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 led 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 et al., 2017). 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 prespecify 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 relative 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 multilab 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. 2016a).

Two strategies for solving the concerns outlined in this section are to (1) raise standards in reporting of experimental detail, such that original papers contain *replication recipes* (Brandt et al. 2014; Dunlap 1926; Popper 1959/2002), and (2) 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., 2017).

5.2. 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 et al. 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 (Smaldino & McElreath 2016).

5.2.1. 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 occur only 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 et al. 2016; Hagger et al. 2016; Shanks et al. 2015; Wagenmakers et al. 2016a). Those who argue that a large set of successful conceptual replications would not be possible in the absence of 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 of 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

https://doi.org/10.1017/S0140525X17001972

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 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. 174-96). 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 et al. 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.

5.3. 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 because of 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 sect. 5.5 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.

5.3.1 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, and 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 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 example, 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 in sect. 5.5), 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, preregistered, 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 preregistered, 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 of 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 a 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.

5.4. 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 the bigger problems psychology is facing (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.

5.4.1. 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 more broadly foundational in science. Also, 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 toward 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

https://doi.org/10.1017/S0140525X17001972

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 over 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 & Heene 2012; Franco et al. 2014; Kühberger et al. 2014; Rothstein & Bushman 2012) or the prevalence and severity of questionable research practices (Fiedler & Schwarz 2015; John et al. 2012, Simmons et al. 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 attempts to replicate effects that publication bias-sensitive meta-analyses suggest are robust should be more successful than attempts to replicate 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 rightly 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. It is just one element of the toolbox of methodological reform.

5.5. 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 for 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, anyone can obtain fluke findings.

In practice, however, the scientific process does not always proceed in 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 of 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 their thorough effort in 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.

5.5.1. 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 et al. 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 Ebersole et al. 2016b). 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, it is hoped, 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 toward 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. 2014a; Open Science Collaboration 2015; Schweinsberg 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.

5.6. 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 produce a statistically significant p value using a predetermined α 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 on which, if any, one should focus. 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?

5.6.1. 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 on 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 et al. 2017; Verhagen & Wagenmakers 2014; Wagenmakers et al. 2016b). 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 et al. 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

https://doi.org/10.1017/S0140525X17001972

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 nonzero 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.

6. 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 distinguishing progressive from degenerative research programs. 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 reported that 52% of scientists across the sciences believe their field has a significant crisis, whereas 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 or failed. We have reviewed the major concerns about direct replication and we have addressed them. Replication cannot solve all of the field's problems, but when used in concert with other approaches to improving psychological science, it helps 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.

ACKNOWLEDGMENT

Alexander Etz was supported by Grant 1534472 from National Science Foundation's Methods, Measurements, and Statistics panel, as well as the National Science Foundation Graduate Research Fellowship Program (Grant 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.

NOTE

1. Current address: Department of Psychology, Psychology Building, 316 Physics Road, Room 249A, Michigan State University, East Lansing, MI 48824.

Open Peer Commentary

If we accept that poor replication rates are mainstream

doi:10.1017/S0140525X18000572, e121

David M. Alexander and Pieter Moors

Brain and Cognition, KU Leuven, Tiensestraat 102, BE-3000 Leuven, Belgium.

david.alexander@kuleuven.be pieter.moors@kuleuven.be

http://www.perceptualdynamics.be

www.gestaltrevision.be

Abstract: We agree with the authors' arguments to make replication mainstream but contend that the poor replication record is symptomatic of a pre-paradigmatic science. Reliable replication in psychology requires abandoning group-level *p*-value testing in favor of real-time predictions of behaviors, mental and brain events. We argue for an approach based on analysis of boundary conditions where measurement is closely motivated by theory.

We relish the authors' arguments to make replication mainstream. They have done a tremendous job in summarizing and countering objections to their cause. Yet acceptance that routine direct replication is crucial (and missing) has not quite reached a critical level among psychological scientists. We view the poor replication record as symptomatic of a pre-paradigmatic science. If we accept that poor replication rates are mainstream, this means the really brutal methodological cold turkey has yet to be braved.

Despite its stated goals, psychology does not in practice aim to establish entirely reproducible effects. Zwaan et al.'s exhortation "to make replication mainstream" arises because of this contrast between goals and practice. Consider any run-of-the-mill journal in physics, chemistry, or some other well-established science in light of the concerns raised in Zwaan et al.'s last paragraph. Applying these to any of the sciences listed would come across as odd; as the tail wagging the dog. Although such an assessment may be unpopular, a science without a core canon of directly reproducible results is not yet a science. Nevertheless, the present pre-paradigmatic phase of psychology arises as a result of the maturity of the field, and not through any particular incompetence or dishonesty on the part of we scientists. Inevitably, psychology will still face a reproducibility problem 20 years from now, even when recommendations such as preregistration, open materials, data, and code are standard (cf. Meehl 1990a). Even those results that now are technically reproducible are not often

reproducible in a predictive sense: that they enable theoretically related problems to be solved in a straightforward fashion.

Many solutions suggested for the concerns highlighted by Zwaan et al. are decades old (Meehl 1967). The reproducibility crisis presents a sober occasion to revisit them, given our accumulating research experience. Our view is that psychology and cognitive neuroscience have succeeded, in part, by picking the low hanging fruit. By this we mean gathering those results that can be distinguished by using assumptions of a linear, low-dimensional measurement space, treating unparceled variance as "noise" and operational definitions of experimental manipulations as sufficient.

We argue reliable that replication requires reformulating the nature of the *ceteris paribus* clause ("holding everything else constant"). This clause is usually interpreted as requiring tight control of subjects' behavior, so everything except the phenomenon of interest is excluded from influencing the experimental outcome. This restriction becomes problematic when applied to a complex nonlinear system like a person embedded in an experimental environment. Instead, we propose that the object to be controlled is the entire experimental (and pseudo-naturalistic) space in which the phenomenon of interest is evoked (Manicas & Secord 1983). The goal is to exactly explore this space, to find out how the phenomenon changes over relevant parameters and where it is only trivially different. Rather than colliding opposing theoretical positions (debates on X vs. Y), the goal is to demarcate when one type of phenomenon (e.g., conscious, directed attention) becomes another (e.g., automatic attention), by defining its boundary conditions. A major focus of theory is then to commit to the experimental space being a certain (potentially nonlinear) shape and dimensionality (Hultsch & Hickey 1978; Wallot & Kelty-Stephen 2018). The "constant" of the *ceteris paribus* clause is the requirement to accurately (and repeatedly) position the subject in a desired portion of the theoretically defined experimental space. Importantly, this introduces direct theoretical criteria for deciding whether an experiment was run correctly and blurs the distinction between direct and conceptual replication.

To prevent new evidential walls made of loose bricks, we believe such a reformulation inherently requires abandoning null hypothesis significance testing (NHST) as the *primary* piece of evidence (Szucs & Ioannidis 2017b). Mere differences are insufficient to characterize the experimental space and to position a subject within it. Theory should provide us point predictions (Lakens 2013; Widaman 2015). This approach allows us to explore the nonlinear nature of the experimental space and to explicitly motivate research practices like data transformation and aggregation. For example, if data appear log-normally distributed, transformation is allowed only if theory states the value range has geometric symmetries. What appears as a practical data-cleaning operation in mainstream NHST could be a gross distortion of the underlying phenomena when experiment, theory, and data analysis are required to be tightly intertwined.

Furthermore, point predictions should be formulated at the individual rather than group level. Much of our statistics was originally developed for agronomy, where individual kernel weights can be aggregated to (trivially) calculate yields for the crop field. This is generally not the case for the relationship between individual behaviors or neural measurements and the concomitant aggregate outcomes across subjects (Alexander et al. 2013; 2015; Estes 1956). Yet, in our experience, it is rare for a paper to be rejected because the authors have not proved that measures are linearly behaved enough to bear the assumptions of aggregation methods such as cross-trial and cross-subject averaging.

A consequence of individual predictions is that replication will then involve running some more subjects over a range of experimental conditions, each of which is a test of theory. Thus, the proposed redefinition of the *ceteris paribus* clause may limit the otherwise onerous resource requirements to reproduce experimental results. Likewise, our redefinition mandates the conditions under which the vast knowledge base of (mostly linear) statistical assessment methods can be justifiably used. Provided experiment and theory-dictated numerical transformations leave the data in a linear space, linear methods are available.

A side effect of adopting something like our present proposal is that it levels the playing field. Results from the history of findings in psychology cannot be regarded as certain until they have achieved successful replications of the kind that Zwaan et al. argue for. We further suggest that this will not occur until a framework is adopted that requires empirical feedback on the validity and success of each experimental manipulation and theoretically mandates every post-experiment transformation of the data. This, in turn, will not occur until the bitter pill that non-replication is mainstream has been swallowed.

Replications can cause distorted belief in scientific progress

doi:10.1017/S0140525X18000584, e122

Michał Białek[a,b]

[a]*Department of Psychology, University of Waterloo, Waterloo, ON N2L 3G1, Canada;* [b]*Centre for Economic Psychology and Decision Sciences, Kozminski University, 03-301 Warsaw, Poland.*

mbialek@uwaterloo.ca

http://mbialek.com.pl

Abstract: If we want psychological science to have a meaningful real-world impact, it has to be trusted by the public. Scientific progress is noisy; accordingly, replications sometimes fail even for true findings. We need to communicate the acceptability of uncertainty to the public and our peers, to prevent psychology from being perceived as having nothing to say about reality.

Zwaan et al. extensively discuss six concerns related to making replication mainstream. I raise a different one – distorted perception of science by the public and, perhaps also, by peer scientists.

Among the public there is a "myth of science": an implicit assumption that scientific findings report true effects, and that, once a study is conducted, all scientists agree on its results (Pitt 1990). For example, when a mathematician is showing a proof for a formula, it makes everybody acknowledge its validity. Similarly, in logic or philosophy, finding a counterexample falsifies the whole theory. People can have similar expectations toward empirical sciences like psychology. Anticipating these expectations, mass media present people with reports of scientific advancements, rarely mentioning any associated uncertainty (Dudo et al. 2011). In this context, it is not that surprising that presenting people with information about the level of scientific consensus on a particular finding, even an unlikely high one (e.g., 98%), sometimes backfires. People interpret the less-than-100% consensus as a degree of uncertainty they did not expect, and, as a result, they reduce their belief in such a finding (Aklin & Urpelainen 2014). In short, people (including my past self) expect scientific findings to be certain, and failing to meet their expectation may lead to disbelief in reported findings, in scientific domains, or even in science as a whole.

This is relevant to the effort to make replications mainstream because the replication movement necessarily introduces a substantial degree of uncertainty into science. For example, the well-cited Open Science Collaboration (2015) was expected to replicate only 65.5% of tested studies under the assumption that *every* original study reported a true effect (Gilbert et al. 2016). Yet, only 47% of the original studies were successfully replicated, becoming a vivid illustration of the "replication crisis." Regardless of whether the Open Science Collaboration would succesfully replicate half or two-thirds of investigated studies, both of these numbers are substantially lower than those expected by the lay audience, namely, that all original studies should replicate. In fact, it is likely that the replication crisis would have arisen even if the Open Science Collaboration had replicated a much higher

proportion of original studies, and even more than the expected 65.5%.

Among scientists, consensus on an issue is likely to depend on congruity of data. Given that replications can fail even for true findings, making replication mainstream backfires so that even experts might be suffering doubt in true findings and might struggle to distinguish between true and false findings. This, in turn, will magnify the doubt of the public and drift the real-world applications of scientific discovery toward zero. Empirical evidence indicates that casting *any* doubt on scientific evidence decreases support for the implementation of a public policy based on such evidence (Koehler 2016). Underlining scientific uncertainty is sometimes used eristically: "serves nothing but defeating or postponing new regulations, allowing profitable but potentially risky activities to continue unabated" (Freudenburg et al. 2008).

To be clear, the present argument is not that scientists should stop replicating studies because they will look bad in the eyes of the public. Rather, we need to actively work on communicating the acceptability of uncertainty associated with scientific findings to the public (and to our peers too). We, as scientists, simply do not want to be perceived as the ones who know nothing and, therefore, are not worth listening to. Quite the opposite, we want to communicate the noisy, but steady progress of science in general, and psychology in particular. We also want our findings to be implemented in public policy, so that we contribute to making the world a better place. To accomplish that, we need to ensure that the public understands how science works and that uncertainty is something natural in science, and not a sign of junk science. The implementation of public policies informed on scientific evidence should be made like judicial verdicts. They should be based on evidence beyond "reasonable doubt," not on absolute certainty.

One thing we could do is to keep in mind how things look among the public, and emphasize the importance of replications not in terms of weeding out "bad science," but the normal self-correction that is the very basis of scientific discovery. Communicating uncertainty associated with scientific progress will determine whether massive replications will have predominantly positive or negative effects.

Strong scientific theorizing is needed to improve replicability in psychological science

doi:10.1017/S0140525X1800078X, e123

Timothy Carsel, Alexander P. Demos, and Matt Motyl

Department of Psychology, University of Illinois at Chicago, Chicago, IL 60607

timothy.carsel@gmail.com ademos@uic.edu matt.motyl@gmail.com
www.timcarsel.wordpress.com www.alexanderdemos.org/
www.mattmotyl.com

Abstract: The target article makes the important case for making replicability mainstream. Yet, their proposal targets a symptom, rather than the underlying cause of low replication rates. We argue that psychological scientists need to devise stronger theories that are more clearly falsifiable. Without strong, falsifiable theories in the original research, attempts to replicate the original research are nigh uninterpretable.

We applaud Zwaan et al. for compiling many of the present concerns researchers have regarding replication and for their thoughtful rejoinders to those concerns. Yet, the authors gloss over an underlying cause of the problem of the lack of replicability in psychological science and instead focus exclusively on addressing a symptom, specifically that the field does not make replications a centerpiece of hypothesis testing. An underlying cause is that psychologists do not actually propose "strong" testable theories. To paraphrase Meehl (1990a), null hypothesis testing of "weak" theories produces a literature that is "uninterpretable." In particular, this is because the qualitative hypotheses generated from weak theories are not formulated specifically enough, just that "X and Y" will interact. Thus, any degree and form of interaction could be used to support the [frequentists'] statistical hypothesis. Further, it is important to remember that the statistical hypothesis, that is, the alternative to the null, is never actually true (Cohen 1994) and can address only the degree of the interaction, not the form. In other words, both a disordinal interaction from an original study and an ordinal interaction from a replication would yield statistical support for the interaction hypothesis. Had the theory been stronger, the hypothesis would have predicted a specific degree and form of the interaction, resulting in the non-replication of the original study by the second. This in part may explain how we came to the conclusion in our own examination of research practices and replication metrics of published research (Motyl et al. 2017; Washburn et al., 2018) that the metrics of replicability seemed to support Meehl's prediction that a poorly theorized scientific literature would produce "uninterpretable" results. Thus, the authors' concern VI (sect. 5.6) regarding point estimation (e.g., effect size, *p*-values) and their confidence intervals implicitly assume that the original study and replication were interpretable results regarding the verisimilitude of the theory. To summarize this argument, take for a moment the example of throwing darts at a dart board. Zwaan et al. were concerned with whether the second dart came near the first. However, based on the way psychology often works, the size of the bullseye may be the whole wall. Thus, replication can only contribute to the falsification of a theory that is well-defined.

The current predicament of weak theorizing may be created in part by the thinking that "humans are too complicated for strong theories." Zwaan et al. speak to the symptom of this problem by stating that "context" needs to be better described in our methods sections. Psychological theories often require substantial auxiliary theories and hypotheses to "derive" the qualitative hypotheses that motivate our studies (Meehl 1990b). In short, this leads to the problem of "theoretical degrees of freedom" such that the ambiguous theory can be re-instantiated in such a way that any result we may find will be used as support for our, in fact, unfalsifiable theories. Zwaan et al. assert "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 the narrow sense" (sect. 2, para. 3). The kind of falsification advocated by Zwaan et al., however, becomes increasingly difficult the further removed a statistical hypothesis is from the qualitative hypothesis (Meehl 1990a), and the finding is rendered uninterpretable when our statistical and qualitative hypotheses become couched in an increasing number of implicit auxiliary hypotheses. Indeed, if our theories are so weak that any contextual change negates them, then those are not theories; they are hypotheses masquerading as theories. Gray (2017) proposed a preliminary method to instantiate our theories visually, which forces the scientist to think through their theory's concepts and relationships. This is a stronger recommendation than the ones made by Zwaan et al., who suggest simply being more careful about statements of generalizability. Concerns II (sect 5.2) and IV (sect. 5.4) would be resolved with stronger theorizing, more careful derivations and discussions of statistical and qualitative hypotheses, as well as both direct and conceptual replications to test the boundary conditions of those theories.

In summary, we contend that the target article authors are right that we need to make replication more mainstream, but argue that we need to go further and encourage stronger theorizing to help make replications more feasible and meaningful.

The costs and benefits of replication studies

doi:10.1017/S0140525X18000596, e124

Nicholas A. Coles,[a] Leonid Tiokhin,[b] Anne M. Scheel,[c] Peder M. Isager,[c] and Daniël Lakens[c]

[a]*Department of Psychology, Austin Peay Building, University of Tennessee, Knoxville, TN 37996;* [b]*School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85281;* [c]*Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, 5600 MB, Eindhoven, The Netherlands.*

colesn@utk.edu ltiokhin@asu.edu a.m.scheel@tue.nl
p.isager@tue.nl D.Lakens@tue.nl
http://leotiokhin.com/
http://www.tue.nl/staff/a.m.scheel
http://www.tue.nl/staff/p.isager
http://www.tue.nl/staff/d.lakens

Abstract: The debate about whether replication studies should become mainstream is essentially driven by disagreements about their costs and benefits and the best ways to allocate limited resources. Determining *when* replications are worthwhile requires quantifying their expected utility. We argue that a formalized framework for such evaluations can be useful for both individual decision-making and collective discussions about replication.

In a summary of recent discussions about the role of direct replications in psychological science, Zwaan et al. argue that replications should be more mainstream and discuss six common objections to direct replication studies. We believe that the debate about the importance of replication research is essentially driven by disagreements about the value of replication studies and the best way to allocate limited resources. We suggest that a decision theory framework (Wald 1950) can provide a tool for researchers to (a) evaluate costs and benefits to determine *when* replication studies are worthwhile, and (b) specify their assumptions in quantifiable terms, facilitating more productive discussions in which the sources of disagreement about the value of replications can be identified.

The main goal of decision theory is to quantify the *expected utility* (the result of a cost-benefit analysis, incorporating uncertainty about the state of the world) of possible actions to make an optimal decision. To determine when a replication study is valuable enough to perform, we must compare the expected utility of a replication study against alternative options (e.g., performing a conceptual replication or pursuing novel lines of research). In this commentary, we explore some of the costs and benefits of direct replications and emphasize how different assumptions can lead to different expected-utility judgments.

Costs and benefits of direct replications. The expected utility of replication studies depends on several factors, such as judgments about the reliability of the literature, the perceived public interest in a finding, or the judged importance of a theory. Importantly, these assessments are subjective and can lead to disagreements among researchers. Consider the concerns addressed by Zwaan et al.: Should we continue to examine highly context-dependent effects or limit ourselves to effects that are robust across contexts? Should we spend more resources on direct or conceptual replications? Are direct replications prohibitively costly in large-scale observational studies? The answer is: It depends.

Highly context-dependent effects might, as Zwaan et al. note, make it "difficult, if not impossible, for new knowledge to build on the solid ground of previous work" (sect. 5.1.1, para. 8, concern I). However, to argue against pursuing these research lines, one must make the case that such costs outweigh the expected benefits. In some research areas, such as personalized medicine, highly context-dependent effects may be deemed worthwhile to pursue. If researchers believe some (perhaps even all) effects are highly context dependent, they should be able to argue why these effects are important enough to study, even when progress is expected to be slow and costly.

Some researchers argue that even a single replication can be prohibitively costly (sect. 5.3, concern III). For example, Goldin-Meadow stated that "it's just too costly or unwieldy to generate hypotheses on one sample and test them on another when, for example, we're conducting a large field study or testing hard-to-find participants" (2016). Some studies may be deemed valuable enough to justify even quite substantial investments in a replication, which can often be incorporated into the design of a research project. For instance, because it is unlikely that anyone will build a Large Hadron Collider to replicate the studies at the European Organization for Nuclear Research (CERN), there are two detectors (ATLAS and CMS) so that independent teams can replicate each other's work. Thus, high cost is not by itself a conclusive argument against replication. Instead, one must make the case that the benefits do not justify the costs.

The expected utility of a direct replication (compared with a conceptual replication) depends on the probability that a specific theory or effect is true. If you believe that many published findings are false, then directly replicating prior work may be a cost-efficient way to prevent researchers from building on unreliable findings. If you believe that psychological theories usually make accurate predictions, then conceptual extensions may lead to more efficient knowledge gains than direct replications (sect. 5.2, concern II). An evaluation of costs might even reveal that neither direct nor conceptual replications are optimal, but that scientists should instead focus their resources on cheaper methods to increase the reliability of science (sect. 5.4, concern IV).

The value of replication studies is also influenced by the anticipated interpretation of their outcomes (sect. 5.6, concern VI). If we cannot reach agreement about how to evaluate a given result, its benefit to the field may be close to zero. The outcome of a replication study should increase or decrease our belief in an effect, or raise new questions about auxiliary assumptions that can be resolved in future studies. Replications may thus have higher subjective value when consensus about the interpretation of outcomes can be determined *a priori* (e.g., via pre-registered adversarial collaboration).

Replication attempts may also have social costs and benefits for researchers who perform replication studies, or whose work is replicated. One strength of decision theory is that it allows us to incorporate such social components in cost-benefit analyses. For example, researchers currently seem to disagree about when, and how much, reputations should suffer when findings do not replicate (sect. 5.5, concern V). If the reputational costs of unsuccessful replications are too high, scholars may be overly reluctant to publish novel or exploratory findings. If the reputational costs are nonexistent, scholars may not exert ideal levels of rigor in their work. The social norms influencing these costs and benefits are shaped by the scientific community. Explicitly discussing those norms can help us change them in ways that incentivize direct replications when they, ignoring the social consequences, would have high utility.

Conclusion. It is unlikely that directly replicating every study, or never directly replicating any study, is optimally efficient. A better balance would be achieved if researchers performed direct replications when the expected utility exceeded that of alternative options. Decision theory provides a useful framework to discuss the expected utility of direct replications based on a quantification of costs and benefits. A more principled approach to deciding when to perform direct replications has the potential to both help researchers optimize their behavior and facilitate a more productive discussion among researchers with different evaluations of the utility of replication studies.

The meaning of a claim is its reproducibility

doi:10.1017/S0140525X18000602, e125

Jan P. de Ruiter

Departments of Computer Science and Psychology, Tufts University, Medford, MA 02115.

jp.deruiter@tufts.edu

https://sites.tufts.edu/hilab/

Abstract: A scientific claim is a *generalization* based on a reported statistically significant effect. The reproducibility of that claim is its scientific meaning. Anything not explicitly mentioned in a scientific claim as a limitation of the claim's scope means that it implicitly generalizes over these unmentioned aspects. Hence, so-called "conceptual" replications that differ in these unmentioned aspects from the original study are legitimate, and necessary to test the generalization implied by the original study's claim.

I commend the authors for carefully addressing some of the canards that have emerged in the recent attempts to downplay the crucial role of replication in psychology. However, they fail to avoid a widespread conceptual confusion that has substantially contributed to the declining reputation of replication in psychology.

In the target article, the words *finding*, *result*, *effect*, and *claim* are used interchangeably, probably for reasons of stylistic variation. That is fine, but it also obscures a number of relevant distinctions. Independent of the words we use, it is important in the context of replication to distinguish between the following concepts: *data*, the raw recordings of the dependent measure(s); *difference*, the descriptive difference between aggregated values of the data for the relevant conditions; and *significant difference*, or *effect*, which is the statistical generalization of an observed difference, demonstrating that the difference cannot be explained by chance alone. The presence of an *effect* is necessary (but not sufficient) for making a *claim*, which is an effect believed to be generalizable to the population and context of interest. Claims either support or undermine theories, which is why we make them in the first place.

Every one of these concepts is an avenue for replication. If we replicate *data*, we essentially double-check for measurement errors or fraud. If we replicate a *difference* found in the data, we double-check for the way the data were aggregated, for example, to avoid Simpson's paradox (Simpson 1951). If we replicate an *effect*, we reproduce the statistical procedure that was used to make sure that the difference was not due to chance alone. Finally, and most importantly for the present discussion, if we do a replication study about a *claim*, we check the generalizability of the effect over the population, task, and other aspects of the context that the claim was about. In that sense, the *reproducibility of a claim is its scientific meaning*. If we establish in a controlled experiment at Tufts University that undergrads in Computer Science perform better at a math test when they have had coffee than when they have not, the fact that those specific undergrads performed better at that specific math test on that specific day after having that specific amount of that specific type of coffee is not particularly interesting. It is the underlying claim (which one hopes is clearly specified in the study) that people, young adults, or students perform better at math, analytic problem solving, or whatever the claim says, when they have had coffee, or caffeine, or whatever the claim says.

So the discussion of direct versus conceptual replication, as well as the assessment of the value of a "conceptual replication," can be elegantly addressed once we realize that all the replication of a claim does is explore the generalizability of that claim.

The more general the claim the finding is held to support, the more "conceptual" the replication of the supporting findings can (and should) be. Suppose we have an effect E that we report to claim evidence for scientific claim C. Then, if C is *identical* to E, such that C is a claim of the type "The participants in our experiment did X at time T in location L performing task X," it is impossible to replicate that claim because the exact circumstances under which E was found were unique and, therefore by definition, irreproducible. But in this case (that C = E), C obviously has no generality at all and is therefore scientifically irrelevant. If, on the other hand, C is more general than E, the level of detail that is provided in the claim should be sufficient to enable readers to attempt to replicate the claim, allowing for variation that the authors do not consider important. If the authors remark that the effect arises under condition A, but acknowledge that it might not arise under condition B (e.g., with participants who are aged 21–24 rather than 18–21), then clearly a follow-up experiment under condition B is not a valid replication. But if their claim does not specify the age for which the claim should hold, then a follow-up study involving condition B is a perfectly legitimate replication. The failure to specify any particular limitation of the claim might reasonably be considered an implicit statement that the claim is so general that changing this aspect in a replication study should not matter.

So assuming the data are accurate and the statistical generalization is solid, if we just use the rule "whatever is not specified in the claim is something the claim is generalizing over," we accomplish three (good) things. First, we create an incentive for authors to be more careful in specifying the generalizability of their claims. Second, we make it easier to replicate studies to assess the validity of their claims. And third, we avoid the possible cop-out for authors of nonreplicated studies that the study did not replicate because of "unknown moderator variables." If these variables were not excluded in the original study by limiting the generality of the claim, they cannot be invoked to discredit a failed replication.

A possible argument against the proposed rule is that it becomes much harder to make claims that hold up under replication. My response to that argument is that this is not a bug, but a feature. Finding general effects in psychology *is* very difficult, and it would be a good first step to address our replication crisis if we stopped pretending it is not.

To make innovations such as replication mainstream, publish them in mainstream journals

doi:10.1017/S0140525X18000614, e126

Boris Egloff

Department of Psychology, Johannes Gutenberg University Mainz, D-55099, Germany.

egloff@uni-mainz.de

http://www.ppd.psychologie.uni-mainz.de/62.php

Abstract: It was a pleasure to read Zwaan et al.'s wise and balanced target article. Here, I use it as a shining example for bolstering the argument that to make innovations such as replication mainstream, it seems advisable to move the debates from social media to respected "mainstream" psychology journals. Only then will mainstream psychologists be reached and, we hope, convinced.

In this commentary, I argue that the important debates in our discipline (e.g., whether and how to replicate) necessarily belong in scientific journals and should not be restricted to the blogosphere or the social media universe. I concede that the issue I am raising (i) is not at odds with Zwaan et al. (in fact, their explicit goal is to move the debate to a journal) and (ii) does not address the main content of their target article (I completely agree with all of Zwaan et al.'s claims about

replication), but instead focuses on a seemingly minor point that was mentioned in the target article (i.e., the outlet in which the debate takes place). However, I am convinced that the outlet is of critical relevance here.

First and most important, I believe that most mainstream scientists still read scientific journals more frequently and more intensely than they follow social media. Thus, it is simply more efficient to publish fresh ideas in journals to gain optimal access to "the silent majority" whom authors would like to convince. A perfect example here is the success of the "False-Positive Psychology" article published in *Psychological Science* (Simmons et al. 2011; see also Simmons et al. 2018). A few additional examples that readily come to mind are the publication of the results of the "Replication Project: Psychology" in *Science* (Open Science Collaboration 2015), the – regrettably renamed – "Voodoo Correlations" paper in *Perspectives on Psychological Science* (Vul et al. 2009), the "Scientific Utopia" article in *Psychological Inquiry* (Nosek & Bar-Anan 2012), and the mind-boggling "Political Diversity" paper in *Behavioral and Brain Sciences* (Duarte et al. 2015).

Of course, it is certainly difficult and all too often very frustrating to try to publish innovative ideas or critiques of established theories in journals because the thorny peer-review process sometimes seems to be abused by established scholars in their roles of reviewers and editors in efforts to block innovations and criticism. By contrast, all ideas can quickly and without filtering be published in blogs, and there have been several additional clever arguments put forward in favor of blogs over journals (e.g., open data, code, and materials, open reviews, no eminence filter, better error correction, and open access; Lakens 2017). On the other hand, established scholars sometimes complain about, for example, a lack of reflection, a lack of peer advice, impulsivity, personalized debates, and personal accusations triggered by the features of social media. Although I believe that the "tone debate" has been largely exaggerated – "Don't dish it out if you can't take it" – there is some evidence that intellectual opponents and especially third parties might be more efficiently convinced if the arguments are presented in a friendly tone. Thus, the more formal and down-to-earth tone used in scientific journals might in fact be helpful for convincing others. Similarly, mainstream journals are, in general, still more highly respected than most social media outlets. Thus, especially more conservative scholars will trust arguments exchanged in journals more than those that come from debates fought out in blogs.

This should by no means be interpreted to mean that blogs and social media do not have their merits in the replication debate and beyond. To the contrary: They are fast, they are subjective, they are mostly short and to-the-point, they may be provocative, and so forth. My argument is instead that the important debates in our discipline (e.g., whether and how to replicate) should not be restricted to these media but should also be published in established mainstream journals. Although such journals are necessarily somewhat slower, they offer another form and style and can potentially present a more elaborated form of the argument. If one mainstream journal rejects your paper, please try another (and so on). There are also newly founded – not yet so well-established – journals such as *Collabra*, *Metapsychology*, or *Advances in Methods, and Practices in Psychological Science* (to name just a few) that might be alternatives in the face of repeated publication failure in more traditional journals.

Taken together, the formal publication of well-crafted and clever articles (e.g., this one on replication in BBS) seems to offer the best and most efficient way to reach a maximal audience and especially to convince as yet undecided individuals to, for example, join the replication movement in order to make replication mainstream, thereby providing one contribution (out of many possible ones) to psychology's renaissance (Nelson et al. 2018).

A pragmatist philosophy of psychological science and its implications for replication

doi:10.1017/S0140525X18000626, e127

Ana Gantman, Robin Gomila, Joel E. Martinez, J. Nathan Matias, Elizabeth Levy Paluck, Jordan Starck, Sherry Wu, and Nechumi Yaffe

Department of Psychology, Princeton University, Princeton, NJ 08544.
agantman@princeton.edu rgomila@princeton.edu joelem@princeton.edu jmatias@princeton.edu epaluck@princeton.edu jstarck@princeton.edu jueyuw@princeton.edu myaffe@princeton.edu
anagantman.com
www.robingomila.com
http://socialbyselection.wordpress.com/
https://twitter.com/natematias
www.betsylevypaluck.com
www.sherryjwu.com

Abstract: A pragmatist philosophy of psychological science offers to the direct replication debate concrete recommendations and novel benefits that are not discussed in Zwaan et al. This philosophy guides our work as field experimentalists interested in behavioral measurement. Furthermore, all psychologists can relate to its ultimate aim set out by William James: to study mental processes that provide explanations for why people behave as they do in the world.

A pragmatist philosophy of psychological science offers to the direct replication debate concrete recommendations and novel benefits that are not discussed in Zwaan et al. Pragmatism starts from the premise that "thinking is for doing" (Fiske 1992). In other words, pragmatic psychological theories investigate the mental processes that predict observable behavior within the "rich thicket of reality" (James 1907, p. 68). This philosophy guides our work as field experimentalists interested in behavioral measurement. Furthermore, all psychologists can relate to its ultimate aim set out by William James: to study mental processes that provide explanations for why people behave as they do in the world.

Recommendations. A pragmatist philosophy of science urges scientists to observe what behaviors emerge in the complexity of real life; it encourages active theorizing about individuals' contexts and the way that individuals construe or interpret them. Specifically, direct replications should research the context of the planned replication site (i.e., James's "thicket of reality") to determine when it is appropriate to use the precise materials of previous experiments and when researchers should translate materials at the new site so that they will replicate the original participants' construal (Paluck & Shafir 2017). Some methods for documenting context and adapting studies include well-designed manipulation checks, pretesting, reporting on the phenomenological experience of participants in any intervention, and collaboration with those who have actually implemented previous studies. An additional recommendation we propose is statistical: Investigators should statistically characterize the field, meaning that every study should report the amount of explained *and* unexplained variance of the treatment effect. In this way, replications and original findings can be explicitly situated by both the effect size and the amount of "noise" (e.g., from measurement error or unmeasured construal, context, and individual differences) that might help identify the source of differences across studies (Martinez et al. 2018).

Benefits. A pragmatist approach draws out the creativity and rigor of replication research. For example, when conducting a replication of a field experiment at a new site, the question of whether to use the same materials or to create translated (construal-preserving) materials arises. Field replications create the most obvious opportunities to develop rigorous standards that describe and compare research settings. These standards could be adopted

by researchers working in many settings. Researchers can break new ground by developing these methodological standards, as opposed to basing replication decisions on unstated assumptions about context similarity. Theorizing the context of a proposed replication also entails creative theoretical integration in our highly differentiated field; specifically, the integration of theories that pertain to context (to situation, identity, culture, and perception) with the focal theory that is to be tested with the replication. Additionally, reporting the total unexplained and explained variance from a study is an explicitly cumulative exercise aimed at meta-analysis. Emphasizing measurement as a point of comparison between studies also addresses the chronology problem (Zwaan et al., sect. 5.1.1) in which studies that are "first" to ask a particular question are prioritized over replications.

Field researchers, who regularly face the challenge of theorizing a broader context, may have a larger leadership role in developing conventions of direct replication than implied by Zwaan et al., who predict fewer replications of field versus laboratory studies. For example, in the digital space, replications of marketing and media experiments proceed at a scale that vastly outstrips normal academic research. These studies represent enormous opportunities to examine the impact of context on causal relationships (Kevic et al. 2017). In the policy world, Campbell's vision for the experimenting society (Campbell 1969; 1991) lays out steps for cost-efficient and politically feasible replication of studies across real-world settings. Such experiments feature contextual variation of deep theoretical importance, including differing levels of economic inequality, demographic diversity, and political contestation (for an example, see Dunning et al., in press). Finally, articles based on field experimental replications can be models of compelling scientific writing, combating claims that replication research is rote and boring, because field studies lend themselves to a rich description of place, participants, history, and more generally the psychological and behavioral equilibrium into which a social scientist intervenes (Lewin 1943/1997).

Don't characterize replications as successes or failures

doi:10.1017/S0140525X18000638, e128

Andrew Gelman

Department of Statistics, Columbia University, New York, NY 10027

gelman@stat.columbia.edu

http://www.stat.columbia.edu/~gelman

Abstract: No replication is truly direct, and I recommend moving away from the classification of replications as "direct" or "conceptual" to a framework in which we accept that treatment effects vary across conditions. Relatedly, we should stop labeling replications as successes or failures and instead use continuous measures to compare different studies, again using meta-analysis of raw data where possible.

I agree wholeheartedly that replication, or the potential of replication, is central to experimental science, and I also agree that various concerns about the difficulty of replication should, in fact, be interpreted as arguments in *favor* of replication. For example, if effects can vary by context, this provides more reason why replication is necessary for scientific progress. I also agree with the target article that it is an error when, following a disappointing replication result, proponents of the original published studies "irrationally privilege the chronological order of studies more than the objective characteristics of those studies when evaluating claims about quality and scientific rigor" (sect. 5.1.1, para. 3). As a remedy to this fallacy I have proposed a "time-reversal heuristic" (Gelman 2016b): the thought experiment of imagining the large, pre-registered replication study coming first, followed by the original, uncontrolled study.

It may well make sense to assign lower value to replications than to original studies, when considered as *intellectual products*, as we can assume the replication requires less creative effort. When considered as *scientific evidence*, however, the results from a replication could well be better than those of the original study, in that the replication can have more control in its design, measurement, and analysis.

It is also good to present and analyze all of the data from an experiment. Selection, forking paths, and researcher degrees of freedom have led us into the replication crisis, but these problems are all much reduced with analyses that use all of the data. Conversely, if we do not have access to raw data, many published results are close to useless, and when there is a high-quality pre-registered replication, I would be inclined to pretty much ignore the original paper, rather than, say, to assume the truth lies somewhere between the original and replication results.

Beyond this, I would like to add two points from a statistician's perspective.

First, the idea of replication is central not just to scientific practice but also to formal statistics, even though this has not always been recognized. Frequentist statistics relies on the reference set of repeated experiments, and Bayesian statistics relies on the prior distribution which represents the population of effects – and in the analysis of replication studies it is important for the model to allow effects to vary across scenarios.

My second point is that in the analysis of replication studies I recommend continuous analysis and multilevel modeling (meta-analysis), in contrast to the target article which recommends binary decision rules, which I think are contrary to the spirit of inquiry that motivates replication in the first place.

The target article follows the conventional statistical language in which a study is a "false positive" if it claims to find an effect where none exists. But in the human sciences, just about all of the effects we are trying to study are real; there are no zeros. See Gelman (2013) and McShane et al. (2017) for further discussion of this point. Effects can be hard to detect, though, because they can be highly variable and measured inaccurately and with bias. Instead of talking about false positives and false negatives, we prefer to speak of type M (magnitude) and type S (sign) errors (Gelman & Carlin 2014). Related is the use of expressions such as "failed replication." I have used such phrases myself, but they get us into trouble with their implication that there is some criterion under which a replication can be said to succeed or fail. Do we just check whether $p<.05$? That would be a very noisy rule, and I think we would all be better off simply reporting the results from the old and new studies (as in the graph in Simmons & Simonsohn 2015). If there is a need to count replications in a larger study of studies such as the Open Science Collaboration, I would prefer to do so using continuous measures rather than threshold-based replication rates.

The authors write, "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 United Kingdom or Japan, for that matter, then a replication carried out with these samples would be considered direct" (sect. 4, para. 3). The difficulty here is that theories are often so flexible that all these sorts of differences *can* be cited as reasons for a replication failure. For example, Michigan is colder than Florida, and outdoor air temperature was used as an alibi for a replication failure of a well-publicized finding in evolutionary psychology (Tracy & Beall 2014). Also there is no end to the differences between the United Kingdom and Japan that could be used to explain away a disappointing replication result in social psychology. The point is that any of these could be considered a "direct replication" if that interpretation is desired, or a mere "extension" or "conceptual replication" if the results do not come out as planned. In social psychology, at least, it could be argued that no replication is truly direct: society, and social expectations, change over time. The authors recognize this in citing Schmidt (2009)

and also in their discussion of why contextual variation does not invalidate the utility of replications; given this, I think the authors could improve their framework by abandoning the concept of "direct replication" entirely, instead moving to a meta-analytic approach in which it is accepted ahead of time that the underlying treatment effects will vary between studies. Rather than trying to evaluate "whether a study is a direct or conceptual" replication, we can express the difference between old and new studies in terms of the expected variation in the treatment effect between conditions.

That said, if the measurements in the original study are indirect and noisy (as is often the case) and it is impossible or inconvenient to reanalyze the raw data, the question is moot, and it can make sense to just take the results from the replication or extension studies as our new starting point.

Three ways to make replication mainstream

doi:10.1017/S0140525X1800064X, e129

Morton Ann Gernsbacher
Department of Psychology, University of Wisconsin, Madison, WI 53706
MAGernsb@wisc.edu
www.GernsbacherLab.org

Abstract: Zwaan et al. argue convincingly that replication needs to be more mainstream. Here, I suggest three practices for achieving that goal: Incremental Replications, which are built into each experiment in a series of experiments; Reciprocal Replications, which are reciprocal arrangements of co-replications across labs; and Didactic Replications, which are replications used for training.

Zwaan et al. provide convincing arguments for the value of replication – and the need to make replication practices mainstream in psychology. However, due most likely to limits of space rather than limits of vision, Zwaan et al. stop short of providing concrete steps researchers can take to make replication mainstream. Here, I suggest three practices researchers can adopt to better incorporate replication into their labs.

Incremental replications. We might think of replication as a practice that occurs in a separate lab with different researchers. Also the other two replication practices I will discuss can occur that way. But our own studies can also benefit from the verification of replication, within our own lab and within the same studies. An obvious step is to conduct exact replications within a series of experiments (e.g., "Experiment 2: Replication. We tested an additional 120 subjects using the same materials and procedures as we used in Experiment 1," and "Experiment 4: Replication. We tested an additional 120 subjects using the same materials and procedures as we used in [Experiment 3]," Gernsbacher & Hargreaves 1988, p. 704 and 706).

More parsimoniously, we can conduct, within the same study, what I am calling incremental replications. For example, in a series of experiments investigating how readers understand pronouns, I probed participants immediately before versus after they read a pronoun in one experiment. In another experiment, I again probed participants immediately after they read a pronoun, but in this second experiment I also probed them after they finished reading the entire sentence (Gernsbacher 1989). In this way, across experiments but within the same study, I tried to incrementally replicate each of the previously tested probe points (see also Garnham et al. 1996, for a similar approach).

As another example, in a series of priming experiments, we manipulated two types of primes in a first experiment and manipulated again one of those two prime types along with a different prime type in a second experiment (Gernsbacher et al. 2001a). In another series, we manipulated three prime types in a first experiment and repeated two of the three prime types across other experiments (Gernsbacher et al. 2001b). These incremental replications in within-subject designs also allowed us to assess the stability of our previous results in slightly different contexts (the value of which Zwaan et al. highlight).

Incremental replication is also valuable in between-subject designs. For example, in a series of between-subject treatment experiments, we repeated the baseline condition in subsequent experiments with other subjects (and juxtaposed with other treatments), which allowed us to assess the baseline condition's stability (Traxler & Gernsbacher 1992). In another series of experiments, we repeated the control condition in subsequent experiments, which allowed us to assess its stability (Traxler & Gernsbacher 1993).

Reciprocal replications. We might also think of replication as a practice that occurs only after a study has been peer reviewed. However, I would rather receive confirmation (or disconfirmation) of the stability of my results earlier rather than later. Zwaan (2017) in material left on the target article's editing floor, describes how this can be done.

> A research group formulates a hypothesis that they want to test. At the same time, they desire to have some reassurance about the reliability of the finding they expect to obtain. They decide to team up with another research group. They provide this group with a protocol for the experiment, the program and stimuli to run the experiment, and the code for the statistical analysis of the data. The experiment is preregistered. Both groups then each run the experiment and analyze the data independently. The results of both studies are included in the article, along with a meta-analysis of the results.

Zwaan (2017) calls this practice as concurrent replication, and my recommendation goes one step further: Make the process reciprocal. Lab A attempts to replicate Lab B's study, while Lab B is doing the same for Lab A's study. Platforms such as *StudySwap* (deemed "a Craigslist for researchers" by Nosek in Chawla 2017) and *Psychological Science Accelerator* are ideal for reciprocal replication. Reciprocal replications should take some of the adversarial sting out of traditional replications.

Didactic replications. Lastly, we can make replication more mainstream by embracing it as a training tool. When I was a first-year doctoral student, in one of my first meetings with my advisor, he walked to his filing cabinet, pulled out a recently published article, and suggested I spend my first semester trying to replicate the results. The fact that this didactic activity occurred nearly 40 years ago might be surprising. More surprising might be the fact that the first author of the study my advisor tasked me to replicate was, indeed, my advisor (Foss & Blank 1980).

As it turned out, the previous study only partially replicated (Foss & Gernsbacher 1983). Learning how to execute an experiment from a published article was an incredibly valuable training experience. (Most likely this is why beginning cooks are encouraged to follow a recipe precisely, before adding their own flourishes.) Deciphering why the previous study only partially replicated was an even more valuable training experience. I believe I learned more about experimental design, stimulus creation, and the myriad other steps involved in doing good science than I would have learned had I joined an in-process study or tried to generate a new study from scratch.

The didactic value of replication has been advocated by others, most notably Grahe and his "Collaborative Replications and Education Project" (Grahe et al. 2014). Along with Incremental Replications, which are replications built into each of a series of experiments to attempt to replicate parts of previous experiments, and Reciprocal Replications, which are reciprocal arrangements of co-replication, Didactic Replications can make replication more mainstream.

Three strong moves to improve research and replications alike

doi:10.1017/S0140525X18000651, e130

Roger Giner-Sorolla,[a] David M. Amodio,[b,c] and Gerben A. van Kleef[c]

[a]*School of Psychology – Keynes College, University of Kent, Canterbury, Kent CT2 7NP, United Kingdom;* [b]*Department of Psychology, New York University, New York, NY 10003;* [c]*Department of Social Psychology, University of Amsterdam, 1018 WS Amsterdam, The Netherlands*

rsg@kent.ac.uk david.amodio@gmail.com G.A.vanKleef@uva.nl
https://www.kent.ac.uk/psychology/people/ginerr/
http://amodiolab.org/
http://www.uva.nl/profile/g.a.vankleef/

Abstract: We suggest three additional improvements to replication practices. First, original research should include concrete checks on validity, encouraged by editorial standards. Second, the reasons for replicating a particular study should be more transparent and balance systematic positive reasons with selective negative ones. Third, methodological validity should also be factored into evaluating replications, with methodologically inconclusive replications not counted as non-replications.

Although we largely agree with Zwaan et al.'s analysis, we want to add to it, drawing on our experiences with replications as authors and editors. Over the past years in psychology, successful reforms have been based on concrete suggestions with visible incentives. We suggest three such moves that Zwaan et al. might not have considered.

Anticipate replication in design. In answering concerns about context variability, Zwaan et al. suggest that original authors' reports should be more detailed and acknowledge limitations. But these suggestions miss what lets us meaningfully compare two studies across contexts: calibration of methods, independent from the hypothesis test.

Often, suspicions arise that a replication is not measuring or manipulating the same thing as the original. For example, the Reproducibility Project (Open Science Collaboration 2015) was criticized for substituting an Israeli vignette's mention of military service with an activity more common to the replication's U.S. participants (Gilbert et al. 2016). All of the methods reporting in the world cannot resolve this kind of debate. Instead, we need to know whether both scenarios successfully affected the independent variable. Whether researchers have the skill to carry out a complex or socially subtle procedure is also underspecified in most original and replication research, surfacing only as a doubt when replications fail.

Unfortunately, much original research does not include procedures to check that manipulations affected the independent variable or to validate original measures. Such steps can be costly, especially if participant awareness concerns require a separate study for checking. Nevertheless, the highest standard of research methodology should include validation that lets us interpret both positive and negative results (Giner-Sorolla 2016; LeBel & Peters 2011). Although the rules of replication should allow replicators to add checks on methods, such checks should also be a part of original research. Specifically, by adopting the Registered Report publication format (Chambers et al. 2015), evaluation of methods precedes data collection, so that planning to interpret negative results is essential. More generally, publication decisions should openly favor studies that take the effort to validate their methods.

Discuss and balance reasons to replicate. Providing a rationale for studying a particular relationship is pivotal to any scientific enterprise, but there are no clear guidelines for choosing a study to replicate. One criterion might be importance: theoretical weight, societal implications, influence through citations or textbooks, mass appeal. Alternatively, replications may be driven by doubt in the robustness of the effect. Currently, most large-scale replication efforts (e.g., Ebersole et al. 2016a; Klein et al. 2014b; Open Science Collaboration 2015) have chosen their studies either arbitrarily (e.g., by journal dates) or by an unsystematic and opaque process.

Without well-justified reasons and methods for selection, it is easy to imagine doubt motivating any replication. Speculatively, many individual replications seem to be attracted by a profile of surprising results, weak theory, and methods. But if replications hunt the weak by choice, conclusions about the robustness of a science will skew negative. This problem is compounded by the psychological reality that findings that refute the status quo (such as failed replications) attract more attention than findings that reinforce the status quo (such as successful replications).

Replicators (like original researchers) should provide strong justification for their choice of topic. When replication is driven by perceptions of faulty theory or implausibly large effects, this should be stated openly. Most importantly, replications should also draw on selection criteria *a priori* based on positive traits, such as theoretical importance, or diffusion in the academic and popular literature. Indeed, we are aware of one attempt to codify some of these traits, but it has not yet been finalized or published (Lakens 2016).

Although non-replication of shaky effects can be valuable, encouragement is also needed to replicate studies that are meaningful to psychological theory and literature. Importance could be one criterion of evaluation for single replication articles. Special issues and large-scale replication projects could be planned around principled selection of important effects to replicate. The Collaborative Replications and Education Project (2018), for example, chooses studies for replication based on *a priori* citation criteria.

Evaluate replication outcomes more accurately. The replication movement also suffers from an underdeveloped process for evaluating the validity of its findings. Currently, replication results are reported and publicized as a success or failure. But "failure" really represents two categories: valid non-replications and invalid (i.e., inconclusive) research. In original research, a null result could reflect a true lack of effect or problems with validity (a manipulation or measure not being operationalized precisely and effectively). Validity is best established through pilot testing, manipulation checks, and the consideration of context, sample, and experimental design, and evaluated through peer review. If validity is inadequate, then the results are inconclusive, not negative.

Indeed, most replication attempts try hard to avoid inconclusive statistical outcomes, often allotting themselves stronger power than the original study. But there has not been as much attention to identifying inconclusive methodological outcomes, such as when a replication's manipulation check fails, or a method is changed in a way that casts doubts upon the findings. One hindrance is the attitude, sometimes seen, that direct replications do not need to meet the same standards of external peer review as original research. For example, the methods of the individual replications in Open Science Collaboration (2015) were reviewed only by one or two project members and an original study author, pre-data collection.

Conclusion and recommendations. Reasons for replicating a particular effect should be made transparent, with positive, systematic methods encouraged. Replication reports and original research alike should include evidence of the validity of measures and manipulations, with standards set before data collection. Methods should be externally peer reviewed for validity by experts, with clear consequences (revision, rejection) if they are judged as inadequate. Also, when outcomes of replication are simplified into "box scores," they should be sorted into three categories: replication, non-replication, and inconclusive. By improving the validity of replication reports, we will strengthen our science, while offering a more accurate portrayal of its state.

Making replication prestigious

doi:10.1017/S0140525X18000663, e131

Krzysztof J. Gorgolewski,[a] Thomas Nichols,[b,c,d] David N. Kennedy,[e] Jean-Baptiste Poline,[f,g] and Russell A. Poldrack[a]

[a]*Department of Psychology, Stanford University, Stanford, CA 94305;* [b]*Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Headington, Oxford OX3 7FZ, United Kingdom;* [c]*Welcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford OX3 7FZ, United Kingdom;* [d]*Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom;* [e]*Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA 01655;* [f]*Montreal Neurological Institute, McGill University, Montréal, QC H3A 2B4, Canada;* [g]*Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720.*

krzysztof.gorgolewski@gmail.com **thomas.nichols@bdi.ox.ac.uk**
david.kennedy@umassmed.edu **jbpoline@gmail.com**
russpold@stanford.edu
http://blog.chrisgorgolewski.org/
https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols
http://www.russpoldrack.org/

Abstract: Making replication studies widely conducted and published requires new incentives. Academic awards can provide such incentives by highlighting the best and most important replications. The Organization for Human Brain Mapping (OHBM) has led such efforts by recently introducing the OHBM Replication Award. Other communities can adopt this approach to promote replications and reduce career cost for researchers performing them.

Zwaan et al.'s otherwise excellent review of issues related to replications paints a rather bleak picture of credit given to researchers involved in replication studies. Even though the status quo is described accurately – that is, replications will not be valued as much as traditional research by hiring and tenure committees – tools and interventions exist that could change this situation ("Rewarding negative results," 2017). We recently introduced a new approach to incentivize replications and provide credit to researchers performing high-quality replications of seminal studies. Tapping into the long tradition of scientific awards, we have designed a Replication Award that can be implemented by scientific communities, journals, institutes or departments. In addition to providing a detailed protocol on how to effectively solicit submissions, score them, and announce the winner (Gorgolewski et al. 2017b) we have also implemented it within the Organization for Human Brain Mapping (OHBM). The first OHBM Replication Award was presented in June 2017 to Wouter Boekel for his thorough investigation of 17 brain-behaviour relationships (Boekel et al. 2015). This study found evidence confirming only one of the investigated relationships (correlation between real-world social network size and grey matter volume in the amygdala (Kanai et al. 2012). No evidence for previously reported effects was found for eight relationships, and results were inconclusive for the remaining eight.

The effectiveness of the OHBM Replication Award in the context of the promotion of replications has been evaluated by a community survey (Gorgolewski et al. 2017a). Of 226 respondents who were familiar with the award, 49% declared that it made them more likely to perform replications in the future, and 41% declared that it made them more likely to publish replication studies. Because prizes, awards, and other honours are often listed on curricula vitae and taken into consideration during hiring and tenure decision, we hope that this award will also have a positive influence on the careers of researchers performing replications.

We look forward to more organizations adopting Replication Awards following the lead of OHBM. The key to making replication mainstream is to provide incentives for researchers and to elevate replications to the status of a first class citizen among other scientific outputs. Replication Awards can provide such incentives a form that is already widely used in the context of evaluating careers. In the future, editorial boards and reviewers should adopt more proactive policies to facilitate the publication of replication studies (similar to *NeuroImage: Clinical* [Fletcher & Grafton 2013]). We also hope that new publishing initiatives such as the one initiated by the OHBM this year will facilitate this change in the future.

A Bayesian decision-making framework for replication

doi:10.1017/S0140525X18000675, e132

Tom E. Hardwicke,[a] Michael Henry Tessler,[b] Benjamin N. Peloquin,[b] and Michael C. Frank[b]

[a]*Meta-Research Innovation Center at Stanford (METRICS), Stanford School of Medicine, Stanford University, Stanford, CA 94305;* [b]*Department of Psychology, Stanford University, Stanford, CA 94305.*

tom.hardwicke@stanford.edu **mhtessler@stanford.edu**
bpeloqui@stanford.edu **mcfrank@stanford.edu**
https://tomhardwicke.netlify.com/
http://stanford.edu/~mtessler/
https://benpeloquin7.github.io/
https://web.stanford.edu/~mcfrank/

Abstract: Replication is the cornerstone of science – but when and why? Not all studies need replication, especially when resources are limited. We propose that a decision-making framework based on Bayesian philosophy of science provides a basis for choosing which studies to replicate.

Direct replications have an important place in our scientific toolkit. Given limited resources, however, scientists must decide when to replicate versus when to conduct novel research. A Bayesian viewpoint can help clarify this issue. On the Bayesian view, scientific knowledge is represented by a probability distribution over theoretical hypotheses, given the available evidence (Strevens 2006). This distribution, called the *posterior*, can be decomposed into the product of the prior probability of each hypothesis and the likelihood of the data given the hypothesis. Evidence from a new study can then be integrated into the posterior, making hypotheses more or less probable. The amount of change to the posterior can be quantified as a study's *information gain*. Using this formalism, one can design "optimal experiments" that maximize information gain relative to available resources (e.g., MacKay 1992). One attraction of this quantitative framework is that it captures the dictum that researchers should design studies that can adjudicate between competing hypotheses (Platt 1964).

Some good intuitions fall out of the Bayesian formulation. A study designed to select between *a priori* likely hypotheses (e.g., those well supported by existing data) can lead to high information gain. By contrast, a study whose data provide strong support for a hypothesis that already has a high prior probability, or weak support for a hypothesis that has a low prior probability, provides much less information gain. Larger samples and more precise measurements will result in greater information gain, but only in the context of a design that can distinguish high-prior-probability hypotheses.

Even well-designed studies can be undermined by errors in execution, reporting, or analysis. If a study is known to be erroneous, then it clearly leads to no information gain, but the more common situation is some uncertainty about the possibility of error. A Bayesian framework can capture this uncertainty by weighting information gain by a study's *credibility*. Concerns about questionable research practices, analytic error, or fraud thus all decrease the overall information gain from a study.

Direct replications are special in this framework only in that they follow a pre-existing set of design decisions. Thus, the main reason to replicate is simply to gather more data in a promising paradigm. In cases where the original study had low credibility or a small sample size, a replication can lead to substantial information gain (Klein et al. 2014c). Replicating the same design again and again will offer diminishing returns, however, as estimates of relevant quantities become more precise (Mullen et al. 2001). If a study design is not potentially informative, for example, because it cannot in principle differentiate between hypotheses, then replicating that design will not lead to information gain. Finally, when a particular finding has substantial *applied* value, replicators might want to consider an *expected value analysis* wherein a replication's information gain is weighted by the expected utility of a particular outcome.

Replications have one unique feature, though: They can change our interpretation of an original study by affecting our estimates of the original study's credibility. Imagine a very large effect is observed in a small study and an identical but larger replication study then observes a much smaller effect. If both studies are assumed to be completely credible, the best estimate of the quantity of interest is the variance-weighted average of the two (Borenstein et al. 2009). But if the replication has high credibility – for example, because of preregistration, open data, and son on – then the mismatch between the two may result from the earlier study lacking credibility as a result of error, analytic flexibility, or another cause. Such explanations would be taken into account by downweighting the information gain of the original study by that study's potential lack of credibility. Of course, substantial scientific judgment is required when sample, stimulus, or procedural details differ between replication and original (cf. Anderson et al., 2016; Gilbert et al. 2016). Often, multiple studies that investigate reasons for the failure of a replication are needed to understand disparities in results (see, e.g., Baribault et al. 2018; Lewis & Frank 2016; Phillips et al. 2015).

The prior probability of hypotheses will not be universally agreed upon and can lead to disagreements about whether a particular result should be replicated. One researcher may believe that a study with low information gain – perhaps on account of a small sample size – deserves to be "rescued" by replication because it addresses a plausible hypothesis. By contrast, a more skeptical researcher who assigned the original hypothesis a lower prior probability might see no reason to replicate. Or that researcher might replicate simply to convince others that the original study lacks credibility, especially in the case that it is influential within academia or the general public. Overall, as long as studies are appropriately conducted and reported, and all studies are considered, then the Bayesian framework will accumulate evidence and converge to an estimate of the true posterior.

Replication is an expensive option for assessing credibility, however. Assessing analytic reproducibility and robustness may be a more efficient means of ensuring that errors or specific analytic choices are not responsible for a particular result (Steegen et al. 2016; Stodden et al. 2016). Forensic tools like *p*-curve or the test for excess significance (Ioannidis & Trikalinos 2007; Simonsohn et al. 2014) can also help in assessing credibility.

How should an individual researcher make use of this Bayesian framework? When thinking about replication, researchers should ask the same questions they do when planning a new study: Does my planned study differentiate between plausible theoretical hypotheses, and do I have sufficient resources to carry it out? For a replication, this judgment can then be qualified by whether a re-evaluation of the credibility of the original study would be a net positive, because downweighting the credibility of an incorrect or spurious study also leads to overall information gain. Adopting such a framework to guide a rough assessment of information value (even in the absence of precise numerical assignments) can help researchers decide when to replicate.

Putting replication in its place

doi:10.1017/S0140525X18000687, e133

Evan Heit[a,1,2] and Caren M. Rotello[b]

[a]*E.H. Division of Research on Learning, Education and Human Resources Directorate, National Science Foundation, Alexandria, VA 22314;* [b]*CMR Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, MA 01003.*

ekheit@nsf.gov **caren@psych.umass.edu**

https://www.umass.edu/pbs/people/caren-rotello

Abstract: Direct replication is valuable but should not be elevated over other worthwhile research practices, including conceptual replication and checking of statistical assumptions. As noted by Rotello et al. (2015), replicating studies without checking the statistical assumptions can lead to increased confidence in incorrect conclusions. Finally, successful replications should not be elevated over failed replications, given that both are informative.

What is the theoretical value of direct replication? In a recent paper, we (Rotello et al. 2015) described several cases where oft-replicated studies repeated the methodological flaws of the original work. In particular, we presented examples from research on reasoning, memory, social cognition, and child welfare in which the standard method of analysis was not justified and indeed could – and in at least two cases, did – lead to erroneous inferences. Repeating the study, along with the flawed analyses, could lead to yet greater confidence in these incorrect conclusions. Most of our examples concerned conceptual rather than direct replications, in the sense that there were various purposeful design and material changes across studies. Our point was about methodology, namely that inferential errors as a result of unjustified analyses can be magnified upon replication. Contrary to the implication of the target article, we would not argue that the theoretical value of direct, or for that matter conceptual, replications is limited.

Indeed, the target article makes a compelling case for the value of replication, as well as its mainstream role in psychology. Yet we would not elevate replication over other worthwhile research practices. Using an example from Rotello et al. (2015), we reported that, beginning with Evans et al.(1983), for three decades, replication studies on the belief bias effect in reasoning have employed analyses such as analyses of variance on differences in response rates without checking the assumptions of those analyses. (In this example, researchers could easily do so by collecting data that would allow them to plot receiver operating characteristic curves to see whether there is a linear or curvilinear relationship between correct and incorrect positive response rates.) Checking statistical assumptions is another worthwhile research practice, the results of which sometimes will contraindicate the strategy of simply running the same analyses again. Researchers should place a high priority on checking the assumptions of their statistical analyses and their dependent measures. Just as the *Reproducibility Project: Psychology* (Open Science Collaboration 2015) has launched a highly successful effort to crowdsource direct replication, we note that other worthwhile research practices, such as checking statistics, could also be crowdsourced. In light of the potential problems with difference scores and analyses of variance that place so many reasoning and recognition memory studies at risk (see also Dubé et al. 2010; Heit & Rotello 2014; Rotello et al. 2008), we would like to see a large-scale effort to check statistical assumptions across of wide range of research domains. We point to statcheck (Nuijten et al. 2016) as a promising example along these lines, although its focus to date has been on checking *p* values. For some research domains, checking statistical assumptions may be a higher priority than direct replications.

Likewise, we would not elevate direct replication over conceptual replication. Philosophers of science have argued that researchers should be particularly confident in a conclusion that can be repeated across diverse contexts and methods (for a

review, see Heit et al. 2005). For example, Salmon (1984) described how early twentieth-century scientists developed a diverse set of experimental methods for deriving Avogadro's number (6.02×10^{23}). These methods included Brownian movement, alpha particle decay, X-ray diffraction, black body radiation, and electrochemistry. Together, these diverse methods – these conceptual replications – provided particularly strong support for the existence of atoms and molecules, going well beyond what direct replications could have accomplished. Turning back to psychology, we pose the question of whether the field learns more from *N* direct replications of a study or from *N* conceptual replications of the same study. Perhaps when *N* is very low there is greater value from direct replications, but as *N* increases the value of conceptual replications becomes more pronounced.

Finally, we would not elevate replication "successes" over replication "failures," namely, successes or failures in obtaining the same results as a prior study. Scientists learn something important from either outcome. This point is perhaps clearer in medical research—finding evidence that a once-promising medical treatment does not work should be just as important as a positive finding. To the degree that psychological research has an influence on health and medical practices, educational practices, and public policy, finding out which results do not replicate will be crucial. Although replication failures can be associated with fluctuating contexts and post hoc explanations, we note that in much research, context is varied purposefully from study to study. In a sense, context itself is an object of study, and failures are informative. Given that a drug is effective for men, does it work for women? Given that an educational intervention is successful for native English speakers, is it successful for English language learners? Here, addressing replication failures is central to the research enterprise rather than being a problematic matter.

To conclude, the pursuit of direct replication is potentially of high theoretical value, and indeed is becoming increasingly mainstream, for example, as psychology journals devote sections to direct replication reports. However, we would place direct replication alongside other worthwhile research practices, such as conceptual replication and careful evaluation of statistical assumptions. Likewise, we would place successful replications alongside failed replications in terms of their potential to inform the field.

NOTES

1. Parts of this commentary are a work of the U.S. Government and are not subject to copyright protection in the United States.

2. This material includes work performed by Evan Heit while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Bayesian belief updating after a replication experiment

doi:10.1017/S0140525X18000699, e134

Alex O. Holcombe[a] and Samuel J. Gershman[b]

[a]*School of Psychology, University of Sydney, Sydney, NSW 2006, Australia;* [b]*Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138.*

alex.holcombe@sydney.edu.au gershman@fas.harvard.edu

http://sydney.edu.au/science/people/alex.holcombe.php

gershmanlab.webfactional.com

Abstract: Zwaan et al. and others discuss the importance of the inevitable differences between a replication experiment and the corresponding original experiment. But these discussions are not informed by a principled, quantitative framework for taking differences into account. Bayesian confirmation theory provides such a framework. It will not entirely solve the problem, but it will lead to new insights.

If they could, researchers would design critical experiments that isolate and test an individual hypothesis. But as Pierre Duhem (1954) pointed out, the results of any one experiment depend not only on the truth or falsity of the central hypothesis of interest, but also on other "auxiliary" hypotheses. In the discussion of replications in psychology, this basic fact is understood (although typically referred to using different terminology).

For an experiment in physics, the auxiliary hypotheses might include that the measurement device is functioning correctly. For an experiment in psychology, in addition to correct functioning of the measurement devices (such as computers), the auxiliaries might include that the participants understood the instructions and, for a replication study, that the effect exists in a new population of participants.

Bayes' theorem dictates how one should update one's belief in a hypothesis as a result of new evidence. However, because the results of actual scientific experiments inevitably depend on auxiliary hypotheses, as well as the hypothesis of interest, the only valid use of Bayes' rule is to update one's belief in an undifferentiated compound of the central hypothesis and all of the auxiliaries. But researchers are interested primarily in the credibility of a central hypothesis, not the combination of it with various auxiliaries. How, then, can data be used to update one's strength of belief in a particular hypothesis/ phenomenon?

The philosopher Michael Strevens used probability theory to answer this. The equation he derived prescribes how, after the results of an experiment, one should update the strength of one's belief in a central hypothesis and the auxiliary hypotheses (Strevens 2001). For a replication experiment, a person's relative strength of belief in the central hypothesis and the auxiliary hypotheses involved determine how one should distribute the blame for a replication failure, with (typically) different amounts going to the auxiliary hypothesis and the central hypothesis. If one has a strong belief in the central hypothesis but a relatively weak belief in the auxiliaries of the replication experiment, then belief in the central hypothesis can and should emerge relatively unscathed.

Strevens' equation for distributing credit and blame emerged from a broader philosophy of science called Bayesian confirmation theory (Hawthorne 2014; Strevens 2017). One might disagree with Bayesian confirmation theory broadly, but still agree that Bayesian belief updating is the ideal in many circumstances.

Strevens' equation provides a way to quantify the evidence for an effect provided by a replication experiment. Many articles on replication, including Zwaan et al., have extensive discussions of the importance of and nature of differences between a replication experiment and an original study, but without a principled, quantitative framework for taking these differences into account. The dream of Bayesian confirmation theory is that scientific inference might proceed like clockwork – in certain circumstances. There will be difficulties with identifying and precisely quantifying the credence of auxiliary hypotheses, but even rough approximations should lead to insights.

Currently, whether scientists' actual belief updating bears much resemblance to the updating prescribed by Strevens' equation is very unclear. Many believe that after a failed replication, researchers often engage in motivated reasoning and irrationally cling to their beliefs, but Strevens' equation indicates that maintaining a strong belief and blaming auxiliaries is rational if one had not put much credence in the auxiliaries of a replication study.

One possible way forward is to implement pilot programs to induce scientists to set out their beliefs before the data of a replication study are collected. In particular, researchers should be asked to specify the strengths (as probabilities) of their beliefs in the central and auxiliary hypotheses of the study. After the data come in, Strevens' equation would dictate how these researchers

should update their beliefs. If it turns out that researchers do not update their beliefs in this way, we will have learned something. These findings, and the comments of the researchers on why they differed from Strevens' prescription (if they do), should illuminate how science progresses and how researchers reason.

Such a program may also help to pinpoint the disagreements that can occur between original researchers and replicating researchers. Presently, after a failed replication, a common practice is for authors of the original study to write a commentary. Frequently, the commentary highlights differences between the replication and the original study, sometimes without giving much indication of how much the authors' beliefs have changed as a result of the failed replication. This makes it difficult to determine the degree of disagreement on the issues.

Our proposal is closely related to several proposed reforms in the literature (and already in the Registered Replication Reports now published by *Advances in Methods and Practices in Psychological Science*, replicating labs are routinely asked what they expect the effect size to be). The key point is the addition of a suitable quantitative framework. Zwaan et al. mention the "Constraints on Generality" proposal of Simons et al. (2017) that authors should "spend some time articulating theoretically grounded boundary conditions for particular findings" as this would mean disagreements with replicating authors "are likely to be minimized" (sect. 4, para. 11). But it may be difficult for an author to testify that a result should replicate in different conditions, as she is likely to be uncertain about various aspects. Rather than making a black-and-white statement, then, it may be better if the author communicates their uncertainty by attaching subjective probabilities to some of the auxiliary hypotheses involved. A further benefit of this system would be that authors, and the theories they espouse, would then develop a track record of making correct predictions (Rieth et al. 2013).

We recognize that in many circumstances, it may not be realistic to expect researchers to be able to quantify their confidence in the hypotheses that are part and parcel of an original experiment and potential replication experiments. Areas that are less mature, in the sense that many auxiliary hypotheses are uncertain, may be especially poor candidates. But other areas may be suitable. There are good reasons for researchers to try.

An argument for how (and why) to incentivise replication

doi:10.1017/S0140525X18000705, e135

Piers D. L. Howe and Amy Perfors

School of Psychological Sciences, University of Melbourne, VIC 3010, Australia.

pdhowe@unimelb.edu.au amy.perfors@unimelb.edu.au
https://www.findanexpert.unimelb.edu.au/display/person340666
http://psychologicalsciences.unimelb.edu.au/research/chdh/ccs

Abstract: Although Zwaan et al. (2018) have made a compelling case as to why direct replications should occur more frequently than they do, they do not address how such replications attempts can best be encouraged. We propose a novel method for incentivising replication attempts and discuss some issues surrounding its implementation.

Zwaan et al. (2018) convincingly argue that replication attempts should become mainstream, but they say little as to how this can best be achieved. The problem is that there are currently few mechanisms in place to encourage replication attempts. For example, a survey conducted in 2015 found that only 3% of psychology journals explicitly state that they will consider publishing replications (Martin & Clarke 2017). Although there have been some notable attempts to encourage more replications (Klein et al. 2014a; Open Science Collaboration 2015), they have been of limited scope, and replications remain scarce: A survey of the top 100 psychology journals found that only 1% of reported studies involved replication (Makel et al. 2012). Given the enormous publication pressures on academics, if replications are rarely publishable, then a mainstream culture of replication will not emerge.

Here, we propose a novel solution to this problem: Make it standard practice for journals to pre-commit to publishing adequately powered, technically competent direct replications (at least in online form) for any article they publish and link to it from the original article. This would be comparatively simple to implement and would have a relatively low cost, but would greatly change the incentive structure for researchers. It would also lead to a virtuous cycle in which the more replications are published, the more other people would be encouraged to perform replications of their own. Indeed, performing replications might become an important part of academic training: Running replications would enable early postgraduate students to gain valuable skills in research implementation and analysis while also contributing to the scientific literature.

If our proposal were to be adopted, one expectation might be that authors of the original article would discuss the extent to which they predict that their findings would replicate. For instance, authors might become more explicit in identifying when they believe that their findings are likely to apply only to a particular demographic or to occur only in particular circumstances. These discussions would enhance the interpretability of the original article and encourage authors to think more clearly about these issues during the design and analysis of their studies.

Why should journals adopt our proposal? We suggest that a simple modification to the calculation of impact would encourage journals to publish replications of original articles, regardless of how those replications turn out. Currently, the Thomas Reuters journal's impact factor is determined by the number of citations of that journal within a designated period, divided by the number of citable documents published overall during that period. Importantly, the denominator does not include documents considered to be "Editorial Material" — a term covering a wide range of document types from true editorials to commentaries such as this one (even when the commentaries report original data). It should be comparatively simple to agree that non–peer-reviewed, online-only, direct replication attempts should also not count toward the denominator. If so, then hosting direct replication attempts on the journal's website would never hurt. Indeed, if these replication attempts could still be cited (just like editorials can be cited), they would only increase the journal's impact factor. This creates an incentive for journals to publish replications, which is a necessity for replications to become mainstream.

What about funding agencies? Like journals, grant agencies greatly value novelty, but they even more greatly value reliable science; a novel finding can have a long-term impact only if it is true. It should, therefore, be in a funding body's interest either to offer grants that are focused solely on replication or to mandate that a certain percentage of each grant be devoted to replicating previous research.

In one sense, our suggestion is a minor alteration in how science is traditionally done but, in another sense, it is a paradigm shift in how to evaluate scientific work. Although novelty and originality are clearly vital, replicability is no less important. Our failure to systematically replicate our findings results in biased estimates of effect sizes, hampers future work, and makes it hard to obtain a realistic evaluation of what we know (Anderson et al. 2017). Because the best way to obtain accurate estimates of a finding's effect size and robustness is to combine multiple independent replication attempts, we need to actively encourage replications. Within our paradigm, the initial publication of an article is just the starting point in an extended conversation that will conclude with a multitude of replication attempts, an increasingly accurate estimate of the effect size, and a much greater understanding of the circumstances for which the findings hold.

How might we appropriately acknowledge replication attempts for the purposes of career advancement? One obvious possibility would be to adopt a convention on curricula vitae in which replication attempts are classified as distinct from other types of publications — much as books, journal articles, and conference proceedings are classified separately now. It would then be up to the individual's university, grant review panel or promotion committee to decide how much to value replication attempts relative to other forms of publication.

Our proposal represents a "win" for academics, journals, and the progress of science as a whole. The ability to easily publish replications would mean that academics would be incentivised to perform replications. Indeed, doing so may become a routine and accepted part of academic training. Within a culture of replicability, the impact of any single replication failure would diminish, making replications less personally threatening and simply part of the process (much as reviews are part of science now). Journals would increase in prestige and citation rates by publishing replications. Fundamentally, incentivising replication attempts is the only way to achieve a mainstream culture of replicability. It is vital for our future that science is built on truth rather than sand.

How to make replications mainstream

doi:10.1017/S0140525X18000717, e136

Hans IJzerman,[a] Jon Grahe,[b] and Mark J. Brandt[c]

[a]*Department of Psychology (LIP/PC2S), Université Grenoble Alpes, BP 47-38040, Grenoble Cedex 9, France;* [b]*Department of Psychology, Pacific Lutheran University, Tacoma, WA 98447;* [c]*Department of Social Psychology, Tilburg University, Tilburg 5000 LE, The Netherlands.*

h.ijzerman@gmail.com **graheje@plu.edu**
m.j.brandt@tilburguniversity.edu
www.hansijzerman.org
www.tbslaboratory.com

Abstract: Zwaan et al. integrated previous articles to promote making replications mainstream. We wholeheartedly agree. We extend their discussion by highlighting several existing initiatives – the *Replication Recipe* and the *Collaborative Education and Research Project* (CREP) - which aim to make replications mainstream. We hope this exchange further stimulates making replications mainstream.

Zwaan et al. integrated previous articles to promote making replications mainstream. We wholeheartedly agree. We extend their discussion by highlighting several existing initiatives that aim to make replications mainstream and that have already helped resolve several of the concerns discussed by Zwaan et al. Specifically, we discuss how to *Increase Replication Quality* and how to *Make Replications Habitual*. These facets should facilitate addressing the concerns of not having a standard method and that expertise of the original and replication authors may not be sufficiently relevant.

Increasing replication quality. Zwaan et al. discussed criticisms of the limited theoretical value of replication and the role of contextual variable in replications. This criticism stems from a well-known discussion in psychology whether quality of research should be results- or theory-centered (e.g., Greenberg et al. 1988; Greenwald et al. 1986). One strategy to resolve the conflict between theoretical value and obtained results is to follow the guidelines outlined in the *Replication Recipe* (RR; Brandt, IJzerman et al. 2014). The RR suggests that replications include 36 "ingredients" for high-quality replications (including, but not limited to, choosing a finding with high replication value, sufficient power, exclusion criteria that are defined *a priori*, identified differences between original and replication studies, and preregistration). Following the RR helps replication researchers identify the central parameters of a study and thus the key components of the replication, so that the replication is as convincing as possible. This not only facilitates communication between original and replication researchers, but also between readers of both the replication and the original research. The RR, for example, suggests that replication researchers list contextual features that likely differ between the original and replication research (e.g., Different cultural setting? Different population?). This helps communicate to the original authors and readers what the differences in the studies are and the degree the study is a direct or more of a conceptual replication. There may not always be agreement on these designations, but at least the information is clearly available for the reader to make up their own mind. The results from the RR can also be used by future scholars to identify (and then test with preregistered studies) potential moderators of the effect across both original studies and replication studies, increasing the theoretical value of replications.

Interestingly, Zwaan et al. misinterpreted the RR as something that should be included in original articles. Our original paper was focused on replications and so we did not discuss original articles, but this misinterpretation highlights the important point that many, if not all, of the qualities of a convincing and high-quality replication are exactly the same as the qualities of a convincing and high-quality original study. Therefore, authors can specify the conditions they consider necessary and relevant for their finding and any limits on generalizability (Simons et al. 2017), resulting in increasingly specified psychological theories.

Making replications habitual. Another key facet to making replication mainstream is making replications habitual. One way of doing so is by developing an appreciation for replication early in the academic career. We created the Collaborative Education and Research Project (CREP; Grahe et al. 2015) with the goal of training undergraduate researchers to conduct high quality replication research through standardized procedures as part of research methods courses. The CREP board selects - through a rigid selection process - impactful studies that are feasible to conduct by bachelor students. Prior to data collection, the CREP board communicates with original authors that we selected their study and invite them to provide any original materials and to comment about any conditions that would facilitate successful replication. Students – often in groups and always under the supervision of a faculty supervisor– create a project page on the Open Science Framework, submit their proposed protocol (including video, methods, and evidence of international review board approval) for review by a CREP review team (three advanced researchers and a student administrative advisor). This review process is at least as stringent (and perhaps sometimes more so) than the journal review process. After receiving approval, they complete a general registration of their study, and then collect data. Upon project completion, they go through a second review where the CREP review team reviews their presentation of their data and findings.

CREP projects directly contribute to the research literature by reporting high-quality replications (with one manuscript published [Leighton et al. 2018] and two more in progress [Ghelfi et al. in preparation; Wagge et al. in preparation]). Additionally, and more importantly, the CREP educates students about modern psychological research methods, training them to be the researchers with the relevant expertise we need. These skills transfer to original research. Students must understand the hypothesis and theory from the original study as they identify which materials are necessary in an original study. They learn to properly document a study (including, but not limited to, obtaining informed consent, collecting and analyzing data, and reporting findings requires the same resources as original research). Further, by interacting with the CREP team, these students experience a review process with faculty at different institutions than their own. As a bonus, instructors are not challenged with reading and supervising poorly conceptualized or poorly planned research that is developed quickly, without adequate preparation that can understandably be typical of students' first research project.

Zwaan et al. integrated perspectives on replication to argue why replications should be made mainstream. The initiatives we describe have and we hope will continue to help make replications mainstream. Over the course of 5 years, 233 RRs have been registered on the Open Science Framework and 356 students at 49 institutions started 106 CREP replications. The RR and the CREP have already substantively contributed to increased replication quality and to making replications habitual. We hope this exchange further stimulates making replications mainstream.

Why replication has more scientific value than original discovery

doi:10.1017/S0140525X18000729, e137

John P. A. Ioannidis

Departments of Medicine, Health Research and Policy, and Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305; Department of Statistics, Stanford University School of Humanities and Sciences, 94305; Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA 94305.

jioannid@stanford.edu

https://profiles.stanford.edu/john-ioannidis

Abstract: The presumed dominance of "original discovery" over replication is an anomaly. Original discovery has more value than replication primarily when scientific investigation can immediately generate numerous discoveries most of which are true and accurate. This scenario is uncommon. A model shows how original discovery claims typically have small or even negative value. Science becomes worthy mostly because of replication.

The plea of making direct replication mainstream by Zwaan et al. seeks to return the scientific method to its roots. Emphasis on direct replication heralded the origins of science. In the Royal Society (established in 1660), early experimenters replicated in front of their colleagues whatever claims they made.

In the modern era, replication fell out of favor. Scientists publish articles that one must trust almost with blind faith. Typically, it is impossible even to perform a re-analysis of the published data (because data, protocols, and scripts are unavailable), let alone replicate in new studies. The implicit assumption is that the large majority of these claimed discoveries are right. If science faces a flood of major discoveries, using resources to replicate or to increase methodological scrutiny is low priority. Similarly, "negative" results seem low priority, if true, important, significant results abound.

The narrative of an oversupply of major, true discoveries and few false findings is untenable (Ioannidis 2005). Sometimes this narrative coexists with additional urges to rush (e.g., "patients are dying, license new drugs immediately"), avoid replication, and even be methodologically sloppy (e.g., use routinely collected data instead of randomized trials).

Clearly, genuine discoveries do occur, but they are inhomogeneous across different scientific fields. Science is a difficult endeavor. Numerous biases affect different designs and disciplines (Fanelli et al. 2017). Even with best intentions, getting wrong or exaggerated (Ioannidis 2008) findings is easy. Many fields make little progress, but they still keep pouring thousands of peer-reviewed papers. Peer review without checking data, protocols, and methods is superficial at best. What we do get, predictably, is an oversupply of statistically significant results; for example, 96% of the papers using *P* values in the biomedical literature have statistically significant *P* values (Chavalarias et al. 2016). However, very few of these millions of papers with seemingly significant and novel results translate into something useful in medicine (Bowen & Casadevall 2015), and similar inefficiencies probably occur in many other fields. The overblown rhetoric of "success" (e.g., the repeated unfulfilled claims of eliminating cancer or Alzheimer's disease) makes science seem untrustworthy in public.

If genuine discoveries are not that torrential, then replication has more value that an original discovery claim. If the proportion of false original findings is relatively high and/or if false findings have substantial consequences (e.g., if they lead researchers and translators of research down the wrong path wasting effort and resources, or if they cause harm for patients or other "users"), then original discovery work may even have negative value.

Let R be the prestudy odds for a research finding, BF the Bayes factor conferred by the discovery study, and h the ratio of the weight of negative consequences from a false-positive (FP) discovery claim versus the positive consequences from a true-positive (TP) discovery. The value of the original research is proportional to $TP-(h*FP)$ or equivalently to $(TP/FP)-h=(R*BF)-h$. Given that R and h are rather field specific and cannot be modified (unless a researcher moves to a different research field or entirely changes investigative strategy), investigators can increase the value of their discoveries mostly by increasing BF, for example, by running larger studies and ensuring greater protection from biases. To avoid negative values, one needs $BF>h/R$. Often this is difficult. With limited resources, most original discoveries come from small studies, where biases are common. A P value slightly less than 0.05 corresponds typically to a BF less than 3 (Benjamin et al. 2017), and biases erode this further. Furthermore, most current research areas lack obvious low-hanging fruit (i.e., domains with high values of R).

In these circumstances, replication can easily have more value than original research, provided that it has reasonable ability to help differentiate eventually what is true and what is false, when properly done, or to identify the proper boundaries where some claimed phenomena hold true.

In some fields, even replications cannot have positive value. Then their conduct in a field where original research already had negative value would not make things better. For example, observational nutritional epidemiology of single nutrients affecting health outcomes may belong to this category (Ioannidis 2013a). Biases in this field are far stronger than whatever true signals may exist. Replication efforts using the same biased observational methods would just add negative value. Such fields should simply be abandoned, acknowledging that the methods that we possess cannot yield positive value. Efforts should be diverted to other methodological strategies (e.g., randomized trials) and other fields of investigation. However, for most fields, it is reasonable to expect that well-done, carefully executed, pre-registered, transparent direct replications in an environment of high methodological standards (Munafò et al. 2017) would have positive value and would help correct the mess of original discovery research. Some fields of modern science, such as population genetics, have learned that replication matters more than discovery (Ioannidis 2013b). We hope this will become more widely recognized.

Replication may have more value than original research even in situations where indeed many discoveries are made and a large proportion reflect true signals. It could still be the case that these true discoveries are then difficult to prioritize for further steps (e.g., translation, application, implementation in the real world to reap benefits) unless the magnitudes of their effects are known with substantial accuracy and their context-sensitivity is well understood. For example, it could be that 200 technologies or tentative interventions are discovered, but implementing all of them is impossible. Understanding which ones are best would require narrowing the uncertainty around the relative benefits and harms of each proposed discovery. This would require large-scale replication evidence. It would also require testing in diverse settings to understand better the architecture of the heterogeneity of the observed effects. It is unlikely that the original study alone will offer such insights, regardless of how well it is designed and executed (IntHout et al. 2016). Again, replication would matter more than the original discovery study. Eventually, the cumulative evidence, comprising both the original and the replication studies, may be more informative than either component.

Introducing a replication-first rule for Ph.D. projects

doi:10.1017/S0140525X18000730, e138

Arnold R. Kochari[a] and Markus Ostarek[b]

[a]*Institute for Logic, Language and Computation, University of Amsterdam, 1090 GE Amsterdam, The Netherlands;* [b]*Max Planck Institute for Psycholinguistics, Nijmegen, 6500 AH Nijmegen, The Netherlands*

a.kochari@uva.nl markus.ostarek@mpi.nl

http://akochari.com/

http://www.mpi.nl/people/ostarek-markus

Abstract: Zwaan et al. mention that young researchers should conduct replications as a small part of their portfolio. We extend this proposal and suggest that conducting and reporting replications should become an integral part of Ph.D. projects and be taken into account in their assessment. We discuss how this would help not only scientific advancement, but also Ph.D. candidates' careers.

Commenting on the role that replications should play in a researcher's career, Zwaan et al. briefly suggest that early career researchers should conduct replications "with the goal of building on a finding or as only one small part of their portfolio" (sect. 5.5.1, para. 4). Extending this, we propose that conducting and reporting replications should become an integral part of Ph.D. projects and should be taken into account in their assessment. Specifically, we suggest adopting a *replication-first rule*, whereby Ph.D. candidates are expected to first conduct a replication when they are building on a previous finding, and only then collect data in their novel study.

One reason we consider it important to specifically address the role of replications for early career researchers is that they face enormous pressure to establish themselves in the scientific community and often fear that their careers could end before they really begin (Maher & Anfres 2016; "Many junior scientists" 2017). Currently, to secure a job in academia after obtaining a doctoral degree, one needs to build an impressive portfolio of publications (Lawrence 2003). Based on our observations of how research projects are carried out in practice, Ph.D. candidates often directly attempt innovative extensions of previous experimental work in the hope of answering a novel research question, because novelty strongly increases publishability (Nosek et al. 2012). When such extensions fail to produce the expected results, they tend to collect more data in several variations of their own experiments before turning to examine the replicability of the original effect. However, it may often turn out that they cannot reproduce the original finding, possibly because the original effect is, in fact, not robust. In these cases, replicating the original effect first would prevent what may turn out to be a substantial waste of time and resources on follow-up experiments. Moreover, the time saved as a result of replicating first can be used to further examine the robustness of the original effect, for example, by conducting an additional high-powered replication. Such replications contribute to a better estimate of effect sizes, which are currently often overestimated on account of publication bias, sampling error, or p-hacking (Fanelli 2011; Ferguson & Brannick 2012; Szucs & Ioannidis 2017a). As such, replications constitute an important scientific contribution and should be regarded as such by Ph.D. project advisors.

The above arguments demonstrate the advantages of replicating first in the case of a failed replication. Likewise, successful replications provide a great opportunity. Pressure to publish operating simultaneously with publication bias means that early career researchers are currently pressed to obtain specifically *positive* findings to publish papers. As a result, in our experience, not knowing whether an experiment will yield positive results causes anxiety in Ph.D. candidates. Incorporating replications as a first step of any new research project can help alleviate this anxiety. If an extension shows no effect or supports the null hypothesis after a successful replication of the original effect, it should be easier to interpret the theoretical significance of this outcome. For example, suppose that one replicates a previously observed priming effect but does not obtain it when the primes are masked. In this case, one can directly compare the effect in both conditions and make a convincing case about the role of visibility for the effect. These two experiments can likely be put together in a strong paper. Similarly, a successful replication and extension make for a solid package that will convince Ph.D. candidates themselves and the fellow researchers who read their work. In this way, replicating first shifts the focus from the *results* to the *underlying scientific process* (how well the work is carried out). In combination with the registered reports format (Chambers 2013), we believe a replication-first rule would minimize Ph.D. candidates' stress caused by the anticipation of negative results and increase the quality of their work.

Finally, we hope that adopting the proposed replication-first rule would result in an important shift in the necessity for early career researchers to learn and demonstrate the ability to conduct replications appropriately. Specifically, evaluating the outcome of replications often involves assessing the strength of accumulated evidence using state-of-the-art meta-analytic tools. We hope demonstration of such skills will increasingly be taken into account in quality assessment of theses and in hiring decisions. Widespread application of the replication-first rule would also generate pressure on graduate schools to organize corresponding courses and seminars.

Even though adopting the replication-first rule may be difficult in cases where data collection is costly for the budget or resources available for a Ph.D. project, this should not be seen as a sufficient reason to omit replications, as also pointed out by Zwaan et al. Because such studies often have smaller sample sizes and more room for arbitrary data analysis choices, replicability is an even larger issue for them (see Poldrack et al. [2017] for a discussion of this for fMRI findings). The growing awareness of this state of affairs in the field will likely lead to greater appreciation and higher rewards for replication in these cases. Ph.D. candidates are thus well advised to go the extra mile and replicate first. If two separate experiments are not feasible, incorporating a replication into the novel study design would be an option.

In sum, we believe that adopting the replication-first rule for Ph.D. projects would not only contribute to scientific progress in the way Zwaan et al. lay out, but also would be beneficial for the Ph.D. candidates themselves. We predict that this will result in a larger number of solid findings and publishable papers, as well as incentivize Ph.D.'s to master the necessary meta-analytic statistical tools for assessing evidence in cumulative science. In this way, we believe conducting replications could be a great boost for early researchers' careers rather than only a "service to the field." That said, we of course do not suggest obliterating the value of creativity and original thinking in doctoral theses and their assessment. The replication-first rule is intended as a constant reminder that a balance between the two is needed to ensure solid science.

Selecting target papers for replication

doi:10.1017/S0140525X18000742, e139

Anton Kuehberger[a] and Michael Schulte-Mecklenbeck[b,c]

[a]*Centre for Cognitive Neuroscience, University of Salzburg, Salzburg 5020, Austria;* [b]*Department of Business Administration, University of Bern, Bern 3012, Switzerland;* [c]*Max Planck Institute for Human Development, Berlin 14195, Germany.*

anton.kuehberger@sbg.ac.at michael@schulte-mecklenbeck.com

https://ccns.sbg.ac.at/people/kuehberger/

http://www.schulte-mecklenbeck.com

Abstract: Randomness in the selection process of to-be-replicated target papers is critical for replication success or failure. If target papers are

chosen because of the ease of doing a replication, or because replicators doubt the reported findings, replications are likely to fail. To date, the selection of replication targets is biased.

Altough running a replication study is difficult to impossible in some domains, it is quite easy in others. Zwaan et al. (2017) state, in their concern III (sect 5.3), that direct replications are not feasible in some domains, for example, large-scale observational studies, or even not possible, for example, for studies capitalizing on rare events. This argument pertains to domains, but more directly to journal articles, that is, experimental studies. We argue that, beyond the larger obstacles described above, studies that are easier to replicate are indeed more frequently replicated, which introduces selection bias into the replication enterprise.

The selection of a to-be-replicated experiment often depends on how easy it is to do a direct replication. An instructive example is the multilab preregistered replication of the ego-depletion effect (Hagger et al. 2016). The authors selected, as the target of their replication, a procedure introduced by Sripada et al. (2014) and not by the original authors of ego depletion (Baumeister et al. 1998; note that Baumeister recommended the alternative procedure!). The reason for selecting this procedure was described as follows: "tasks used in the original experiments were deemed too elaborate or complex to be appropriate for a multilab replication" (Sripada et al. 2014, p. 548). Why is selection bias at work here? Ease of application is frequently related to the quality of manipulating the independent variable, such that the strength of the manipulation is often limited in easy-to-administer operationalization. A vicious circle is generated: Ease of application breeds a multitude of primary studies (e.g., using simple procedures like questionnaires or vignettes). Many of these studies lead to significant results and therefore publication, but often they are false positives and effect sizes are overestimated (e.g., because of publication bias). If such studies become predominantly targets for direct replication, these replications have little power and are doomed to fail. We end up with many failed replications that are also published, even using the new gold standards of preregistration and multilab collaborations. This circle artificially increases the number of replication failures. The choice of the to-be-replicated target study thus is crucial.

Another selection criterion could be even more harmful: doubt. Many papers become replication targets, not because they are theoretically interesting or important, but because other researchers doubt their results. If there is something to researchers' intuitions of whether a result is likely to be true or not, less likely results have a lower base rate to be true. Even after a significant result, the posterior probability of the hypotheses tested in studies with a small prior is low. Selecting "doubted studies" as targets for direct replication also is doomed to result in failure under most definitions of successful replication. Again, if people select replication targets because they doubt the original findings, and if their doubt is reasonable, the literature will be filled with many failed replications.

The process for choosing the to-be-replicated target study thus is crucial. Ease of application and doubt may contribute to the selection of target papers, leading to an overestimation of replication failures. The best way to avoid this is random selection of replication targets. We pick the Replication Project: Psychology (Open Science Collaboration 2015) as an example of such a procedure. However, inspection of this selection process reveals a variety of judgments, deviating from a purely random choice. The decision tree in Figure 1 illustrates the selection that cuts down an overall 488 articles in the 2008 issues of three journals (*Psychological Science; Journal of Experimental Psychology: Learning, Memory and Cognition; Journal of Personality and Social Psychology*) to an ultimate 100 completed replications.

We identify the following nonrandom selections in the Replication Project: Psychology: (a) publication (only published papers are included); (b) year (papers published in 2008); (c) journal (*Psychological Science, Journal of Experimental Psychology: Learning, Memory and Cognition, Journal of Personality and Social Psychology*); (d) type (488 original research papers); (e) eligibility (158/488, i.e., 32.4%); (f) claim (113 of 158 claimed by replicators, i.e., 71.5%); (g) completion (100 of 113 papers completed and data uploaded to the Open Science Framework within given time frame, i.e., 88.5%). Eventually, a fifth (100/488, i.e., 20.5%) of all possible replications were run and ultimately published. Bias caused by the difficulty of doing a replication surely exists for eligibility (step e, see description in Open Science Collaboration 2015, Methods appendix) and is likely for claiming (step f) and completion (step g). Bias caused by doubt influences claiming (step f). In

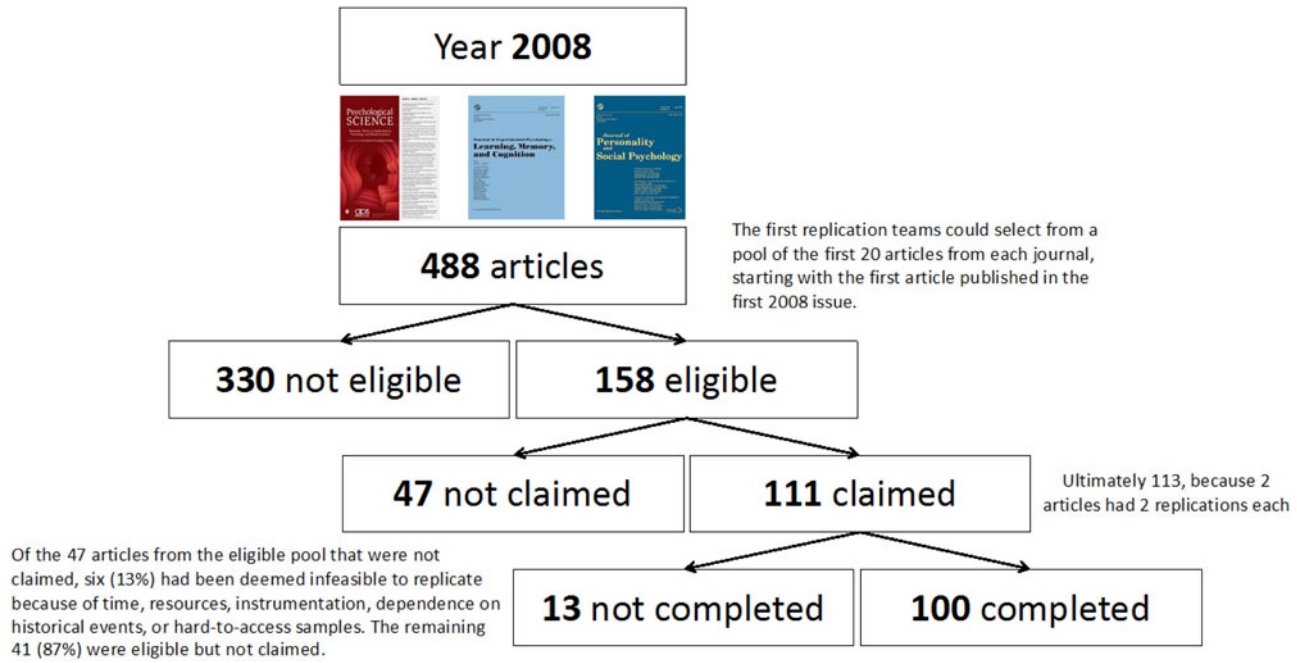

Figure 1. (Kuehberger & Michael Schulte-Mecklenbeck). Selection process for to-be-replicated papers in the Replication Project: Psychology (RPP). Texts in small print are citations from the Open Science Collaboration (2015).

sum, there is plenty of room for bias in the selection of replication targets. In our opinion, this problem has not yet been addressed adequately.

Direct replication and clinical psychological science

doi:10.1017/S0140525X18000754, e140

Scott O. Lilienfeld

Department of Psychology, Emory University, Atlanta, GA 30322; School of Psychological Sciences, University of Melbourne, Melbourne, VIC 3010, Australia

slilien@emory.edu

http://psychology.emory.edu/home/people/faculty/lilienfeld-scott.html

Abstract: Zwaan et al. make a compelling case for the necessity of direct replication in psychological science. I build on their arguments by underscoring the necessity of direct implication for two domains of clinical psychological science: the evaluation of psychotherapy outcome and the construct validity of psychological measures.

In their clearly reasoned target article, Zwaan et al. make a persuasive case that direct replication is essential for the health of psychological science. The principle of the *primacy of internal validity* (Cook et al. 1990) underscores the point that one must convincingly demonstrate a causal effect (internal validity) before generalizing it to similar settings, participants, measures, and the like (external validity). Some scholars appear to have overlooked the importance of this mandate. In an otherwise incisive article, my Ph.D. mentor David Lykken (1968) wrote that "Since operational replication [what Zwaan et al. term *direct replication*] must really be done by an independent second investigator and since constructive replication [what Zwaan et al. term *conceptual replication*] has greater generality, its success strongly impl[ies] that an operational replication would have succeeded also" (p. 159). Lykken, like many scholars, underestimated the myriad ways (e.g., p-hacking, file-drawering of negative results) in which conceptual replications can yield significant but spurious results (Lindsay et al. 2016). Hence, an apparently successful conceptual replication does not imply that the direct replication would have succeeded, as well.

I build on Zwaan et al.'s well-reasoned arguments by extending them to a subdiscipline they did not explicitly address: clinical psychological science. Probably because recent replicability debates have been restricted largely to scholars in cognitive, social, and personality psychology (Tackett et al. 2017a), the implications of these discussions for key domains of clinical psychology, especially psychotherapy and assessment, have been insufficiently appreciated. I contend that an overemphasis on conceptual replication at the expense of direct replication can generate misleading conclusions that are potentially detrimental to clinical research and patient care.

In the psychotherapy field, attention has turned increasingly to the development and identification of empirically supported therapies (ESTs; Chambless & Ollendick 2001), which are treatments demonstrated to be efficacious for specific disorders in independently replicated trials. Their superficial differences notwithstanding, all EST taxonomies require these interventions to be manualized or at least delineated in sufficient detail to permit replication by independent researchers. Although direct replications of psychotherapy outcome studies are often impractical (Coyne 2016) given the formidable difficulties of recruiting comparable patients and ensuring comparably trained therapists, investigators can still undertake concerted efforts to ascertain whether a carefully described psychotherapy protocol that yields positive effects in one study does so in future studies. Herein lies the problem: Without an independently replicated demonstration that the original protocol generates positive effects, practitioners and researchers can interpret a successful conceptual replication of a modified protocol as evidence that the treatment is ready for routine clinical application. Such a conclusion would be premature and potentially harmful, because the original protocol has demonstrated its mettle in a single study alone.

Conversely, practitioners and researchers may assume that a conceptual replication failure implies that the initial psychotherapy protocol was ineffective, but this conclusion could likewise be erroneous. Admittedly, research on the extent to which adaptations of EST protocols tend to degrade their efficacy is inconsistent (Stirman et al. 2017). Nevertheless, in certain instances, seemingly minor changes in psychotherapy protocols may produce detrimental effects. For example, studies of exposure therapy for anxiety disorders suggest that the commonplace practice of encouraging patients to engage in safety behaviors (e.g., practicing relaxation skills) during exposure often adversely affects treatment outcomes (Blakey & Abramowitz 2016). The same overarching conclusion may hold for self-help interventions. Rosen (1993) observed that even seemingly trivial changes to self-help programs can result in unanticipated changes in treatment compliance, effectiveness, or both. For example, in one study the addition of a self-reward contracting manipulation to an effective program for snake phobia decreased treatment compliance from 50% to zero (Barrera & Rosen 1977), perhaps because clients perceived the supplementary component as onerous. Consequently, failed conceptual replications can lead to the mistaken conclusion that effective treatment protocols are impractical, ineffective, or both.

In the clinical assessment field, an overemphasis on conceptual replication can contribute to what Pinto and I (Lilienfeld & Pinto 2015) termed the *illusion of replication*. This illusion can arise when investigators fail to delineate an explicit nomological network (Cronbach & Meehl 1955) of predictions for the construct validation of a measure, permitting them to engage in a program of ad hoc validation (Kane 2001). In such a research program, psychologists are free to hand-pick from an assortment of findings on diverse indicators to justify support for a measure's construct validity. In some cases, they may conclude that a measure has been validated for a given clinical purpose even in the absence of a single directly replicated finding.

Research on the widely used "Suicide Constellation" of the Rorschach Inkblot Test affords a potential illustration. Based on a meta-analysis of Rorschach variables, an author team concluded that the Suicide Constellation is a well-validated indicator of suicide risk (Mihura et al. 2013, p. 572). Nevertheless, this conclusion hinged on only four studies (see Wood et al. [2015] for a discussion), one on completed suicides, one on attempted suicides, one on ratings of suicidality, and one on levels of serotonin in cerebrospinal fluid (low levels of which have been tied to suicide risk [Glick 2015]). As a result, the validity of the Suicide Constellation is uncertain given that its support rests on correlations with four ostensibly interrelated, but separable, indicators, with no evidence of direct replication.

Conversely, researchers may assume that a conceptual replication failure following a seemingly minor change to a measure calls into question the initial positive finding. For example, in efforts to save time, investigators frequently administer abbreviated forms of well-established measures, such as the Minnesota Multiphasic Personality Inventory–2. Nevertheless, such short forms often exhibit psychometric properties inferior to those of their parent measures (Smith et al. 2000). Hence, failed conceptual replications using such measures do not mean that the original result was untrustworthy.

When it comes to psychological treatments and measures, generalizability cannot simply be assumed. Direct replications of initial positive results, or at least close approximations of them, are not merely a research formality. They are indispensable for drawing firm conclusions regarding the use of clinical methods.

Replication is already mainstream: Lessons from small-*N* designs

doi:10.1017/S0140525X18000766, e141

Daniel R. Little and Philip L. Smith

Melbourne School of Psychological Sciences, University of Melbourne, Parkville, VIC 3010, Australia.

daniel.little@unimelb.edu.au philipls@unimelb.edu.au

http://psychologicalsciences.unimelb.edu.au/research/research-groups/knowlab

Abstract: Replication is already mainstream in areas of psychology that use small-*N* designs. Replication failures often result from weak theory, weak measurement, and weak control over error variance. These are hallmarks of phenomenon-based research with sparse data. Small-*N* designs, which focus on understanding processes, treat the individual rather than the experiment as the unit of replication and largely circumvent these problems.

The claim that psychology has not given due consideration to replication treats psychology as a homogeneous discipline in which the focus is on demonstrating the presence or absence of experimental effects. In contrast, we argue that replication has long been a part of standard research practice, and is already mainstream in several areas of psychology, including visual and auditory psychophysics, animal learning, and mathematical psychology, and many parts of cognitive psychology. A common feature of research in these areas is the systematic use of small-*N* designs, in which a small number of expert participants (or highly trained animals) are tested over many sessions. The effects of interest in these designs are thus replicated over trials, over sessions, and between participants. As a result, the questions of theoretical interest can be tested at the individual participant level rather than the group level. The individual participant rather than the group then becomes the replication unit, and the study effectively becomes its own replication.

Failure to replicate is often a symptom of deeper problems that arise from the three vices of weak measurement, weak theory, and weak control over error variance. It is typical in much of psychology for the relationship between the measurement scale and the underlying theoretical constructs to be at best ordinal. Ordinal-level theories can at best predict that performance in one condition will be greater (more accurate, faster, etc.) than performance in a second condition; they cannot predict strong functional relationships. They also tend to be sparse in the sense that inferences are made using single point estimates. Because effects vary from individual to individual, the typical response to measurement variability is to increase the sample size. Without addressing these more basic problems, however, a focus on increased replication will only squander limited resources on ill-thought-out questions. Although often discussed in different terms, replication can be viewed simply as another way to increase the sample size to try to obtain a better estimate of the effect size. When viewed in this way, replication continues to serve the questionable goal of establishing the existence of an effect that is defined only in weak ordinal terms.

In contrast to this type of phenomenon-driven research (Meehl 1967; 1990a), the goal of small-*N* research is usually not to demonstrate some effect but to elucidate the underlying process-based mechanism that leads to the behavior of interest. This typically entails strong measurement and hypothesizing on a stronger-than-ordinal scale. In visual psychophysics, for example, variables such as contrast, summation time, motion direction or speed, orientation thresholds, and response time are measured on ratio scales and used to define functional predictions across the range of stimulation (e.g., psychometric functions or response-time distributions).

The focus of process-based research is on testing theoretical model predictions and not on testing the significance of experimental effects. Because the mechanisms of interest are typically defined at the individual level, it is most appropriate to test predictions about them at the same level (Grice et al. 2017). To appropriately control error variance across individuals, at least two methods are commonplace: first, participants are typically highly practiced; second, stimulus manipulations are tailored to the specific sensitivities of the individuals. These methods act to counteract the lack of precise control that arises with naïve participants. Individuals are tested extensively so that the distribution of responses (and other characteristics of those responses, like timing) is estimated with high power. Testing a small number of participants, each of whom acts as a replication of the entire experiment, controls for contextual variation across both time (between sessions but within individuals) and individuals (between individuals but within sessions).

The upshot of this style of research is that rich contact can be made between theory and data. This contact facilitates the use of strong inference methods to falsify specific models (see e.g., Little et al. 2017) and testing of strong out-of-sample and out-of-context predictions (Yarkoni & Westfall 2017). These kinds of systematic tests of strong quantitative relationships are characteristic of mature sciences that psychology should be striving to become. Although replication might weed out spurious effects, it often begs the question why we should care about these effects in the first place.

We (Smith & Little 2018) recently demonstrated the advantages of individual-level analysis in cognitive settings by simulating effects of different sizes using the additive factors method (a method for characterizing the stages of processing in a cognitive task [Sternberg 1969]) and then estimating the power of either individual-level analysis (e.g., maximum likelihood model estimation) or group-level statistical analysis (i.e., analysis of variance [ANOVA]). The goal of the additive factors method is to determine the presence or absence of an interaction which provides either falsification or confirmation, respectively, of the point prediction of a serial, sequential-stages processing model. Our results indicated that the individual-level analysis could detect the presence of an interaction even with small effect sizes. The group-level statistical analysis, by contrast, reached similar levels of power only when the group sample size was increased substantially. Further, the individual-level analysis also provides an estimate of the value of the interaction and the consistency with which it appeared across individuals.

Small-*N* designs will not be appropriate for all areas of psychology. They will not be appropriate with reactive measures that allow only a single measurement per person or when multiple measurements are made on individuals but the resulting data are sparse. In the latter eventuality, the best approach is to model the individual variation at the group level (i.e., hierarchically [Lee & Wagenmakers 2005]). Our argument is that the level of replication must be appropriate for the question being asked. In the areas of psychology with which we are concerned, this is at the individual level. The fact that areas that routinely use small-*N* paradigms have so far remained immune to the replication crisis afflicting other areas of psychology can be seen as an object lesson on the kind of methodological reform that the discipline requires, which goes deeper than just the routine practice of replication.

Enhancing research credibility when replication is not feasible

doi:10.1017/S0140525X18000778, e142

Robert J. MacCoun

School of Law, Department of Psychology, and Freeman-Spogli Institute, Stanford University, Stanford, CA 94305.

rmaccoun@stanford.edu

https://law.stanford.edu/directory/robert-j-maccoun/

Abstract: Direct replications are not always affordable or feasible, and for some phenomena they are impossible. In such situations, methods of

blinded data analysis can help minimize p-hacking and confirmation bias, increasing our confidence in a study's results.

In their target article, Zwaan et al. make a spirited and persuasive case for an increased emphasis on replication of psychological studies. Yet I want to challenge their conclusion that "there are no theoretical or statistical obstacles to making direct replication a routine aspect of psychological science" (Abstract).

The word "experiment" appears 52 times in the target article, and it appears that the authors are mostly concerned with theory-testing experiments (often using college students or recruited Web samples). Direct replications are almost always feasible for such studies. They are not without cost, including the opportunity cost of replicating an old result rather than testing a new result. But the costs are not great, which is perhaps why Zwaan et al. mention only one type of cost – "reputational costs."

But nonexperimental and/or field methods make up a substantial minority of published psychological research (Cortina et al. 2017; Lipsey & Wilson 1993). For such studies, both words in the phrase "direct replication" aspirations are problematic.

The notion of a "direct" replication is problematic in field studies, because standardization is so difficult to achieve – even if it were desirable. Different studies of the same question occur in different locations, in different years, with different implementation details, client caseloads, resources, and political environments. A cautionary tale is provided by a National Institute of Justice effort to replicate an influential experiment on the effect of arrests in domestic assault cases, forcing analysts to search for covariates to account for dramatically disparate results across different cities (see Berk et al. 1992). Such situations show that such categories as "construct validity," "statistical conclusion validity," and "external validity," although conceptually distinct, are very fuzzy in practice.

In any case, field studies and archival analyses tend to be much more costly and much less feasible to replicate. Indeed, replication is impossible (or nearly so) for empirical studies based on real-world events, such as homicide rates, gun purchases, presidential elections, terrorist events, and responses to natural disasters, or changes in law. Moreover, maximum sample sizes are often constrained by circumstances beyond the control of the researcher.

Despite all of these problems, field research is indispensable for addressing many important questions. When we cannot replicate such studies, are there other ways of enhancing their statistical reliability and validity? Some areas of physics have had to confront a similar dilemma. Major experiments in particle physics can be conducted at only a few facilities (sometimes only one), at enormous expense. Some questions in cosmology can only be answered at present using a small set of observations (e.g., of supernovae) that, although not rare in the universe, are very costly and difficult to detect.

To make the best use of such data, and to minimize the risk of p-hacking and other forms of biased inference, many physicists routinely perturb their data (e.g., by adding noise and/or bias, scrambling data labels, or "salting" the data with fake events [see Klein & Roodman 2005]). Blinding methods are increasingly used in forensic science for similar reasons – the conditions under study were historically unique and cannot be replicated to compensate for analyst bias (Robertson & Kesselheim 2016). Recently, physicist Saul Perlmutter and I have examined such methods and their suitability for psychological science (MacCoun & Perlmutter 2015; 2017). Different methods blind different features of the results (e.g., point estimates, effect sizes, statistical significance). More research is needed to examine their suitability and effectiveness for the behavioral sciences. But it seems likely that blind analysis can make results more credible and reduce some of the biases that reduce the replicability of research findings.

Verify original results through reanalysis before replicating

doi:10.1017/S0140525X18000791, e143

Michèle B. Nuijten, Marjan Bakker, Esther Maassen, and Jelte M. Wicherts

Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, 5037 AB, Tilburg, The Netherlands

m.b.nuijten@tilburguniversity.edu
m.bakker_1@tilburguniversity.edu **e.maassen@tilburguniversity.edu**
j.m.wicherts@tilburguniversity.edu
https//mbnuijten.com
http://marjanbakker.eu
https//www.tilburguniversity.edu/webwijs/show/e.maassen/
http://jeltewicherts.net

Abstract: In determining the need to directly replicate, it is crucial to first verify the original results through independent reanalysis of the data. Original results that appear erroneous and that cannot be reproduced by reanalysis offer little evidence to begin with, thereby diminishing the need to replicate. Sharing data and scripts is essential to ensure reproducibility.

Zwaan et al. (2017) provide an important and timely overview of the discussion as to whether direct replications in psychology have value. Along with others (see, e.g., Royal Netherlands Academy of Arts and Sciences 2018), we agree wholeheartedly that replication should become mainstream in psychology. However, we feel that the authors missed a crucial aspect in determining whether a direct replication is valuable. Here, we argue that it is essential to first verify the results of the original study by conducting an independent reanalysis of its data or a check of reported results, before choosing to replicate an earlier finding in a novel sample.

A result is successfully reproduced if independent reanalysis of the original data, using either the same or a (substantively or methodologically) similar analytic approach, corroborates the result as reported in the original paper. If a result cannot be successfully reproduced, the original result is not reliable and it is hard, if not impossible, to substantively interpret it. Such an irreproducible result will have no clear bearing on theory or practice. Specifically, if a reanalysis yields no evidence for an effect in the original study, it is safe to assume that there is no effect to begin with, raising the question of why one would invest additional resources in any replication.

Problems with reproducibility in psychology. Lack of reproducibility might seem like a non-issue; after all, it may seem like a guarantee that running the same analysis on the same data would give the same result. However, there is increasing evidence that reproducibility of published results in psychology is relatively low.

Checking reproducibility of reported results in psychology is greatly impeded by a common failure to share data (Vanpaemel et al. 2015; Wicherts et al. 2006). Even when data are available, they are often of poor quality or not usable (Kidwell et al. 2016). Yet some issues with reproducibility can be assessed by scrutinizing papers. Studies repeatedly showed that roughly half of all published psychology articles contains at least one inconsistently reported statistical result, wherein the reported p value does not match the degrees of freedom and test statistic; in roughly one in eight results this may have affected the statistical conclusion (e.g., Bakker & Wicherts, 2011; Nuijten et al. 2016; Veldkamp et al. 2014; Wicherts et al. 2011). Furthermore, there is evidence that roughly half of psychology articles are inconsistent with the given sample size and number of items (Brown & Heathers 2017), coefficients in mediation models often do not add up (Petrocelli et al. 2012), and in 41% of psychology articles reported degrees of freedom do not match the sample size description (Bakker & Wicherts 2014).

Problems that can be detected without having the raw data, are arguably just the tip of the iceberg of reproducibility issues. Studies that intended to reanalyze data from published studies also often ran into problems (e.g., Ebrahim et al. 2014; Ioannidis et al. 2009). Beside the poor availability of raw data, papers usually do not contain details about the exact analytical strategy. Researchers often seem to make analytical choices that are driven by the need to obtain a significant result (Agnoli et al. 2017; John et al. 2012). These choices can be seemingly arbitrary (e.g., choice of control variables or rules for outlier removal; see also Bakker et al. [2012] and Simmons et al. [2011]), which makes it hard to retrace the original analytical steps to verify the result.

Suggested solution. Performing a replication study in a novel sample to establish the reliability of a certain result is time consuming and expensive. It is essential that we avoid wasting resources on trying to replicate a finding that may not even be reproducible from the original data. Therefore, we argue that it should be standard practice to verify the original results before any direct replication is conducted.

A first step in verifying original results can be to check whether the results reported in a paper are internally consistent. Some initial screenings can be done quickly with automated tools such as "statcheck" (Epskamp & Nuijten 2016; http://statcheck.io), "*p*-checker" (Schönbrodt 2018), and granularity-related inconsistency of means ("GRIM" [Brown & Heathers 2017]). Especially if such preliminary checks already flag several potential problems, it is crucial that data and analysis scripts are made available for more detailed reanalysis. One could even argue that if data are not shared in such cases, the article should be retracted.

If a result can successfully be reproduced with the original data and analyses, it is interesting to investigate its sensitivity to alternative analytical choices. One way to do so is to run a so-called multiverse analysis (Steegen et al. 2016), in which different analytical choices are compared to test the robustness of the result. When a multiverse analysis shows that the study result is present in only a limited set of reasonable scenarios, you may not want to invest additional resources in replicating such a study. Note that a multiverse analysis still does not require any new data, and is therefore a relatively cost-effective way to investigate reliability.

Reanalysis of existing data is a crucial tool in investigating reliability of psychological results, so it should become standard practice to share raw data and analysis scripts. Journal policies can be successful in promoting this (Giofrè et al. 2017; Kidwell et al 2016; Nuijten et al. 2017), so we hope that more journals will start requiring raw data and scripts.

In our proposal, the assessment of replicability is a multistep approach that first assesses whether the original reported results are internally consistent, then sets out to verify the original results through independent reanalysis of the data using the original analytical strategy, followed by a sensitivity analysis that checks whether the original result is robust to alternative choices in the analysis, and only then involves the collection of new data.

Direct replications in the era of open sampling

doi:10.1017/S0140525X18000808, e144

Gabriele Paolacci[a] and Jesse Chandler[b,c]

[a]*Rotterdam School of Management, Erasmus University Rotterdam, 3062 PA, Rotterdam, The Netherlands;* [b]*Mathematica Policy Research, Ann Arbor, MI 48104;* [c]*Institute for Social Research, University of Michigan Ann Arbor, MI 48109.*

gpaolacci@rsm.nl **jjchandl@umich.edu**
https://www.rsm.nl/people/gabriele-paolacci/
https://www.jessechandler.com

Abstract: Data collection in psychology increasingly relies on "open populations" of participants recruited online, which presents both opportunities and challenges for replication. Reduced costs and the possibility to access the same populations allows for more informative replications. However, researchers should ensure the directness of their replications by dealing with the threats of participant nonnaiveté and selection effects.

When the "crisis of confidence" struck psychology, giving a new pace to the academic debate on replications, a parallel revolution was happening in the field: Data collection rapidly moved away from near exclusive dependence on traditional participant pools (e.g., undergraduate samples) and towards sampling from online marketplaces where adults complete tasks (e.g., academic surveys) in exchange for compensation. About five years later, virtually any major journal in psychology and beyond routinely publishes studies conducted on Amazon Mechanical Turk, Prolific, or other third-party platforms (Chandler & Shapiro 2016; Stewart et al. 2017). Importantly, these marketplaces are typically "open" on both ends: Compared to any traditional participant pool (e.g., psychology undergraduates in a Midwestern university), few restrictions exist about who can join the participant pool and who can recruit participants from these populations.

Zwaan et al. provide a compelling case for direct replications, emphasizing both the necessity of being able to reproduce the procedures used in the original experiments and the lack of structural obstacles to make replications a habit in the field. However, Zwaan et al. do not discuss how direct replications are affected by the current practices of data collection, and in particular by researchers' increasing reliance on open sampling. We build on the target article by highlighting how open sampling presents opportunities to make direct replication mainstream and the challenges of conducting a proper direct replication using these samples.

Open sampling can remove barriers to making replications habitual, while also making attempted replications more conclusive and compelling. First, data collection from open populations is comparatively faster and cheaper (even controlling for pay rate, Goodman & Paolacci 2017). This reduces concerns about committing scarce resources to replication, and allows recruiting larger samples given the same time and budget allocated to conducting a replication. This is beneficial for any study and particularly for replication studies that demand even more participants than original studies to make conclusive statements (Simonsohn 2015).

Second, original studies conducted on open populations can be replicated by different researchers using the same population. Sharing the same population does not make replications perfect, and we discuss below how this is also true of open populations; however, a shared population is a necessary precondition for more informative failed replications. Samples from different sources vary substantially on many characteristics, which can sometimes have a substantive impact on results (Krupnikov & Levine, 2014). All else being equal, a failed replication on the same population is both less suggestive of hidden moderators and less ambiguous about which "hidden moderators" (if any) might be at play. When the replicator's goal is to increase the directness of a replication, rather than discovering population-level moderators of the target effect, open populations further reduce the "Context Is Too Variable" concern that Zwaan et al. address in the target article.

Despite these advantages, open sampling only increases the directness of a replication if researchers pay appropriate attention to the sampling methodology. First, despite intuitions of the contrary, open populations have a large but limited number of participants (Difallah et al. 2018; Stewart et al. 2015). Combined with researchers using these populations to conduct many studies that are often high-powered, this has resulted in concerns about participant nonnaiveté that are relevant to replication. Open populations include many participants who are experienced with research participation, and who become more experienced over time with specific research paradigms and instruments. Illustratively, popular paradigms are known to a large majority of participants (e.g., Chandler et al. 2014, Thomson & Oppenheimer

2016). Zwaan et al. highlight how some findings in cognitive psychology (i.e., perception/action, memory, and language) replicate even with participants who were previously exposed to them (Zwaan et al. 2017). However, this is not necessarily the case for any paradigm, and may be particularly not true of replications in other psychological fields. There is evidence that experimental manipulations in social psychology and decision-making that convey experience (e.g., tasks conducted under time pressure) or factual knowledge (e.g., numeric estimates following different numeric anchors) become less strong with repeated exposure. This can result in replications that are less statistically powerful than intended (Chandler et al. 2015, Devoe & House 2016, Rand et al. 2014), and participant nonnaiveté should therefore be accounted for by direct replicators (Chandler et al. 2014).

Second, samples obtained from open populations are not probability samples, and thus can still vary as a result of procedural differences in sampling. Participants of open populations self-select into studies by choosing from many that differ across observable characteristics (e.g., payment, task description) that may make them more or less attractive to different people. Researchers may place explicit constraints on participant eligibility that have a measurable impact on data quality (e.g., worker reputation scores; Peer et al. 2014 or nationality; Chandler & Shapiro 2016) but may not be reported. Other recruitment criteria that are not deliberately selected may still be impactful. The diversity of open populations compounds this concern, because it suggests a comparatively high potential for procedural differences to meaningfully affect sample composition. Though evidence is still scarce, researchers have found that sample demographics fluctuate with time-of-the-day and day-of-the week (Arechar et al. 2017; Casey et al. 2017). This implies the need for direct replicators to consider aspects of the original design (e.g., timing, study compensation) that are not typically assumed to be hidden moderators in undergraduate samples that are less diverse and less characterized by self-selection. It also emphasizes that, in the era of open samples, original authors are as responsible as direct replicators to support replicability efforts by reporting their sampling choices in sufficient detail to ensure meaningful replication.

In sum, we applaud the target article on convincingly addressing the most commonly raised concerns about replication, and put some of the target article's insights within the context of today's dominant practice in data collection—open sampling. We hope this commentary will contribute to make *informative* replication mainstream, by encouraging researchers to both embrace the advantages of open sampling and consider what transparent reporting of sampling methods and direct replication means when using these samples.

You are not your data

doi:10.1017/S0140525X1800081X, e145

Gordon Pennycook

Department of Psychology, Yale University, New Haven, CT 06520-8205.

gordonpennycook.net

gordon.pennycook@yale.edu

Abstract: Scientists should, above all else, value the truth. To do this effectively, scientists should separate their identities from the data they produce. It will be easier to make replications mainstream if scientists are rewarded based on their stance toward the truth – such as when a scientist reacts positively to a failure to replicate – as opposed to a particular finding.

Zwaan et al. have provided a service to science by synthesizing the strong case for making replications mainstream. However, they have missed – and perhaps even subtly perpetuated – what may be the central underlying impediment to having a strong culture of replications: The idea that "you are your data."

The underlying value that ought to be common among scientists is respect for the truth. Scientific theories that are not backed up by evidence – by data or logic that supports belief in a proposition or set of propositions – are simply not scientific. Moreover, a *lack* of concern for the truth is linked to (and, for some, necessary for) the very antithesis of science: bullshitting (Frankfurt 2005). Naturally, however, people (and presumably scientists) differ in the extent to which they are receptive to bullshit (Pennycook et al. 2015), value logic and evidence (Ståhl et al. 2016), and acknowledge that their beliefs could be wrong (i.e., intellectual humility [Leary et al. 2017]).

If respect for the truth is the crucial underlying value that, at least, *ought* to bind scientists together, what implications does this have for making replications mainstream?

Zwaan et al. adroitly point out that "researchers may feel a sense of ownership of 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" (sect. 5.5, para. 3).

It is unfortunately true that some may make inferences about the value of a researcher based on a failure to replicate their work. However, scientists may overestimate the negative reputational consequences of a failure to replicate (Fetterman & Sassenberg 2015) as scientific reputation is based more on process than outcome (Ebersole et al. 2016b). Indeed, being subject to a failure to replicate does *not* provide a strong signal as to one's respect for the truth. Quite the contrary, ones' *response* to a failure to replicate is a substantially stronger signal. Do you care more about your reputation than the empirical result?

Unfortunately, at least based on anecdote, abject fear is often the intuitive reaction to discovering that one is to be replicated. This reaction is precisely the opposite of what it should be. For the most part, scientists research topics in which they are interested. Moreover, setting aside petty narcissism for a moment, the highest aspiration that a scientist can have is to make a meaningful impact on her field and perhaps even the outside world. Discovering that your finding is to be subject to a replication attempt should be doubly *exciting*, for it indicates that (a) you are to gain information about something you are interested in, and (b) your work is impactful enough to be considered worth replicating. Producing research that causes independent scientists to spend time and money pursuing is, plainly, an accomplishment.

Why, then, are replication attempts sometimes met with trepidation? As noted by Zwaan et al., failed replications are viewed as having negative reputational consequences (Bohannon 2014; Fetterman & Sassenberg 2015). The fear may be that a failure to replicate indicates that an individual engaged in questionable research practices – or even outright fraud – to get a significant result. However, this is an untenable conclusion when based on a single or small number of observed results. Anecdotal evidence is not acceptable in our science; it should not be acceptable in our evaluation of scientists.

Only when a *pattern* of nonreplicable results occurs is it justifiable for a researcher's reputation to be affected. Also, even in such a case, this is because researchers who cannot produce replicable results are likely to have put concerns about their reputation and career success (in the form of publications, tenure, grants, awards, etc.) above the pursuit of science (in the form of robust research that tells us something about the world). The offending researchers only "become their data" because (with enough observations) it becomes a signal of their stance toward the truth.

It is important to point out that the replicator, too, should care more about the empirical result than the reputation of the person being replicated. A focus on "debunking" fellow scientists is guilty of the same toxic concern about reputation (and, therefore, nonprimacy of the pursuit for truth) that is needed to justify such a targeting in the first place. The fact such debunking attempts produce data about the world is not justification for perpetuating the (identity-focused) stance that may have been the source of bad data in the first place. If we want people to stop assuming that one

failure to replicate will ruin their reputation, it is imperative that we do not treat failures to replicate (outside of extreme circumstances) as having reputational consequences.

Although it is perhaps not feasible for every scientist to fully buy in to the "you are not your data" mantra, it is nonetheless important to increase its aggregate influence. To this end, scientists who demonstrate a willingness to divorce themselves from their data should be celebrated (see https://lossofconfidence.com/ for a group focused on this very thing). Awards and accolades should go to scientists who, beyond having a significant influence on their respective fields, can also provide evidence of identifying with the *process* of science and the pursuit of truth (e.g., via dedication to open science or revision of a previous stance based on new data; see Nosek et al. [2012]). Prestigious academic positions should be given to those who do research that is both impactful *and* sound (a notion that seems sufficiently obvious, but that does not necessarily correspond to the selection of individuals who have successfully created a "brand"). Finally, the significance of valuing the truth should be emphasized to graduate students and future generations of scientists, particularly in cases when the relaxing of scientific values is expedient. Ultimately, making replications mainstream will be easier if scientific incentive structures begin to align with a separation of identity and data.

ACKNOWLEDGMENTS

For comments on an earlier version, I thank Nathaniel Barr, Adam Bear, Shadi Beshai, Michal Bialek, Justin Feeney, Jonathan Fugelsang, Gordon Kraft-Todd, Srdan Medimorec, Sandeep Mishra, David Rand, Paul Seli, Nick Stagnaro, and Valerie Thompson.

The importance of exact conceptual replications

doi:10.1017/S0140525X18000821, e146

Richard E. Petty

Department of Psychology, Ohio State University, Columbus, OH 43210.
petty.1@osu.edu
https://richardepetty.com/

Abstract: Although Zwaan et al. argue that original researchers should provide a replication recipe that provides great specificity about the operational details of one's study, I argue that it may be as important to provide a recipe that allows replicators to conduct a study that matches the original in as many conceptual details as possible (i.e., an *exact conceptual replication*).

Zwaan et al. make the classic distinction between *exact replications* (using the same operations as in an original study) and *conceptual replications* (using different materials to instantiate the independent variables [IVs] and/or dependent variables [DVs]). They argue that exact replications are superior and therefore original authors should provide a "replication recipe" providing considerable detail about the specific operations used so others can duplicate one's study. Furthermore, Zwaan et al. claim that a finding is "not scientifically meaningful until it can be replicated *with the same procedures* that produced it in the first place" (sect. 6, para. 1). Instead, I argue that for much theoretical work in psychology, use of the same *operations* is not what is critical, but rather instantiation of the same *concepts*. Thus, theory testing researchers should emphasize conducting *exact conceptual replications* (ECRs) where the goal is to repeat as closely as possible not the precise methods of the original study, but to instantiate the same conceptual meaning of the original variables in the same conceptual context (Petty 2015).

In the physical sciences, the emphasis on carefully replicating operations is often reasonable. For instance, when mixing hydrogen and oxygen to create water, the choice of operations to represent the hydrogen and oxygen concepts is constrained because there is a tight link between concepts and operations (i.e., the operations and concepts are basically the same). Furthermore, the operations chosen are likely to represent the concepts across virtually all contexts. Thus, if you reasonably do the same thing, you should get the same result. In contrast, in many theory testing psychology studies, the choice of operations to represent concepts is vast and the link between the two is not assured. Thus, conducting a replication that is as close as possible to the original study will not necessarily help with replicability because the meaning of the original IVs and DVs in the new context may have changed.

Consider a psychologist mixing a credible source with a persuasive message to produce favorable attitudes toward some proposal. When Hovland and Weiss (1951) did this, Robert Oppenheimer was used as a credible source, and the Russian newspaper, *Pravda*, was the low credible source on the topic of building atomic submarines. Oppenheimer produced more favorable attitudes than *Pravda*. It seems unlikely that the same operations would produce the same result today. Does this render the original study scientifically meaningless? No. The initial result is meaningless only if you cannot conduct an ECR. ECRs are important because what we ultimately want to know is not whether Oppenheimer produces more favorable attitudes toward submarines than *Pravda*, but whether credibility affects persuasion.

The initial credibility study results would be meaningful if the study can be replicated in an ECR. Original authors can specify the criteria any replication study should meet. Namely, provide the *conceptual recipe*. This differs from the *operational recipe* that Zwaan et al. favor. Thus, if manipulating credibility, instead of only articulating operational details like replicators must have people see an 8 X 10 picture of the source with an 18 word description, original authors could also indicate that the high credibility manipulation should produce a rated level of credibility of 7 on an 11 point credibility scale and the low credibility condition should be at 4. But, it is not sufficient for replicators to produce a successful manipulation check. If the original study had high and low credibility means of 7 and 4 but the replication study had means of 1 and 4, the manipulation check in the replication study would seem "successful" (and the effect size of the manipulation check might be comparable to the original), but, the placement of the manipulation along the credibility continuum would be quite different and thus inappropriate for an ECR. In addition to providing information about the statistical properties of the IV manipulation check, original authors should specify what constructs the IV should *not* vary. Thus, original authors should not only provide the IV information just noted, but also what concepts should be assessed to ensure they are not confounded (e.g., source attractiveness and power).

Critically, similar arguments apply to the DV. In the chemistry example, the dependent variable (water) is easily assessed. However, there are multiple ways to assess favorable evaluations (e.g., explicit vs. implicit measures). Now consider a different original study in which investigators are examining the frustration to aggression link. These researchers should indicate how to determine if the dependent measure taps aggression. The original study might have measured how many teaspoons of hot sauce were administered, but in a replication attempt in Mexico, giving hot sauce may not signal aggressiveness. Thus, specifying what criteria the DV should meet (to gauge its conceptual meaning) is as important as specifying this for the IVs. For example, participants can rate how aggressive it is to give hot sauce.

Finally, the overall level of the DV on the aggression continuum in the new context is important. This is because unlike the chemistry example where there is only one way to produce water, psychology DVs are often multiply determined. There are many ways to produce aggression and there may be factors in the replication context that affect the hot sauce DV that were not present in the original research. Some of these may be alternative causes of

aggression (e.g., hot temperature), but others may influence giving out hot sauce for other reasons (e.g., its popularity or price in the culture). Each can be problematic and lead to replication failure. Thus, a replication recipe should focus on describing contextual factors that are plausibly linked to the DV. Most simply, one can report the mean level of the operational (amount of hot sauce) and conceptual (link to aggressiveness) DV in a control condition in which none of the critical IVs are varied. This is needed to assure that relevant background variables in the replication study that affect the DV are set at a similar level to the original study.

In sum, conceptually driven psychology research is different from the physical sciences, and our replication recipes should reflect this.

The replicability revolution

doi:10.1017/S0140525X18000833, e147

Ulrich Schimmack

Department of Psychology, University of Toronto, Mississauga, ON L5L 1C6, Canada.

Ulrich.schimmack@utoronto.ca

https://replicationindex.wordpress.com/

Abstract: Psychology is in the middle of a replicability revolution. High-profile replication studies have produced a large number of replication failures. The main reason why replication studies in psychology often fail is that original studies were selected for significance. If all studies were reported, original studies would fail to produce significant results as often as replication studies. Replications would be less contentious if original results were not selected for significance.

The history of psychology is characterized by revolutions. This decade is marked by the replicability revolution. One prominent feature of the replicability revolution is the publication of replication studies with nonsignificant results. The publication of several high-profile replication failures has triggered a confidence crisis.

Zwaan et al. have been active participants in the replicability revolution. Their target article addresses criticisms of direct replication studies.

One concern is the difficulty of re-creating original studies, which may explain replication failures, particularly in social psychology. This argument fails on three counts. First, it does not explain why published studies have an apparent success rate greater than 90%. If social psychological studies were difficult to replicate, the success rate should be lower. Second, it is not clear why it would be easier to conduct conceptual replication studies that vary crucial aspects of a successful original study. If social priming effects were, indeed, highly sensitive to contextual variations, conceptual replication studies would be even more likely to fail than direct replication studies; however, miraculously they always seem to work. The third problem with this argument is that it ignores selection for significance. It treats successful conceptual replication studies as credible evidence, but bias tests reveal that these studies have been selected for significance and that many original studies that failed are simply not reported (Schimmack 2017; Schimmack et al. 2017).

A second concern about direct replications is that they are less informative than conceptual replications (Crandall & Sherman 2016). This argument is misguided because it assumes a successful outcome. If a conceptual replication study is successful, it increases the probability that the original finding was true and it expands the range of conditions under which an effect can be observed. However, the advantage of a conceptual replication study becomes a disadvantage when a study fails. For example, if the original study showed that eating green jelly beans increases happiness and a conceptual replication study with red jelly beans does not show this effect, it remains unclear whether green jelly beans make people happier or not. Even the nonsignificant finding with red jelly beans is inconclusive because the result could be a false negative. Meanwhile, a failure to replicate the green jelly bean effect in a direct replication study is informative because it casts doubt on the original finding. In fact, a meta-analysis of the original and replication study might produce a nonsignificant result and reverse the initial inference that green jelly beans make people happy. Crandall and Sherman's argument rests on the false assumption that only significant studies are informative. This assumption is flawed because selection for significance renders significance uninformative (Sterling 1959).

A third argument against direct replication studies is that there are multiple ways to compare the results of original and replication studies. I believe the discussion of this point also benefits from taking publication bias into account. Selection for significance explained why the reproducibility project obtained only 36% significant results in direct replications of original studies with significant results (Open Science Collaboration 2015). As a result, the significant results of original studies are less credible than the nonsignificant results in direct replication studies. This generalizes to all comparisons of original studies and direct replication studies. Once there is suspicion or evidence that selection for significance occurred, the results of original studies are less credible, and more weight should be given to replication studies that are not biased by selection for significance. Without selection for significance, there is no reason why replication studies should be more likely to fail than original studies. If replication studies correct mistakes in original studies and use larger samples, they are actually more likely to produce a significant result than original studies.

Selection for significance also explains why replication failures are damaging to the reputation of researchers. The reputation of researchers is based on their publication record, and this record is biased in favor of successful studies. Thus, researchers' reputations are inflated by selection for significance. Once an unbiased replication produces a nonsignificant result, the unblemished record is tainted, and it is apparent that a perfect published record is illusory and not the result of research excellence (a.k.a flair). Thus, unbiased failed replication studies not only provide new evidence; they also undermine the credibility of existing studies. Although positive illusions may be beneficial for researchers' eminence, they have no place in science. It is therefore inevitable that the ongoing correction of the scientific record damages the reputation of researchers, if this reputation was earned by selective publishing of significant results. In this way direct replication studies complement statistical tools that can reveal selective publishing of significant results with statistical tests of original studies (Schimmack 2012; 2014; Schimmack & Brunner submitted for publication).

Constraints on generality statements are needed to define direct replication

doi:10.1017/S0140525X18000845, e148

Daniel J. Simons,[a] Yuichi Shoda,[b] and D. Stephen Lindsay[c]

[a]*Department of Psychology, University of Illinois, Champaign, IL 61820;* [b]*Department of Psychology, University of Washington, Seattle, WA 98195;* [c]*Department of Psychology, University of Victoria, Victoria, BC V8W 2Y2, Canada.*

dsimons@illinois.edu **yshoda@uw.edu** **slindsay@uvic.ca**

www.dansimons.com

http://www.psych.uw.edu/psych.php?p=358&person_id=2569

https://www.uvic.ca/socialsciences/psychology/people/faculty-directory/lindsaysteve.php

Abstract: Whether or not a replication attempt counts as "direct" often cannot be determined definitively after the fact as a result of flexibility in how procedural differences are interpreted. Specifying constraints on generality in original articles can eliminate ambiguity in advance, thereby leading to a more cumulative science.

Exact? Direct? Conceptual? How we label a replication attempt depends on its similarity to the original study (Zwaan et al.). But "similarity" in whose view and with respect to which parameters? When a replication study reproduces the original finding, the distinction between direct and conceptual is less fraught – any discrepancies between the original and replication procedures provide evidence for generality. But, when a replication result differs from the original, any procedural changes become candidate explanations for the discrepancy; whether or not it counts as "direct" depends on how you interpret those differences in procedures.

Kahneman (2014) hoped to eliminate this ambiguity by encouraging researchers conducting replication studies to consult with the original authors about the appropriate procedures. The Registered Replication Reports developed at *Perspectives on Psychological Science* and now hosted at *Advances in Methods and Practices in Psychological Science* consult the original authors extensively to verify the accuracy of the replication protocol. *ManyLabs* projects attempt to do so, as well, as do the new Preregistered Direct Replications in *Psychological Science*. Although such contacts are a professional courtesy that can and often do improve the precision of a replication protocol, they are not always possible (i.e., the original researcher may be deceased or otherwise unresponsive). Furthermore, in principle they should not be necessary –published articles are the currency of our field, and they should stand alone as a statement of what is needed to reproduce a finding.

Much as preregistration makes clear whether an analysis was planned in advance or exploratory, an original article can make clear whether a subsequent replication is direct or conceptual if it explicitly defines constraints on generality (a COG statement; see Simons et al. [2017] for details). Most psychology papers generalize from the tested samples – of participants, stimuli, testing contexts – to broader populations (e.g., all college students, any emotional scene, any testing cubicle). All papers should explicitly define these target populations and indicate which presumed generalizations are derived from empirical evidence, which are theoretical predictions, and which are hunches. A COG statement defines the scope of the original finding, clarifies which aspects of the samples must be reproduced in a direct replication, and identifies claims of generality that need further testing in conceptual replications. It indicates those aspects of the original study its authors believe to be essential, thereby defining the conditions that the authors believe would constitute a direct replication.

Current publishing practices incentivize making the boldest and most general claims possible: A study testing undergraduate psychology students is generalized to all adults; a laboratory effect in a controlled, computer-based task is generalized to real-world decision making; interactions with a trained confederate are generalized to romantic relationships; studies with WEIRD samples are generalized to non-WEIRD populations (Henrich et al. 2010). Such broad conclusions are inherently more interesting and provocative, but rarely are justified and often are not justifiable. When constraints on generality are left unstated, authors implicitly encourage readers (including editors and reviewers) to generalize broadly, even if they would explicitly discourage readers from doing so.

Without a COG statement, both original authors and those conducting replications can dispute the meaning of a replication study, assigning the burden of proof to "do it right" to each other. An original author might assume narrow generality for their finding across contexts, so they place the burden of proof on the replicator to ensure that the replication mimics all aspects of the original context. They can dismiss unsuccessful replications as fatally flawed on account of changes in context that, in hindsight, they believe are important. In contrast, a replicator might assume maximal generality in the absence of stated constraints. When an original author posits a change in context as the reason for an unsuccessful replication, the replicating author will expect the original author to demonstrate that it was not a false positive (e.g., by making the effect come and go with changes to a moderator).

Adding a COG statement overcomes the ambiguity of classifying replications as direct or conceptual after the fact because it specifies the target populations for the original claim, allowing anyone to draw from those same populations. It bypasses the need to consult the authors of a publication to identify those constraints, allowing the published finding to serve as the unit of analysis. It also focuses our collective research enterprise on what should be our goal – not just determining whether or not effects exist, but identifying boundary conditions and mechanisms. Direct replications will test the reliability of those procedures thought to be essential, narrowing the purported scope of an original effect or perhaps demonstrating conditions under which it does not occur. Conceptual replications will test whether the proposed constraints on generality are accurate, leading to a more refined understanding of the robustness of the effect. A systematic program of research would evaluate how the size of an effect varies as a function of those constraints; how large is the effect within the narrowest, idiosyncratic confines of the original study procedures and how quickly and in what ways does it change when the restraints are loosened.

What the replication reformation wrought

doi:10.1017/S0140525X18000857, e149

Barbara A. Spellman[a] and Daniel Kahneman[b]

[a]*University of Virginia School of Law, Charlottesville, VA 22903;* [b]*Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, 08544.*

bas6g@virginia.edu **kahneman@princeton.edu**

http://content.law.virginia.edu/faculty/profile/bas6g/1211027

Abstract: Replication failures were among the triggers of a reform movement which, in a very short time, has been enormously useful in raising standards and improving methods. As a result, the massive multilab multi-experiment replication projects have served their purpose and will die out. We describe other types of replications – both friendly and adversarial – that should continue to be beneficial.

As two old(er) researchers who were involved early in the current science reform movement (pro-reform, to the chagrin of many of our peers), we believe that the target article barely addresses an essential point about the "replication crisis": In a very short time, the resulting reform movement, including all of the fuss, anger, and passion that it generated, has been enormously useful in raising standards and improving methods in psychological science. Rather than believing that the field is still in crisis, some highly influential members of our community recently announced that psychology is now experiencing a "renaissance" (Nelson et al. 2018). One of us calls what has happened a civil war–like revolution (Spellman 2015), suggesting an insurrection in which one group overthrows the structures put in place by another group. But here we use the term "reformation," suggesting that the profession has become enlightened and changed itself for the better.

The reform movement has prompted changes across the entire research and publication process. As a result, experimental results are more reliable because researchers are increasing sample sizes. Researchers are posting methods, data, and analysis plans (sometimes encouraged by journals), thus promoting more accurate replications and vetting of data integrity. Researchers are preregistering hypotheses and journals are pre-accepting registered reports, making conclusions more credible. Also the experimental record is more complete because of preprint services, open access journals, and the increasing publication of replication studies. Therefore, we believe that the reformation's success results from actions by individuals, journals, and societies, combined with various environmental factors (e.g., technology, demographics, the cross-disciplinary recognition of the problem [Spellman

2015]) that allowed the changes to take hold now, whereas reform movements with similar goals in the past had failed.

Amazingly, all of this has transpired in seven plus or minus two years. The early revelations that an assortment of high-profile studies failed to replicate, and then the later various mass replications – both those in which many different labs worked on many different studies (e.g., Nosek et al. 2015) and those in which many different labs worked on the same studies (e.g., Simons et al. 2014)–provided existence proofs that non-replicable published studies were widespread in psychology. The ground-breaking gem of a paper by Simmons et al. (2011) gave our field a way to understand how this could have happened by scientists simply following the norms as they understood them, without any evil intent. But the norms were defective.

We believe that the quality of psychological science has been improving so fast and so broadly–mainly because of the replication crisis–that replications are likely to become rarer rather than routine. The massive multi-lab multi-experiment replication projects have served their purpose and will die out. What should happen, and indeed become mainstream, is the extension of original research should routinely include replication. The design of experiments and their execution are separable: Friendly laboratories should routinely exchange replication services in a shared effort to improve the transparency of their methods. Most replications should be friendly and adversarial replications should be collegial and regulated. How might this be done?

In one approach (Kahneman 2014), after developing but before running a study, replicators send the original authors a complete plan of the procedure. Original authors have a set time to respond with comments and suggested modifications. Replicators choose whether and how to change the protocol but must explain how and why. These exchanges should be available for reviewers and readers to use when evaluating the claims of each side (e.g., whether it was a "faithful" replication).

In a second approach, the negotiation is refereed. For example, journals that take pre-registered replications may require careful vetting of the replicator's protocol before giving it a go-ahead stamp of "true direct replication." But journal intercession is not necessary; authors and replicators could agree to mediation by knowledgeable individuals or teams of appointed researchers.

The two proposals above, however, are limited to checking the replicability of individual studies – individual "bricks in the wall" – in the same way current reforms directly affect only the integrity of individual studies (Spellman 2015). Science involves groups–groups of studies that connect together to define and develop (or destroy) theories (i.e., to create buildings or bridges from individual bricks) and communities of scientists who can work together, or in constructive competition (note: not opposition), to hone their shared ideas. Below we suggest two ways in which communities of scientists can engage in replication and theory development.

A third approach to replication is the daisy-chain approach. A group of laboratories that share a theoretical orientation that others question could get together, with each lab offering its favorite experiment for exact replication by the next lab in the chain – with all results published together, win or lose. Even if not all of the replications are successful, such an exercise would improve the quality of communications about research methods within a field, and improve the credibility of the field as a whole.

A fourth ambitious form of replication, called "paradigmatic replication," has been implemented by Kathleen Vohs (2018). Vohs recognizes that massive replication attempts of one study, particularly in an area where different researchers use different methods that change over time, is not a useful indicator of an evolving theory's robustness. In this procedure, the major proponents of a theory jointly resolve what the core elements of the theory are, and then decide what the best methods had been (or would be) to demonstrate/test its workings. A few diverse methods (e.g., varying independent or dependent measures) are devised, the protocols are pre-registered, and then multiple labs, both "believers" and "non-believers," run the studies. Data are analyzed by a neutral third party.

Overall, we believe that the replication reform movement has already succeeded in valuable ways. Improvements of research methods are raising the credibility of results and reducing the need for replications by skeptics. We also believe that routine exchanges of "replication services" between cooperating laboratories (e.g., through StudySwap [https://osf.io/view/StudySwap/]) will further enhance the community's confidence in the clarity and completeness of methods, as well as in the stability of findings.

Verifiability is a core principle of science

doi:10.1017/S0140525X18000869, e150

Sanjay Srivastava

Department of Psychology, University of Oregon, Eugene, OR 97403-1227.
sanjay@uoregon.edu
http://psdlab.uoregon.edu/

Abstract: Scientific knowledge is supposed to be verifiable. Replications promote verifiability in several ways. Most straightforwardly, replications can verify empirical claims. Replication research also promotes dissemination of information needed for other aspects of verification; creates meta-scientific knowledge about what results to treat as credible even in the absence of replications; and reinforces a broader norm of scientists checking each other's work.

A distinguishing feature of science is that its claims are expected to be verifiable. This has been long appreciated by scientists and philosophers. In 1660 the Royal Society, the oldest continuously operating scientific society in the world, adopted as its motto the Latin phrase *nullius in verba*, meaning "take nobody's word for it." Merton (1942) wrote that scientific norms are upheld through scrutiny and verification of our work by other scientists. For Popper (1959), scientific statements must be intersubjectively testable, capable of being evaluated similarly by any person thinking rationally about them. Lupia and Elman (2014) argued that scientific claims get their credibility from being publicly available for inspection and critique.

Verifiability is a broad idea that can be applied to every step in the derivation of a scientific claim, and it is the basis for many current norms, practices, and expectations in scientific discourse. If an author of a paper asserts that a prediction follows from some theory, scientific norms hold that the arguments should be presented in enough detail that readers can evaluate whether the theory is coherent and the prediction does indeed follow. If an author asserts that a measurement procedure produces valid scores, scientists expect the author to support that claim with data or through citations to previous validity studies. If an author employs a novel statistical method, scientists expect to be able to verify how the method works by inspecting proofs or the results of simulations.

Verifiability is important for empirical results, as well. Consider a famous example from philosophy, the black swan. Once upon a time, Europeans believed that all swans were white. But in the seventeenth century, Dutch explorers reported that they had observed black swans in Australia, forcing Europeans to update their beliefs about the possible colors of swans. In the stylized telling of this story, this was a purely logical operation: The empirical statement "black swans exist" falsifies the theoretical statement "all swans are white" via *modus tollens*. But the logic applies only after we accept the premises. Philosophers have long recognized that scientists are not required to accept a premise like "black swans exist" just because someone says they do, or else theories would have to accommodate all kinds of

outlandish claims (Mill 1882/2014). Instead, scientists establish methodological rules to decide whether to refuse or admit new findings into the corpus of scientific knowledge, with verifiability being one key principle guiding those rules (Meehl 1990b; Popper 1959). There are many reports from people who say they have seen black swans, but there are also many reports from people who say they have seen Bigfoot. Part of why scientific taxonomies include black swans but not Bigfoots is because of differences in the verifiability of the reports.

Replication research makes scientific knowledge more verifiable in several ways. The most straightforward way, and the one that receives the greatest emphasis by Zwaan et al. in the target article, is that direct replication studies are a way to evaluate empirical claims and decide which ones to accept as dispositive for theories. If a result is a statistical fluke, or if the necessary conditions for getting a result include critical elements that were not specified in the original report, replication research will produce discrepant results and scientists may be justifiably cautious in updating their theories. Conversely, because new empirical results are almost never definitive (Srivastava 2011), a successful direct replication can increase scientists' confidence in the definitiveness of a claim.

A second way replication research can advance verifiability is by promoting open dissemination of data, analysis code, materials, and details of experimental procedures. All of these things help researchers carry out high-quality direct replications, so norms and policies that are designed to enable replications should promote their dissemination. But all of them are useful for other kinds of verification beside replication. Open data and code allow other scientists to evaluate how well statistical analyses support conclusions. Open materials and procedures allow other scientists to evaluate whether research protocols work as intended. Thus, even if no replication study is ever conducted, when original authors provide the information that makes replication possible, they make other aspects of verification possible too.

A third way that replication research promotes verifiability is through the accumulation of meta-scientific knowledge. The claim "this experimental result is verifiable" can be treated like a hypothesis and tested empirically in a direct replication. But in practice, scientists do not have the time or resources to replicate every single experiment. Instead they must rely on meta-scientific ideas about which results are more or less credible. Should scientists put more trust in findings with low p values? Studies with more thorough disclosures? Papers by high-status authors? When scientists read one another's work, they make informal judgments about how much they believe the results. Replication research, especially large-scale systematic replication research like the Many Labs projects (e.g., Ebersole et al. 2016a; Klein et al. 2014a) can tell us what makes research more replicable and help scientists make those judgments in a more principled way.

A fourth way that replication research promotes verifiability is by reinforcing it as a norm of scientific work. As Merton (1942) noted, it is unrealistic to expect individual scientists to have only pure scientific motives. Many scientists want to discover true things about nature; but they also want to make money, gain status, look and feel smart, confirm their preconceptions, and all sorts of other entirely human things. Psychologists should especially appreciate that individual scientists cannot be expected to perfectly manage their own conflicting motivations and biases. So scientific disinterestedness must be upheld through the social process of scientists scrutinizing and verifying each other's claims. The more replication research is supported formally through professional incentives, and informally through scientists speaking favorably about replication as valued work and treating it as an expected part of their jobs, the more it will promote a broader culture that values verification of scientific knowledge above the advancement of scientists' individual interests.

What *have* we learned? What *can* we learn?

doi:10.1017/S0140525X18000870, e151

Fritz Strack[a] and Wolfgang Stroebe[b]

[a]*Department of Psychology, University of Würzburg, Würzburg 97070, Germany;* [b]*Department of Social Psychology, University of Groningen, 9712 TS Groningen, The Netherlands*

fritz_strack@yahoo.de **wolfgang.stroebe@rug.nl**

http://www.i2.psychologie.uni-wuerzburg.de/

http://stroebe.socialpsychology.org/

Abstract: We advocate that replications should be an integral part of the scientific discourse and provide insights about the conditions under which an effect occurs. By themselves, mere nonreplications are not informative about the "truth" of an effect. As a consequence, the mechanistic continuation of multilab replications should be replaced by diagnostic studies providing insights about the underlying causes and mechanisms.

As Zwaan et al. have repeatedly emphasized, replication is a cornerstone in the edifice called "science." It is therefore somewhat surprising that its structure has not yet collapsed under the burden of the virulent "replication crisis." Instead, lively discussions have sprung up focusing on various characteristics of replication endeavors, for example, about the distinction between "direct" procedural and theoretically driven "conceptual" replications (Stroebe & Strack 2014). In this comment, we discuss the benefits of the frequent replication exercises and suggest a more fruitful way to pursue in the future.

There is no doubt that thus far, the number of "replication failures" has attracted great public attention. "Over half of psychology studies fail reproducibility test" was the title of a cover story in *Nature* (Baker 2015), and frequently, the news that "yet another classic study was not replicated" makes the headlines in popular newspapers and magazines. Typically, these replication failures were based on an attempted copy of the original study that did not reach conventional significance levels and/or yielded effect sizes below the original standard. Rarely were there significant reversals of the original outcomes.

What can be learned from such procedural nonreplications? There are two possibilities. Other than suspecting fraud, the original effect may be diagnosed as fragile and not sufficiently "robust" to emerge under changing conditions. Second, the effect is declared as being not "real" or "true," but merely a "false positive."

To be sure, it is certainly important to ascertain the robustness of an effect if a procedure is meant to be employed as an intervention in a natural setting. There, it is more important to determine the *effectiveness* of an intervention than to understand its *underlying mechanisms*. For this purpose, procedural replications are diagnostic and highly appropriate. For example, to assess the therapeutic value of a cancer drug, it is necessary to ascertain its effectiveness under varying contexts.

Understanding underlying mechanisms, however, requires us to resort to the theoretical level and to assess the validity of the theory. Unlike experimental effects, theoretical statements have a truth value that can be supported or undermined by empirical findings. Theories, however, are formulated on a level that transcends the concrete evidence, and their validity does not rest on the outcome of one specific experimental paradigm. Thus, even if a procedural nonreplication implies a lack of reliability for a concrete effect, consequences about truth and falseness of an otherwise well-supported theory must be weighed in the context of supporting evidence from these other investigations.

In particular, inferences about truth or falseness that are based solely on selected parameters from Null Hypothesis Significance Testing are highly problematic. Recently, this statistically myopic approach of determining truth or falsity has been sharply criticized. For example, Blakeley et al. (2018) have advocated moving "beyond the paradigm of routine 'discovery,' and binary statements about there being 'an effect' or 'no effect,' to one of continuous and inevitably flawed learning that is accepting

of uncertainty and variation" (p. 9). In a similar vein, the Deutsche Forschungsgemeinschaft (German National Science Foundation) has issued a statement on the "Replicability of Research Results" (2017), arguing that "ascertaining the replicability or non-replicability of a scientific result is itself a scientific result. As such, it is not final; rather, like all scientific knowledge, it is subject to methodological skepticism and further investigation" (p. 2).

These insights imply the lack of a direct route from statistical parameters to scientific truth (Strack 2017). Instead, empirical data must be recognized as arguments that must enter the scientific discourse to persuade its participants. For this purpose, the collection of data must be guided by theoretically grounded hypotheses that generate specific predictions. Such hypothesis-guided approaches are conspicuously missing when it comes to procedural replications. As a consequence, little insight is gained about the underlying causal dynamics.

If science has the goal of finding the causes of things ("rerum causas cognoscere"), it is not enough to merely devalue a finding as "random." Instead, replicators must make the effort to identify the actual determinants of the original finding if they believe that it was not caused by the factors claimed in the original study. In other words, the evidence of a replication should consist of an interaction where the original finding is replicated under certain conditions, but not under others. To be sure, making a theoretical statement contingent on the fulfillment of other conditions may decrease its falsifiability. Zwaan et al. are right in pointing out that repeatedly adding conditions may cause a research program to become "degenerative." On the other hand, the price of being closer to the truth may be a decreased theoretical power. At the same time, proposing new theories that go beyond what has been known so far and include previous causalities as special cases is always more desirable.

When it comes to theory testing, it is not justified to yield replications a special status that sets them apart from the theoretical discourse. It is not justified to assign the original researcher the onus of identifying the contextual confound that may have prevented the effect from occurring in the replication. Like any other basic research, replications should be driven by theoretical assumptions generating results that go beyond demonstrating a mere failure to replicate and identifying yet another "false positive." To advance scientific knowledge, replicators should be able to identify the reasons why an effect did not replicate. Without such evidence, simple nonreplication of a finding that is part of a well-supported theory is uninformative. Thus, to answer the question stated in our title, the theoretical knowledge gained from replication failures has been negligible. Therefore, mechanistically continuing multilab replications of yet another set of empirical studies is a waste of valuable scientific resources that would better be employed in developing and testing better and more powerful theories.

Conceptualizing and evaluating replication across domains of behavioral research

doi:10.1017/S0140525X18000882, e152

Jennifer L. Tackett[a] and Blakeley B. McShane[b]

[a]*Psychology Department, Northwestern University, Evanston, IL 60208;* [b]*Kellogg School of Management, Northwestern University, Evanston, IL 60208.*

jennifer.tackett@northwestern.edu
b-mcshane@kellogg.northwestern.edu
https://www.jltackett.com/
http://www.blakemcshane.com/

Abstract: We discuss the authors' conceptualization of replication, in particular the false dichotomy of direct versus conceptual replication intrinsic to it, and suggest a broader one that better generalizes to other domains of psychological research. We also discuss their approach to the evaluation of replication results and suggest moving beyond their dichotomous statistical paradigms and employing hierarchical/meta-analytic statistical models.

We thank Zwaan et al. for their review paper on replication and strongly endorse their call to make replication mainstream. Nonetheless, we find their conceptualization of and recommendations for replication problematic.

Intrinsic to Zwaan et al.'s conceptualization is a false dichotomy of direct versus conceptual replication, with the former defined as "a study that attempts to recreate the critical elements (e.g., samples, procedures, and measures) of an original study" (sect. 4, para. 3) and the latter as a "study where there are changes to the original procedures that might make a difference with regard to the observed effect size" (sect. 4.6). We see problems with both of Zwaan et al.'s definitions and the sharp dichotomization intrinsic to their conceptualization.

In terms of definitions, first, Zwaan et al. punt in defining direct replications by leaving unspecified the crucial matter of what constitutes the "critical elements (e.g., samples, procedures, and measures) of an original study" (sect 4., para. 3). Specifying these is nontrivial if not impossible in general and likely controversial in specific. Second, they are overly broad in defining conceptual replications: Under their definition, practically all behavioral research replication studies would be considered conceptual. To understand why, consider large-scale replication projects such as the Many Labs project (Klein et al. 2014a) and Registered Replication Reports (RRRs; Simons et al. 2014) where careful measures were taken such that protocols were followed identically across labs in order to achieve near exact or direct replication. In these projects, not only did observed effect sizes differ across labs (as they always do), but so too did, despite such strict conditions, true effect sizes; that is, effect sizes were heterogeneous or contextually variable – and to roughly the same degree as sampling variation (McShane et al. 2016; Stanley et al. 2017; Tackett et al. 2017b). This renders Zwaan et al.'s suggestion of conducting direct replication infeasible: Even if defining the "critical elements" were possible, recreating them in a manner that maintains the effect size homogeneity they insist on for direct replication seems impossible in light of these Many Labs and RRR results.

In addition, and again in light of these results, the sharp dichotomization of direct versus conceptual replication intrinsic to Zwaan et al.'s conceptualization is unrealistic in practice. Further, even were it not, replication designs with hybrid elements (e.g., where the theoretical level is "directly" replicated but the operationalization is systematically varied) are an important future direction – particularly for large-scale replication projects (Tackett et al. 2017b) – not covered by Zwaan et al.'s conceptualization.

Instead, and in line with Zwaan et al.'s mention of "extensions," we would like to see a broader approach to conceptualizing replication and, in particular, one that better generalizes to other domains of psychological research. Specifically, large-scale replications are typically only possible when data collection is fast and not particularly costly; thus they are, practically speaking, constrained to certain domains of psychology (e.g., cognitive and social). Consequently, we know much less about the replicability of findings in other domains (e.g., clinical and developmental) let alone how to operationalize replicability in them (Tackett et al. 2017a; in press). In these other domains,

where data collection is slow and costly but individual data sets are typically much richer, we recommend that in addition to the prospective approach to replication employed by large-scale replication projects thus far, a retrospective approach that leverages the large amount of shareable archival data across sites can be valuable and sometimes even preferable (Tackett et al. 2017b; 2018).

This will require not only a change in both infrastructure and incentive structures, but also a better understanding of appropriate statistical approaches for analyzing pooled data (i.e., hierarchical models) and more complex effects (e.g., curve or function estimates as opposed to point estimates); lab-specific moderators most relevant to include in such analyses; additional method factors that drive heterogeneity (e.g., dropout mechanisms in longitudinal studies); and how to harmonize measurements across labs (e.g., if they use different measures of depression).

It may also require a change in procedures for statistically evaluating replication. Zwaan et al. suggest three ways of doing so, all of which are based on the null hypothesis significance testing paradigm and the dichotomous p-value thresholds intrinsic to it. Such thresholds, whether in the form of p-values or other statistical measures such as confidence intervals and Bayes factors (i) lead to erroneous reasoning (McShane & Gal 2016; 2017); (ii) are a form of statistical alchemy that falsely promise to transmute randomness into certainty (Gelman 2016a), thereby permitting dichotomous declarations of truth or falsity, binary statements about there being "an effect" or "no effect," a "successful replication" or a "failed replication"; and (iii) should be abandoned (Leek et al. 2017; McShane et al. 2018).

Instead, we would like to see replication efforts statistically evaluated via hierarchical / meta-analytic statistical models. Such models can directly estimate and account for contextual variability (i.e., heterogeneity) in replication efforts, which is critically important given, as per the Many Labs and RRR results, that such variability is roughly comparable to sampling variability even when explicit efforts are taken to minimize it as well as the fact that it is typically many times larger in more standard sets of studies when they are not (Stanley et al. 2017; van Erp et al. 2017). Importantly, they can also account for differences in methods factors such as dependent variables, moderators, and study designs (McShane & Bockenholt 2017; 2018) and for varying treatment effects (Gelman 2015), thereby allowing for a much richer characterization of a research domain and application to the hybrid replication designs discussed above. We would also like to see the estimates from these models considered alongside additional factors such as prior and related evidence, plausibility of mechanism, study design, and data quality to provide a more holistic evaluation of replication efforts.

Our suggestions for replication conceptualization and evaluation forsake the false promise of certainty offered by the dichotomous approaches favored by the field and by Zwaan et al. Consequently, they will seldom if ever deem a replication effort a "success" or a "failure," and indeed, reasonable people following them may disagree about the degree of replication success. However, by accepting uncertainty and embracing variation (Carlin 2016; Gelman 2016a), we believe these suggestions will help us learn much more about the world.

Making prepublication independent replication mainstream

doi:10.1017/S0140525X18000894, e153

Warren Tierney,[a] Martin Schweinsberg,[b] and Eric Luis Uhlmann[c]

[a]*Kemmy Business School, University of Limerick, Castletroy, Limerick V94 T9PX, Ireland;* [b]*ESMT Berlin, 10178 Berlin, Germany;* [c]*Organisational Behaviour Area, INSEAD, 138676, Singapore.*

warrentierney@hotmail.com
martin.schweinsberg@esmt.org eric.luis.uhlmann@gmail.com
https://ie.linkedin.com/in/warrentierney
https://www.esmt.org/person/martin-schweinsberg
http://socialjudgments.com/

Abstract: The widespread replication of research findings in independent laboratories prior to publication is suggested as a complement to traditional replication approaches. The pre-publication independent replication approach further addresses three key concerns from replication skeptics by systematically taking context into account, reducing reputational costs for original authors and replicators, and increasing the theoretical value of failed replications.

The reproducibility of scientific findings, whereby a study is replicated by independent investigators in order to assess the robustness of the research and of its findings, is fundamental to the scientific process (Dunlap 1926; Popper 1959). Overall, we strongly agree with the authors of the target article that replication should be made mainstream. Although replication is typically discussed in terms of reproducing previously published work, we further advocate for making mainstream the independent replication of findings prior to publication (see also Schooler 2014). Pre-publication independent replication (PPIR) is a collaborative, crowdsourced approach to science in which original study authors nominate their own findings to be replicated in independent laboratories around the world. This approach complements existing replication initiatives that focus on published findings and has different strengths and weaknesses. Importantly, PPIR further addresses three of the key concerns from replication skeptics counterargued so effectively in the target article.

In our first pre-publication independent replication initiative (Schweinsberg et al. 2016; Tierney et al. 2016), 10 unpublished moral judgment effects from the last author's research pipeline were replicated by 25 independent research groups who collected data from more than 11,000 participants. The findings were mixed – while some studies replicated successfully, others did not replicate according to the *a priori* established criteria. Overall, 6 findings successfully replicated; one study replicated but with a much smaller effect size than the original (a decline effect [Schooler 2011]), two findings were not supported, and one study was culturally moderated (replicating consistently in the original country but not in five other countries). The culturally moderated effect provides evidence that contextual factors can play an important and unexpected role in replications. In total, 40% of the original findings failed at least one major criterion for reproducibility.

We have expanded the scope of our crowdsourcing approach in a second PPIR initiative, the Pipeline Project 2. This initiative opens pre-publication independent replication to the world, providing original authors the opportunity to nominate their unpublished work for replication in partner laboratories as well as graduate methods classes.

We currently have 13 original findings being replicated at more than 50 sites around the world (Schweinsberg et al., in preparation). Original authors opt into the PPIR process and help select replicators they regard as suitable and as having access to relevant subject populations, leading to collaborative rather than adversarial interactions. Notably, original authors are asked to specify beforehand in what cultures and research sites they do and do not expect their effect to emerge. We are further conducting a prediction market (Dreber et al. 2015) to see if members of the scientific community at large can anticipate contextual variability in effects. These aspects of the PPIR process further address a key challenge raised by replication skeptics, by systematically taking context into account.

Concerns have also been raised about reputational damage to those involved in replications, both to original authors whose published findings are not reproduced by other research groups, and replicators whose results question established findings (Bohannon 2014; Kahneman 2014; Schnall 2014a; 2014b; 2014c). By replicating findings in independent laboratories before (rather than after) the findings are published, PPIRs minimizes reputational costs to both original authors and replicators because (1) no one's reputation depends on the outcome, and (2) original authors voluntarily opt into the PPIR process and help select their replicators.

Another common argument is that failed replications are uninterpretable and low in theoretical value (Schnall 2014a; 2014b; 2014c). Although in our view replications are always informative and valuable (Dreber et al. 2015), it is at the same time true that there are other plausible explanations for null findings other than the original effect being false (Open Science Collaboration 2015). We suggest that the theoretical value of PPIR in terms of identifying false positives is even higher than for traditional replications, because most alternative explanations for null effects are ruled out. In particular, defenders of the original finding have little basis to attribute an unsuccessful replication to a lack of replicator expertise or use of irrelevant subject populations, because the original authors helped select what they regarded as qualified replicators and specified a priori which participant populations they expected to exhibit the effect. However, informational value is correspondingly lower for successful PPIRs, relative to traditional replications, because the original authors participate in selecting their own replicators, who may be biased in favor of the hypothesis. Indeed, research demonstrates that the theories investigators endorse strongly predict the effect sizes they obtain (Berman & Reich 2010).

The biggest challenge to making pre-publication independent replication mainstream is the lack of professional incentives, especially for replicators. One potential solution is to build PPIRs into the education of graduate students (Everett & Earp 2015) as part of crowdsourced projects on which they and the instructors of their methods courses are co-authors. These student PPIRs can examine findings that the original authors identify as straightforward for a junior researcher to conduct. To facilitate the integration of pre-publication independent replication into graduate methods courses, as part of the Pipeline Project 2 we have developed an open source curriculum on Crowdsourcing Science including instructions for student PPIR projects (https://osf.io/hj9zr/). Researchers of any level of experience who wish to initiate projects can use the Study Swap website (https://osf.io/view/StudySwap/), a new forum where interested parties can engage with the PPIR process, both as original authors looking for labs to replicate their findings or as independent investigators looking to replicate findings. Networks of partner laboratories such as the Psychological Science Accelerator (Chartier 2017) might also be leveraged to conduct replications of unpublished, rather than published findings.

In sum, conducting independent replications earlier in the research process – before findings are even submitted for publication – can further address what the target article identifies as three of the key concerns raised by skeptics of replication. The pre-publication independent replication approach minimizes reputational costs to original authors and replicators, systematically takes into account context, and maximizes the informational value of failed replications.

Scientific progress is like doing a puzzle, not building a wall

doi:10.1017/S0140525X18000900, e154

Alexa M. Tullett[a] and Simine Vazire[b]
[a]*Department of Psychology, University of Alabama, Tuscaloosa, AL 35487-0348;* [b]*Department of Psychology, University of California, Davis, California 95616.*
svazire@ucdavis.edu alexa.tullett@gmail.com
http://alexatullett.com
http://simine.com

Abstract: We contest the "building a wall" analogy of scientific progress. We argue that this analogy unfairly privileges original research (which is perceived as laying bricks and, therefore, constructive) over replication research (which is perceived as testing and removing bricks and, therefore, destructive). We propose an alternative analogy for scientific progress: solving a jigsaw puzzle.

Many scientists, including the authors of the target article, use the metaphor of "building a wall" for accumulating scientific knowledge. In this analogy, correcting a false positive through replication is compared to removing a faulty brick from the wall. This leads to the widespread view that conducting replications only has the potential to eliminate knowledge and not to add to knowledge. Here we contend that this view is not only misguided, but also detrimental to the field.

In their target article, Zwaan et al. note that a common concern about direct replications is that they have little theoretical value (concern II, sect. 5.2). For example, according to Stroebe and Strack (2014, p. 63), "one reason why exact replications are not very interesting is that they contribute little to scientific knowledge." Their account portrays original research as a constructive process – the laying of the bricks – and replication as a secondary process – the testing of the bricks. Replication research, which can only tear down weak parts of the wall (or confirm that existing bricks are strong), is considered of lower theoretical and scientific value than original research, which can help build new parts of the wall.

There are some senses in which this analogy is quite apt. For example, because conducting an original study requires design, it often involves more work (and more creativity) than does running a replication. The original study is also generative in the sense that it posits the idea, whereas the replication tests an existing idea. For these reasons, our

field's emphasis on original research doesn't seem entirely misplaced. One skill we should value among scientists is the ability to come up with novel ideas, or with novel ways of testing existing ideas.

There are other aspects of this analogy, however, that we see as more problematic. First, it creates a false distinction between original and replication studies, treating the first study as a greater contribution to knowledge than those that follow. Science does not care whether data come from a replication or an original study. The order in which these explorations take place is orthogonal to the degree to which they advance our understanding. Gelman (2016b) illustrates the problem with this way of thinking using the "time-reversal heuristic" in which a reader is asked to imagine that the replication study was done first. This exercise is meant to show that the order in which studies are conducted should be irrelevant to our evaluation of their scientific value.

A further weakness of the "building a brick wall" analogy is that equating an original study to a new brick suggests that researchers are in the business of inventing psychological phenomena. In reality, we are simply trying to understand the psychological phenomena that already exist. If we think of effects as existing in the world, prior to any particular scientific investigation of them, we realize that the distinction between original research and replication research is arbitrary from the point of view of quantifying the evidence provided.

In addition to being inaccurate, it may also be harmful to think of replication research as fundamentally different from original research. When original studies are presented as constructive and knowledge-producing (i.e., "laying the bricks"), whereas replication studies are presented as a mechanical, auditing activity (i.e., "testing the bricks"), this reduces the perceived value of, and therefore incentive to conduct, replication research (Crandall & Sherman 2016). This is problematic given that replication is one of the defining features of a scientific field. Moreover, giving privileged status to original research contributes to the persistence of false-positive original findings in the literature, in textbooks, and in the media. When the scientific evidence produced by replication studies is perceived as different from (and often lesser than) the evidence produced by original studies, false claims from original studies become even harder to correct.

With these ideas in mind, we propose a new metaphor for scientific progress. Rather than likening scientific progress to building a wall, we suggest the analogy of solving a jigsaw puzzle. First, this highlights the fact that we are not builders but discoverers; we are not creating phenomena, but instead trying to reveal a pre-existing reality. Second, it highlights the nonlinear nature of progress; realizing that a piece is in the wrong spot is just as valuable as putting it in the right spot.

With this metaphor, it becomes apparent that original research and replication research are not as different as we might think. The order in which studies are carried out is not important for their evidentiary value; what is important is whether the result improves our understanding of reality (i.e., whether the puzzle piece is in the right place). Moreover, the misguided ideas of "constructive" and "destructive" research are avoided when thinking about solving a jigsaw puzzle. Any new evidence that moves us closer to an accurate solution is a constructive step in the process.

Holding replication studies to mainstream standards of evidence

doi:10.1017/S0140525X18000912, e155

Duane T. Wegener[a] and Leandre R. Fabrigar[b]

[a]*OSU Psychology Department, Columbus, OH 43210;* [b]*Department of Psychology, Queen's University, Kingston, Ontario, K7L 3N6, Canada.*

Wegener.1@osu.edu **Fabrigar@queensu.ca**

https://psychology.osu.edu/people/wegener.1

http://www.queensu.ca/psychology/people/faculty/lee-fabrigar

Abstract: Replications can make theoretical contributions, but are unlikely to do so if their findings are open to multiple interpretations (especially violations of psychometric invariance). Thus, just as studies demonstrating novel effects are often expected to empirically evaluate competing explanations, replications should be held to similar standards. Unfortunately, this is rarely done, thereby undermining the value of replication research.

Zwaan et al. provided a useful summary of key issues in the role of replication in psychological research. The article will serve as a useful resource for scholars. However, their coverage of some issues failed to address important qualifications to their conclusions. In the interest of brevity, we highlight one such example.

The authors argue that direct replications should be more prominent in the literature, in part, because they have substantial theoretical value (see in particular concern II, sect. 5.2). We agree that direct replications can sometimes make valuable theoretical contributions. However, such contributions only become likely to the degree that replications are held to the same standards of evidence as studies demonstrating novel effects. Unfortunately, the present discussion (along with many others) implicitly adopts a different standard of evidence than is customary for studies of novel effects.

It is useful to consider the nature of psychological hypotheses and the evidence typically required of studies exploring those hypotheses. The hypotheses being tested generally link two or more psychological or behavioral constructs. For example, "frustration leads to aggression" links the psychological experience of frustration to the outcome of aggression. When an original study claims support for such a hypothesis, it is because a measure or manipulation of frustration is empirically associated with a measure of aggression. For any given study, though, reviewers or editors might question the extent to which the chosen manipulation or measure adequately reflects the construct of interest or question the proposed mechanism linking the constructs. Selective journals routinely require the researcher to empirically evaluate the viability of competing explanations. That is, the demonstration of a novel "effect" is considered to be of limited theoretical value if it is open to multiple interpretations, particularly if one or more of those interpretations is uninteresting or falls outside the focal theory (such as demand artifacts, placebo effects, confounds in a manipulation or measure, or an alternative psychological mechanism). As a result, the testing of a novel theory routinely requires a programmatic approach involving multiple studies.

Unfortunately, results of direct replications are frequently open to multiple interpretations, particularly when they fail to produce the original effect, and many potential explanations are uninteresting or fall outside the replicator's preferred explanation. For example, statistical

problems (e.g., inadequate power or severe violations of underlying statistical assumptions) or violations of psychometric invariance (e.g., differences between studies in the construct validity of a manipulation or measure) would often be of little substantive interest (e.g., see Fabrigar & Wegener 2016; Stroebe & Strack 2014). Replicators have paid attention to statistical power but have often ignored other alternative accounts of their effects (such as failures of psychometric invariance [Fabrigar & Wegener 2016]). Imagine that a researcher attempted to replicate a study originally conducted in the early 1980s using a clip from *Three's Company* to produce positive mood. If a replication study fails to show the effect because the positive mood induction is no longer humorous to contemporary participants, this would not constitute a notable theoretical advance (presuming the goal of the original research was to understand mood effects rather than the psychology of *Three's Company* or 1980s American sitcoms).

Other explanations for failing to replicate might be theoretically interesting, such as differences in the characteristics of the study participants or features of the experimental context changing the nature of the relations between the psychological constructs of interest. However, such insights are possible only if the relevant participant differences or contextual influences are identified. Likewise, concluding that the original study was a false positive could be a valuable contribution. However, that statistical explanation is convincing only after alternative explanations have been evaluated and rejected (just as support for a novel theory becomes convincing only after alternative plausible explanations have been evaluated and rejected).

Zwaan et al. did acknowledge that changes in contexts or participants might require changes in study materials even when the goals of the research are of "direct replication." That acknowledgment takes a step toward the approach we are advocating compared with some direct replication efforts. However, neither the present article nor many others place a strong emphasis on evaluating competing explanations for a replication study's findings. In failing to do so, such articles suggest that replication studies advance theory even when the implications of their findings are highly ambiguous. To the contrary, we suggest that replication studies open to many alternative explanations are no more theoretically valuable than an original study open to many alternative explanations. Replication advocates often seem to view alternatives to false-positive conclusions as if they are "excuses" or "dodges" offered by the original researchers. Excuses or dodges might sometimes be offered, but psychometric invariance of manipulations and measures, contextual moderators, and individual difference moderators are not "dodges." They are standard methodological and theoretical considerations. They can often be specified in advance and evaluated before and after a replication study has been undertaken. These considerations parallel the kinds of considerations routine in evaluating alternative explanations for original research results. Putting aside such considerations in the case of replications only weakens their empirical and theoretical utility.

In practice, researchers undertaking direct replications have rarely attempted a systematic exploration of competing explanations for their findings. For example, the Many Labs initiative conducts tests of previously demonstrated effects, but does not follow up these tests with multistudy assessments of plausible explanations (e.g., Ebersole et al. 2016a). Instead, it has been left to the original researchers to explain discrepant findings and provide initial empirical evaluations of the alternatives (e.g., Luttrell et al. 2017; Petty & Cacioppo 2016). In some cases, replication failures have stemmed from violations of psychometric invariance comparing the replication with the original research (Ebersole et al. 2017; Luttrell et al. 2017).

In summary, we do not have a problem with replication being an important part of mainstream psychological science, but benefits of that effort will be most likely to the extent that replications are evaluated in ways that parallel evaluation of original research (i.e., holding each to the same standards of evidence).

Data replication matters to an underpowered study, but replicated hypothesis corroboration counts

doi:10.1017/S0140525X18000924, e156

Erich H. Witte[a] and Frank Zenker[b]

[a]*Institute for Psychology, University of Hamburg, 20146 Hamburg, Germany;* [b]*Departments of Philosophy and Cognitive Science, Lund University, SE-221 00 Lund, Sweden.*

witte_e_h@uni-hamburg.de frank.zenker@fil.lu.se

https://www.psy.uni-hamburg.de/personen/prof-im-ruhestand/witte-erich.html

http://www.fil.lu.se/en/person/FrankZenker/

Abstract: Before replication becomes mainstream, the potential for generating theoretical knowledge better be clear. Replicating *statistically significant nonrandom data* shows that an original study made a discovery; replicating a *specified theoretical effect* shows that an original study corroborated a theory. Yet only in the latter case is replication a necessary, sound, and worthwhile strategy.

Bakker et al. (2012) state the average replication probability of empirical studies in psychology as $1-\beta$ error = 0.36. Originating in Neyman-Pearson test theory (NPTT), the $1-\beta$-error is also known as test power. Prior to collection of data, it estimates the probability that a replication attempt duplicates an original study's data signature. Because $1-\beta$-error = 0.36 "predicts" the estimated actual replication rate of 36% that Open Science Collaboration (2015) report, we may cautiously interpret the rate as a consequence of realizing NPTT empirically.

In seeming "fear" that a random process (H_0) may have generated our data (D), we (rightly) demand that D feature a low α-error, $p(D, H_0) \leq \alpha = 0.05$. We nevertheless regularly allow such "low α" data to feature a high average β-error = 0.64 (Open Science Collaboration 2015). A similarly high β-error value is unproblematic, of course, if we use a *composite* H_1 hypothesis. For it simply fails to point-specify the effect size (postulated by the H_1) that calculating the β-error presupposes. So we cannot but ignore the replication probability of data.

By contrast, point-specifying three parameters – the α-error, the actual sample size (N), and the effect-size – lets NPTT infer the β-error. By the same logic, point-specifying the effect size, as well as the α- and β-error (e.g., $\alpha = \beta \leq 0.05$) lets NPTT infer the minimum sample size sufficing to register this effect as a statistically significantly nonrandom data signature. Hence, NPTT generally serves to plan *well-powered* studies.

To an underpowered original study – one featuring $1-\beta$-error < 0.95, that is–successful data replication thus matters, for this raises our confidence that the original study made a discovery. A well-powered original study, however, already features $\alpha = \beta < 0.05$. Hence, if the replication attempt's error probabilities are at least as large (as is typical), then replicating a well-powered study's nonrandom data signature *restates*, but it cannot raise our confidence that successful data replication is highly probable. Except where we can decrease the error probabilities, therefore, fallible knowledge that replicating the data signature of a well-powered study is highly probable does *not* require actual data replication.

What the target article's authors call "direct replication" thus amounts to a data replication. For rather than use a theory to point-predict an effect, we use the actual N, the actual α-error, and a stipulated β-error to *induce* the effect size from data. A direct replication we must assess by estimating its test power, itself calculable only if the H_0 and H_1 hypotheses are both point-specified. Here, the H_0 invariably states a random data distribution. In case the point effect the H_1 postulates is uncertain, we may alternatively predict an interval H_1 hypothesis. (Its endpoints qualify as theoretical predictions, and the midpoint as a theoretical assumption.) We consequently obtain test power either as a point value or as an interval.

In both cases, calculating test power lets our methodological focus shift from the *discovery* to the *justification* context (Witte & Zenker 2017b). In the former context, we evaluate data given hypotheses by studying the error rates of data given the H_0 and H_1 distributions, and so compare $p(D, H_0)$ with $p(D, H_1)$. In the latter context, by contrast, we evaluate hypotheses given data by studying the likelihood ratio (LR) $L(H_1|D)/L(H_0|D)$. Because a *fair* test assigns equal priors, $p(H_0) = p(H_1)$; this makes the LR numerically identical to the Bayes factor. Moreover, setting the hypothesis corroboration threshold to $(1-\beta\text{-error})/\alpha\text{-error}$ makes it a Wald test (Wald 1947). Desirably, as N increases, test results thus asymptotically approach the percentages of false-positive and false-negative errors.

Data replication then matters, but what counts is the replicated corroboration of a theoretical hypothesis, as per $LR_{H1/H0} > (1-\beta\text{-error})/\alpha\text{-error}$. This the target article's authors call "conceptual replication." Compared with an H_1 that merely postulates significantly nonrandom data, the theory-based point-specified effect a conceptual replication presupposes is more informative, of course. We can hence do more than run mere twin experiments. Crucially, as one accumulates the likelihoods obtained from individual experiments, *several* conceptually replicated experiments together may (ever more firmly) jointly corroborate, or falsify, a theoretical prediction (Wald 1947; Witte & Zenker 2016a; 2016b; 2017a; 2017b). (Psychology could only gain from accumulating such methodologically well-hardened facts; see Lakatos [1978].) Provided we test a point effect *fairly*, then, conceptual replication is a genuine strategy to probabilistically support, or undermine, a theoretical construct.

As to how psychological tests correspond to theoretical variables, several different measures currently serve to validate tests (Lord & Novick 1968). In fact, accepting one such test as a measurement procedure for a dispositional variable (e.g., personality, intelligence) lets this test dictate how we estimate the focal variable practically. A comparable strategy to validate experiments, by contrast, seems to be missing. Perhaps it is for this reason that disagreements regarding an experiment's quality often appear purely subjective.

From a theoretical viewpoint, however, test validation strategies are equivalent to experiment validation strategies, for "[t]he validity coefficient *is* the correlation of an observed variable with some theoretical construct (latent variable) of interest" (Lord & Novick 1968, p. 261, italics added). Indeed, this identity is what warrants our interpreting an experimental setting as the empirical realization of a theoretical construct. We may consequently treat the difference, or correlation, between the individual measurements in the experimental and control groups as an experiment's *validity coefficient*.

This difference/correlation is valid only if we can exclude alternative explanations that cite various internal or external influences. Compared with the significant workload that preregistration approaches require (Hagger et al. 2016), for instance, validating an experiment is yet more effortful. For we must establish that (1) participants can, and do, interpret our experimental setting as intended; (2) they are motivated to display the corresponding behavior; and (3) an independent observer can adequately evaluate their reactions (Witte & Melville 1982). Indeed, an overly simplistic manipulation check is "an obstacle toward cumulative science" (Fayant et al. 2017, p. 125). Therefore, successfully replicating a point-specified effect is *sound* only if each individual experiment is valid.

In sum, an experiment validation strategy that renders worthwhile the efforts of constructing a valid experiment should rest not on data alone, but also on how well we theoretically predict a focal phenomenon. Only if several labs then achieve a replicated hypothesis corroboration (by testing the LR fairly) could replication provide the gold standard that a theoretically progressive version of empirical psychology requires.

Authors' Responses

Improving social and behavioral science by making replication mainstream: A response to commentaries

doi:10.1017/S0140525X18000961, e157

Rolf A. Zwaan,[a] Alexander Etz,[b] Richard E. Lucas,[c] and M. Brent Donnellan[c,1]

[a]*Department of Psychology, Education, and Child Sciences Erasmus University, 3000 DR, Rotterdam, The Netherlands;* [b]*Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100;* [c]*Department of Psychology, Michigan State University, East Lansing, MI 48824*

zwaan@essb.eur.nl
etz.alexander@gmail.com
lucasri@msu.edu
donnel59@msu.edu
https://www.eur.nl/essb/people/rolf-zwaan
https://alexanderetz.com/
https://www.msu.edu/user/lucasri/
https://psychology.msu.edu/people/faculty/donnel59

Abstract: The commentaries on our target article are insightful and constructive. There were some critical notes, but many commentaries agreed with, or even amplified our message. The

first section of our response addresses comments pertaining to specific parts of the target article. The second section provides a response to the commentaries' suggestions to make replication mainstream. The final section contains concluding remarks.

Replication facilitates scientific progress but has never occupied a central role in the social and behavioral sciences. The goal of our target article was to change this situation. Science is not a collection of static empirical findings that have passed some threshold for statistical significance. Rather, it should rest on a set of procedures that reliably produce specific results to help advance theories. We presented direct replication as just one of many ways that can improve research in the social and behavioral sciences, along with, for instance, preregistration and greater transparency.

We are encouraged by the many thoughtful and constructive commentaries on our target article. Taken as a whole, we believe the commentaries affirm our position. Several commentaries amplify and enhance what we said in the target article. Other commentaries bring up new topics that researchers should consider as they move forward with their own replication attempts. Still others sound a more critical note about the value of direct replication, at least as currently practiced. In the first section of this response we discuss comments that pertain to specific sections of the target article or that raise novel points that we had not previously addressed. In the second section, we highlight and respond to additional issues raised by the commentators. In the final section we provide an integrative overview of the target article, the commentaries, and our response to them.

R1. Response to specific comments

In the target article we presented an overview of what direct replication studies are, how they relate to other forms of inquiry, and what terminology has been used to describe these various investigations. We also presented a series of six concerns that have been raised about the value of replication studies and their implementation, along with our response to each of these concerns. In this section we follow the structure of the target article to discuss how the commentaries further shape our thinking about these specific concerns.

R1.1. Concern I: Context is too variable

One response to replication attempts (especially attempts that fail to achieve the same result as the original) is to posit that some unspecified contextual factor, a *hidden moderator*, affected the results of the replication study such that the original result was not replicated. This claim is then used to argue that differences in results are difficult to interpret. Taken to the extreme, this line of reasoning can be used by critics to question the entire enterprise of direct replication. Indeed, the hidden moderator argument is sometimes used to suggest that for entire areas of research, contextual factors are so influential and so difficult to predict that replication studies should not be expected to arrive at the same results as the original study. In the target article, we explained that the extreme form of this argument is antithetical to mainstream beliefs about how scientific knowledge is supposed to accumulate and how it is to be applied. Essentially, it means that entire experimental lines of research could be rendered immune from independent verification. However, as the commentaries note, when considered in relation to specific replication attempts, some nuance is required when considering contextual factors.

One of the most consistent, and most important, themes emerging from the commentaries regarding this issue is that although concerns about context sensitivity are often presented as an issue for replicators to consider, they can also be addressed by researchers conducting original studies. For instance, **de Ruiter** rightly points out that a replication is a test of a scientific claim. If the scope of this claim has not been specified, then the scientific community should take that claim to mean that the original authors implicitly generalized across the unmentioned details. Debates about context sensitivity as an explanation of failed replication studies highlight that original researchers have an important responsibility for specifying the conditions that are essential for the predictions from their theory to hold. **Howe & Perfors** likewise propose that authors specify the extent to which they expect their findings to replicate. **Simons, Shoda, & Lindsay** (**Simons et al.**) take this idea even further and propose that specifying *constraints on generality* should be an essential component of original articles. Such statements can eliminate ambiguity in advance, thereby leading to a more cumulative science. We wholeheartedly endorse the proposal of specifying constraints on generality and return in more detail to this issue in our discussion of concern IV.

One virtue of statements specifying constraints on generality is that they provide authors with greater incentive (and explicit guidance) to think carefully about contextual influences in the initial stages of an original study rather than when interpreting failed replication attempts. Likewise, when replications are routine, there are more incentives to thoroughly document study procedures and identify the kinds of expertise needed to conduct studies. Authors may also wish to increase the rigor of their studies by adopting preregistration and within-lab direct replications. These practices will increase their own confidence in the evidentiary value of their original work. Collectively, these practices will help move research forward. They are among the reasons motivating our target article. This spirit of optimism is evident in the commentary by **Spellman & Kahneman** when they noted that replications efforts have been useful in improving research practices and evidentiary standards.

Some of the commentators appear to be unconvinced that it will ever be possible to specify conditions that allow for direct replications to occur. **Petty**, for instance, notes that the use of the same operations does not guarantee that a study counts as a replication, precisely because the same operations may mean different things in different contexts. This sentiment is echoed by **Wegener and Fabrigar**. We agree with these authors that (1) the appropriateness of specific procedures in new samples or new settings must certainly be evaluated when conducting replications and (2) evaluating these issues is necessary when interpreting results from replication studies. Fortunately, as noted by those commentators, the conceptual and statistical tools needed to conduct such analyses are available. The need to make sure that procedures of a direct replication study have validity does not invalidate the importance of direct replication.

https://doi.org/10.1017/S0140525X17001972

Furthermore, it is telling that the two examples of problematic reliance on original operations that **Petty** and **Wegener** & **Fabrigar** highlight are (1) not based on actual failed replication attempts (as far as we can tell) and (2) represent extreme and arguably implausible examples. For instance, Petty asks what would happen if stimuli from the 1950s were used today, and Wegener and Fabrigar ask readers to imagine a film clip from a 1980s sitcom being used to elicit humor almost 40 years later. Arguments about the importance of these contextual factors are more compelling when accompanied by evidence that a failed replication resulted from inattention to contextual factors that changed the meaning of the operations used in the study. These kinds of general concerns are occasionally raised when a replication study fails. However, these arguments are more convincing when researchers propose tests of those ideas in future work.

The extent to which context matters may indeed depend on many factors, including the domain in which the research is being conducted. **Gantman, Gomila, Martinez, Matias, Levy Paluck, Starck, Wu, & Yaffe** suggest that the problem of context sensitivity is especially thorny in field research. If true, then specifying constraints on generality is especially important in this type of research. Importantly, this concern with context highlights an additional benefit of a greater emphasis on replication.

Gelman proposes to abandon the notion of direct replication and move to a meta-analytic approach, given that direct replications, in his view, are impossible in psychology. This might be too extreme of a perspective. We already stated in the target article why it is important to retain the notion of direct replication, along the lines proposed by Nosek and Errington (2017). Moreover, specifying constraints on generality will make it easier to conduct direct replications. We do, however, agree with Gelman's observation that meta-analytic approaches are important for advancing research and theory and see our proposal as fully congruent with this idea.

R1.2. Concern II: Direct replications have limited theoretical value

Several commentators (**Alexander & Moors**; **Little & Smith**; **Carsel, Demos, & Motyl** [**Carsel et al.**]; **Witte & Zenker**) argue that the focus on replication ignores a more serious underlying problem, namely, the poor status of theorizing in the field. Carsel et al., for instance, suggest that stronger theories will make replications more feasible and more informative because such theories generate more testable hypotheses. They note that statistical hypotheses are never really true in the strictest sense and that for such hypotheses to be of any use we must ensure that they map as closely as possible to our substantive (qualitative) hypotheses; we are reminded of Box's (1979) famous quote: "All models are wrong, but some are useful." We agree with these commentators' general sentiment and reiterate our belief that a theory can be tested only when it is clearly and unambiguously specified (see also Etz et al., in press). The greater the specificity of a prediction, the more informative is the research. This idea applies to both original and replication studies.

Direct replication has an important role to play with respect to the development of stronger theories. It is key, as the epigraph to the target article indicates, to have a procedure that reliably produces an effect and not to just have a single experiment with $p<.05$. Direct replication is the way to determine whether the procedure is reliable. The next step is then to examine, in a theory-driven systematic way, the limits of that effect in increasingly more stringent tests. This, as **Gernsbacher** notes, has already been the common practice in some areas of psychology for years (see also Bonett 2012). The proposal to have researchers state constraints of generality of their findings (**Simons et al.**) is a step in this direction. Having to state constraints on generality are beneficial will force researchers to make their theories more explicit by making distinctions between factors that are thought to be essential for the effect to occur and those that are incidental.

An interesting and novel refinement of replication research that builds directly on the notion of constraints on generality is the meta-study (Baribault et al. 2018). Researchers distinguish between factors that are deemed essential for the effect, for instance, whether the meaning of a color word matches the color in which it is presented in a Stroop experiment and factors that might be moderators of the effect, for instance, the use of words that are not color words but are strongly associated with a color (e.g., blood and grass), the font in which the words are presented, the number of letters that are colored, or the geographical location of the lab in which the experiment is carried out. These latter factors are randomized in a series of micro-experiments, each of them being a potential moderator of the effect. In other words, a meta-study is an attempt to sample from the set of possible experiments on a given topic. A series of meta-studies allows for a stronger test of a theory than a single experiment (original or replication) in that it assesses the robustness of an effect across various subtly different incarnations of the experiment (see also the commentary by **Witte & Zenker**). Accordingly, a meta-study can provide empirical support for a statement of constraints on generality and allows for further theoretical specification. Moreover, when a moderator appears, being able to account for it enhances the explanatory power of the theory, thus resulting in a progressive research program (Lakatos 1970). We agree with **Alexander & Moors** and **Little & Smith** that new avenues for more ambitious testing of theories should be explored; meta-studies are one such approach.

R1.3. Concern III: Direct replications are not feasible in some domains

We addressed this concern in the target article and several commentators expanded on that theme. **Kuehberger & Schulte-Mecklenbeck** point out that the fact that direct replications are more feasible in some domains than in others creates a selection bias: Studies that are easy to reproduce are more likely to become the target of replication efforts than studies that are difficult or expensive to reproduce. Similarly, **Giner-Sorolla, Amodio, & van Kleef** (**Giner-Sorolla et al.**) point out that if the reasons for selecting specific studies are not made explicit, then studies that are most frequently targeted for replication attempts may be those for which there is the most doubt. Whether or not this is a problem, however, depends on the goals of the replication attempt.

It is important to distinguish between two distinct perspectives related to this concern. The first is the perspective

of the meta-scientist, who may want to estimate the replicability of a paradigm, domain, research area, or even an entire field. We agree that to accomplish such a goal, having a sound sampling plan is critical. If only the weakest studies or those that are easiest to reproduce are selected for replication, then surely estimates regarding the strength of an entire field or domain of study would not be accurate. We also agree that this sampling issue has not been given sufficient attention in the literature; the suggestions in these commentaries provide an important step in this direction.

The second perspective is that of the researcher in the field who is interested in the robustness of a specific research finding. For this researcher, there are myriad potential reasons why a specific study is selected. Also this is perfectly acceptable. Indeed, even explicit doubt about the veracity of the original finding can be a valid reason for conducting a replication study. Blanket suggestions about which studies can or should be chosen for replication limit the freedom of researchers to follow theoretical and empirical leads that they believe will be most interesting and fruitful. These suggestions place constraints on replicators that are not placed on original researchers. A single investigator may be interested in the replicability and robustness of a single minor finding; and just as the original investigator was free to produce that minor finding, someone wishing to replicate that result should be free to do so without others raising concerns about how the study was selected.

In short, we are unsympathetic to suggestions for the need to more tightly regulate replications versus other kinds of research. Movements in this direction are antithetical to making replication mainstream. Nevertheless, a number of tools are available to help replicators approach their task in a more rigorous fashion. Many of these tools are useful to scientists who simply want to evaluate the existing literature without conducting a direct replication. For instance, we endorse the suggestion by **Nuijten, Bakker, Maassen, & Wicherts** (**Nuijten et al.**), who suggest that those who wish to replicate a specific finding first check to make sure that the results of the original studies can actually be reproduced with the original data to ensure that these original results themselves can be verified. Our hope is that these issues will become less relevant when replication is more common and original studies are pushed by the field to have more evidentiary value.

We acknowledge (as we did in the target article) that there are some domains and some types of studies for which widespread replication will be difficult. In those domains, however, it will be especially important to incorporate additional safeguards that accomplish some of the same goals that direct replication is designed to accomplish. Many of the commentaries provided novel suggestions that may help in this respect. **MacCoun**, for instance, echoed the idea that direct replications are not always affordable or feasible and, for some phenomena, may even be impossible. In such situations, he argues, methods of blinded data analysis can help minimize p-hacking and confirmation bias, increasing our confidence in a study's results. We agree and note that, in fact, **Spellman & Kahneman** express the view that such strengthening of original studies is already happening.

R1.4. Concern IV: Replications are a distraction

Researchers who have raised reservations about direct replications often question their theoretical value and practical feasibility. A specific incarnation of this view is that direct replications are largely wasted efforts given the limited resources available to researchers in terms of time and energy. In response to this perception, several commentators provided suggestions for ways that replicators could increase the value of replication efforts. However, some of the more critical commentaries on our paper place what we see as puzzling demands on replicators.

Notably, **Strack & Stroebe** suggest that the onus of explaining why an effect was not replicated should be shouldered by the researchers performing the replication. In the target article, we called this an attempt to "irrationally privilege the chronological order of studies over the objective characteristics of those studies when evaluating claims about quality and scientific rigor" (sect. 5.1.1, para. 3). In their commentaries, **Gelman** and **Ioannidis** refer to this privileging of the original result as a "fallacy" and an "anomaly," respectively. The requirement for replicators to explain why they did not duplicate the effect poses several practical problems. Foremost, original findings can be flukes. In such cases, it is difficult to provide any sort of explanation for the failed attempt beyond noting that it is possible that a random process generated a $p<.05$ result in a single original study. Indeed, neither replicator nor original author can be certain when random processes are responsible for findings. Less extreme but no less thorny situations occur when replicators (and likely original authors themselves) are unaware of the myriad contextual factors that might have coalesced to produce an original effect. Thus, we are not in favor of placing so much onus on replicators relative to original authors. We believe that replicators (just like original authors) should simply provide their interpretations of results and findings in their own papers in the way they believe is faithful to the data and the literature. The research community can then decide whether particular interpretations are reasonable and empirically supported. Furthermore, pushing replicators to come up with strong statements explaining why they failed to replicate a result may increase concerns about the reputational consequences of replications. Sometimes the best response when reporting a failed replication is simply to get the finding into the literature, to provide a constraints on generality statement, and to issue necessary caveats about the need for additional research.

Indeed, we prefer to adopt a multipronged approach to evaluating replications, which would ideally culminate with multiple replications of specific findings that are combined in a meta-analysis. This seems to contrast with the scenario outlined by **Strack & Stroebe**, who describe what appears to be a case where there is one successful study and one failed direct replication. Without knowing more about the relative merits of the two studies in question, it is impossible to provide sound advice about how replicators should interpret the results and, thus, what they should or should not do in a discussion section. For instance, when the original study employed a between-subjects design with a sample size of 40 participants and the replication was a seemingly faithful recreation of the design but with a sample size of 400 participants, the weight of the evidence might lean in favor of the results of the replication. If both

https://doi.org/10.1017/S0140525X17001972

studies had modest sample sizes, the interpretation might need to be quite constrained. If, in contrast, the original had a sample size of 400 and the replication had a sample size of 40, there might be a strong need for the replication authors to compare effect size estimates and contemplate the power of their replication study before drawing strong conclusions! We think it is unwise to tie the hands of replicators by placing blanket requirements about how they interpret their results.

What we would consider fluke findings are sometimes part of the existing literature. Consider the results of several Registered Replication Reports, which show that a large effect size from an initial small-sample, between-subjects design can be reduced to near zero in large-scale, multi-lab replication attempts (e.g., Eerland et al. 2016; Hagger et al. 2016; Wagenmakers et al. 2016a). It is also possible to test whether there is heterogeneity in effect size estimates to try to find evidence in support of the existence of moderators. In cases where the effect size estimate is indistinguishable from zero and there are few indications of heterogeneity, the simplest explanation is that the original finding was a false positive or a grave overestimate (**Gelman**'s type M error). We furthermore like to underscore the relevance of Gelman's time-reversal heuristic here. Suppose the registered replication report had been conducted first and then the original study came second. What weight would we then assign to the original study? Very little, we surmise. Indeed, this is a subtext of the commentary by **Ioannidis**.

Strack & Stroebe are right to note that theories are formulated on a level that transcends the concrete evidence and that their validity does not rest on the outcome of one specific experimental paradigm. This mirrors our view (as we already outlined above); we hope nothing in our target article suggests otherwise. Direct replications provide a specific kind of information about the ability of a set of procedures to reliably produce the same results upon repetition. The process of evaluating the evidence for or against theoretical propositions involves a complex judgment involving multiple strands of evidence. Further, we also believe that null results are useful for the evaluation of a theory. That is, when explanations can be formulated for the absence of an effect and empirical support for them can be obtained, then the theory would actually be strengthened. This is akin to the process for evaluating the discriminant validity of measures in psychometric work. Theory specifies there should be no relation between two constructs and evidence is then gathered to test such a prediction.

Other commentators pushed the field to consider additional important elements besides direct replication. In many ways, we have no issues with these lines of thought. It was not our intent to say that direct replication is the one and only thing that will improve psychological science. **Heit & Rotello** seem to agree with us that direct replication is valuable, but argue that it should not be elevated more than, and thus shift attention away from, other worthwhile research practices, including conceptual replication and checking statistical assumptions. They also point out that replicating studies without checking the statistical assumptions can lead to increased confidence in incorrect conclusions, and that successful replications should not be elevated more than failed replications, given that both are informative. Indeed, our goal was simply to emphasize the benefits of, and call for increased attention to, what is an underused practice: direct replications. It is perhaps not surprising that we also agree that direct replications should be performed in a sensible manner. Indeed, replicating studies without checking statistical assumptions is unwise. Also we obviously agree that successful replications should not be elevated over failed replications (or vice versa).

Witte & Zenker raise the interesting issue of when replication is a necessary and worthwhile endeavor. They argue that replication efforts should be limited to specified theoretical effects. A successful replication would then show that an original study corroborated a theory. This is a reasonable point but we return to our earlier concern–we do not want to limit the freedom of replicators to study the effects of their own choosing. We also agree with their observation that several conceptually replicated experiments together may (ever more firmly) jointly corroborate, or falsify, a theoretical prediction.

R1.5. Concern V: Replications affect reputations

The fifth concern addressed in the target article focused not on the accumulation of scientific knowledge *per se*, but on the extra-scientific concerns about the people involved. Specifically, many have worried publicly about the reputational impact of replication studies, both for those whose works are targeted and for those who conduct the replications themselves. **Pennycook** agrees that scientists should separate their identities from the data they produce. More importantly, he points out that although people often fear that their work will fail to replicate, reputational consequences are often based more on whether the original authors approach such results with an open, scientific mindset than on whether the replication attempt affirmed or contradicted the original work. Pennycook rightly remarks that the same is true for the reputation of the replicators, and we agree that dispassionate, descriptive approaches to reporting replication results will make this research less fraught. We very much appreciate Pennycook's suggestions regarding ways that social and behavioral sciences can move toward this goal.

Along a similar line of thinking, **Tullett & Vazire** provide an idealistic new metaphor for scientific progress that challenges the idea that replications contribute to the literature only when they "tear down" bricks in an existing wall of scientific knowledge. They prefer an alternative metaphor where different participants in any scientific endeavor should be thought of as jointly solving a jigsaw puzzle. This metaphor captures the idea that the goal of scientific research is to uncover some underlying phenomenon and that both novel and replication studies provide critical information for achieving that goal. A welcome feature of the puzzle metaphor is that it puts replicators and original researchers on equal footing as two kinds of agents trying to solve a common problem. The brick analogy sets up a somewhat adversarial or regulatory dynamic in which original researchers are builders and replicators are those who further test the bricks for soundness. Reputational concerns are likely to be less relevant in the former conceptualization. It is also important to underscore that making replication mainstream means that *original researcher* and *replicator* are roles in the scientific enterprise that will normally be

https://doi.org/10.1017/S0140525X17001972

played by the same individual at different times and in different contexts. This captures what we mean by making replication mainstream.

Ultimately, the conceptual separation of data and researchers will be an important part of making replication mainstream. In fact, we can expect this process to be reciprocal. To the extent that replication becomes more mainstream, reputation will become more separated from data, which will make replication more mainstream, and so on. This cycle would seem to pay dividends for scientists and science as a whole.

One commentary brought up a reputational concern that we did not consider in the target article, but that is nonetheless interesting and important: replications might have reputational consequences for science itself. Scientific results must, of course, ultimately be conveyed to the public. This creates special challenges as popular coverage of science sometimes invites and perhaps even demands more certainty and clarity than is warranted by the existing evidence. **Białek** is concerned that false negatives may inspire lower confidence in science, which would undercut the effort to make replication mainstream. He does not advocate that scientists should stop replicating studies simply because they will look bad in the eyes of the public. Rather, he argues for a concerted effort to communicate the acceptability of uncertainty associated with scientific findings to the public (and to our peers too). We agree with this view but hasten to add that we suspect that a great deal of the responsibility lies with the original researchers. A quote from **de Ruiter**'s commentary nicely expresses this sentiment: "Finding general effects in psychology *is* very difficult, and it would be a good first step to address our replication crisis if we stopped pretending it is not" (last para.). There are many examples in which researchers broadcast their findings (often based on a single experiment with p barely<.05), trumping up their relevance to a variety of domains, only to resort to complaints about context being too variable after a failed replication. As we noted earlier, researchers should be realistic about the generalizability of their findings.

One additional virtue of making replication mainstream is that science will ideally produce more findings relevant to the kinds of claims covered in the popular media. Journalists may not have to wait too long to see if seemingly newsworthy findings are credible by virtue of having a track record of replicability. We see this increased knowledge base as an important practical consequence of the ideas we advocated in our target article.

A flipside to our argument is that research findings conveyed to the public but not backed by a solid evidentiary base risk generating grave reputational consequences. For instance, members of the public may become disillusioned when they implement the findings described in the popular media to change their behaviors and find that their efforts prove unsuccessful. This could prove catastrophic as these are the people who support science by funding governmental investments in research and the existence of many universities. Thus, we believe that researchers have strong incentives to make sure the public is provided with scientific claims that have a strong evidentiary base. As we have argued, direct replication is a component in making sure that the evidence base for claims is strong.

R1.6. Concern VI: There is no standard method to evaluate replication results

A concern about replications, noted in the target article, is that researchers are faced with myriad ways to statistically evaluate original and replication studies. Several commentators pick up on this theme and on related concerns about effect sizes and the kinds of errors that characterize research. For instance, **Gelman**, as well as **Tackett & McShane**, suggests that no real "null" effects are being studied in psychology, and thus, the concepts of false positives and false negatives are not very useful and should be abandoned. They suggest that we begin with an assumption that all effects we study are nonzero to some extent and recommend transitioning toward using multilevel models that allow us to characterize the variability of effects between studies in a rich fashion. Their preference to essentially abandon null hypothesis testing reflects a long-standing issue in the field as hypothesis testing has always been a contested issue among statisticians and methodologists. The fact remains, however, that many well-informed users of statistics still consider a hypothesis test (and ideas of false positives and false negatives) relevant to answering their scientific question in many cases. Thus, again we are reluctant to be too directive about the kinds of statistical tools researchers use. There might be cases where hypothesis testing is useful.

We agree with those commentators who push the value of thinking beyond type I and II errors to have researchers consider estimation errors of the sort Gelman has proposed. There are many ways to get things wrong in scientific research! Type M errors occur when researchers overestimate or underestimate the magnitude of an effect, and type S errors occur when researchers get the sign of the effect wrong. This kind of error could be particularly problematic when dealing with interventions as the sign may indicate iatrogenic effects. As we hope was clear in both the target article and this response, we see replication studies as providing additional information that can help reduce errors of all sorts. Thus, to the extent that we used terms like false positives in the original target article, critics who prefer the type M and S framework can think of a false positive as occurring when researchers are dealing with effects that are tiny in comparison to the original estimate or when there is type S error of any magnitude. If this mindset is adopted, we believe the vast majority of our arguments and perspectives hold.

Tackett & McShane chastise us for suggesting "three ways of statistically evaluating a replication, all of which are based on the null hypothesis significance testing (NHST) paradigm and the dichotomous p-value thresholds" that are "a form of statistical alchemy that falsely promise to transmute randomness into certainty" (para. 7). We worry this is a misreading of our target article. We assume these comments are in reference to our summary of three of the methods used to evaluate the results of the Reproducibility Project: Psychology (Open Science Collaboration 2015). However, we were merely describing the various methods that the authors of that study had used to evaluate replication success, and elsewhere in our article, we (briefly) discuss additional approaches that could be used. Tackett & McShane are right in noting limitations in some of the existing methods for evaluating replications and we are glad that they brought attention to those issues.

https://doi.org/10.1017/S0140525X17001972

We reduced our critical coverage of those issues in the target article in the interest of space. We are inclined toward the two approaches we discuss in detail (the small telescopes approach and the replication Bayes factor approach), but we agree that other approaches such as a meta-analytical (multilevel) approach can be useful, as well. We did not want to have our case for making replication mainstream bogged down by statistical arcana or the "framework wars" that sometimes derail debates over frequentist versus Bayesian methods.

Holcombe & Gershman reiterate that the result of an experiment depends on the status of not only the primary research hypotheses being investigated, but also the other auxiliary hypotheses (i.e., moderators). As such, their proposal is also relevant to concern I: Context is too variable. A failure to replicate a past finding can be a result of either the falsity of the primary research hypothesis or the failure to satisfy the conditions specified by auxiliary hypotheses. It would appear that a replication failure can provide evidence against only the conjunction of the research hypothesis and relevant auxiliary hypotheses–an instance of the Duhem-Quine problem. Without a way to distinguish between the primary and auxiliary hypotheses when interpreting a replication result, researchers are left wondering about the status of the theory. Holcombe & Gershman suggest a reformulation of Bayes' theorem derived by Strevens (2001, p. 525) can solve this problem. Call *H* the truth status of the primary hypothesis, *A* the truth status of an auxiliary hypothesis, and *HA* the conjunction of *H* and *A*. Strevens showed that the posterior belief in *H* given the falsification of *HA* can be determined entirely by (1) our prior belief in *H* and (2) our prior belief in *A* given *H*. Based on this result, Holcombe & Gershman recommend implementing "pilot programs to induce scientists to set out their beliefs before the data of a replication study are collected," allowing Strevens's result to be used and the belief in the theory updated.

It is fruitful to consider how our beliefs in our theories can be disentangled from our beliefs in auxiliary hypotheses in a quantitative way. However, we must admit to being surprised by how apparently simple the result summarized by **Holcombe & Gershman** is. We only need to specify prior probabilities, and need not consider such things as our model for the data-generating process? Consulting Strevens (2001), the precise result referred to by Holcombe & Gershman (clarified to us via personal communication) is

$$P(H|\neg(HA)) = P(H) \times \frac{1 - P(A|H)}{1 - P(A|H)P(H)}$$

Our (brief) impression based on the derivation by Strevens is that the above result can be used to evaluate replication success only when the following two conditions are met. First, it is possible for the data from the replication experiment to provide strict falsification of *HA*. If this condition is not met, the expression above becomes dependent on more than the two prior probabilities. Second, there exists either a single auxiliary hypothesis or a small number of independent auxiliary hypotheses that capture the differences between original and replication study. If this condition is not met we again see the above result become dependent on more than the relevant prior probabilities. We suspect that neither of these conditions is usually met in the context of psychological research. Hypotheses in psychology tend to make only probabilistic predictions, meaning strict falsification of the conjunction *HA* is usually not possible. Moreover, we doubt that differences between studies can be fully captured by a small number of independent auxiliary hypotheses.

R2. Additional suggestions for making replication mainstream

Although we anticipated that some of the commentaries might present dissenting views about the value of direct replication or the specific suggestions we made for resolving controversies about these issues, we were especially gratified to see that many commentators went beyond our suggestions to provide novel ways to make replication more mainstream. **Srivastava**, for example, points out that replication research promotes dissemination of information needed for other aspects of verification. Making replication more normative creates the expectation that others will need to know the details of original research, including previously opaque details about specific measures, procedures, or underlying data from original research. Thus, making replication mainstream promotes meta-scientific knowledge about what results to treat as credible even if a specific study never happens to be replicated. More broadly, endorsing replication as a method for ensuring the credibility of research reinforces the idea that scientists ought to be checking each other's work. **Lilienfeld** argues that direct replication is not only feasible but, in fact, also necessary for two domains of clinical psychological science: the evaluation of psychotherapy outcome and the construct validity of psychological measures. We agree, and we hope that replication attempts become more commonplace in these areas.

Howe & Perfors propose to make it standard practice for journals to pre-commit to publishing adequately powered, technically competent direct replications (at least in online form) for any article they publish and link to it from the original article. This is a practice that journals could readily implement. It is known colloquially as the *Pottery Barn Rule* following Srivastava (2012) and is close to the editorial policy adopted at the *Journal of Research in Personality* when Richard Lucas was the Editor-In-Chief (Lucas & Donnellan 2013). Our one caveat is that we are not convinced replications should only be relegated to online archives. **IJzerman, Grahe, & Brandt** (**IJzerman et al.**) and also **Gernsbacher** point to an initiative to make replication habitual by integrating replication with undergraduate education. Given that several of us are already using this practice, we support this initiative. Similar to the proposal by IJzerman et al., but targeted at a stage somewhat later in the educational process is **Kochari & Ostarek**'s proposal to introduce a replication-first rule for Ph.D. projects. We think this is an interesting proposal that specific graduate programs consider for themselves. We suspect including replication studies could actually improve the quality of the dissertations themselves, while also becoming a valuable element of graduate training. **Gorgolewski, Nichols, Kennedy, Poline, & Poldrack** suggest making replication mainstream by making it prestigious, for instance, by giving awards and note that such an approach has already been implemented by the Organization for Human Brain

https://doi.org/10.1017/S0140525X17001972

Mapping. Although we are not sure which of these proposals will be seen as most feasible and effective, we were impressed with the diverse suggestions that the commenters provided in this regard and look forward to seeing them implemented in some form.

Other commentaries focus on how the use of direct replication was synergistic with other proposed reforms in the ongoing discussions about methodological improvements in the field. **Little & Smith** favor the use of small-*N* designs. We think there is much to like about such an approach, but we also worry that such designs are not feasible in many areas of psychology. Thus, to the extent that such designs are appropriate for a given research question, they should be encouraged. **Paolacci & Chandler** advocate the use of open samples (e.g., those recruited from platforms such as Mechanical Turk), and they discuss how to use them in a methodologically sound way. These authors rightly note that although open samples provide greater opportunity for close replication (as more researchers have access to the same population), they also pose unique challenges for replication research. Original researchers and replicators alike would benefit from considering the issues that Paolacci & Chandler raise in this regard. **Tierney, Schweinsberg, & Uhlmann** (**Tierney et al.**) and also **Gernsbacher** suggest an approach to which we are particularly sympathetic (and, indeed, a description of which we had to cut out of an earlier version of the manuscript; Zwaan 2017) called *concurrent replication*. This practice involves the widespread replication of research findings in independent laboratories prior to publication or in a reciprocal fashion. As Tierney et al. argue, this addresses three key concerns discussed in the target article: it explicitly takes context into account, reduces reputational costs for original authors and replicators, and increases the theoretical value of failed replications. **Spellman & Kahneman** question whether replications will need to continue in the way they have recently been conducted. Specifically, they argue that large-scale, multi-lab replications that assess the robustness of one or two critical findings may die out as replication becomes more integrated into the research process. This may indeed be a consequence of making replication mainstream. Spellman & Kahneman present several proposals about how various research labs could collaborate to simultaneously investigate new phenomena and conduct direct replications of the results that are found. These are well worth considering, and journal editors may wish to commission special issues to incentivize tests of their proposals.

As many commentators note, improving research practices more broadly would also facilitate embedding replication in the mainstream of research. Improvements can be targeted at different stages of the research process. For instance, **Schimmack** notes that the current practice of basing publication decisions on whether the main results are significant or not is problematic, as this provides a built-in guarantee that most replication results will yield smaller effect sizes than original studies. In turn, this means that replication studies will inevitably be perceived as a challenge to the original effect. Publication bias also renders meta-analyses problematic, which is an issue that the field needs to address. Tools such as registered reports can help reduce publication bias. Many commentators furthermore note that reporting standards for original research should be improved because they are currently not sufficiently stringent (**Giner-Sorolla et al.** and **Simons et al.**). With reporting standards as they are, it is important to verify the original results before launching a replication effort (**Nuijten et al.**). With respect to the evaluation of replications, Giner-Sorolla et al. state that methodologically inconclusive replications ought not to be counted as non-replications. We concur. With respect to the dissemination of replications, **Egloff** suggests that debates surrounding the larger reform movement in psychology, in contrast to what often occurs right now, should choose mainstream outlets for their work (rather than, e.g., blogs and social media). We agree that this would help make replication more mainstream, and researchers should definitely be encouraged to send their contributions to this important debate to conventional outlets in addition to their blogs. If the debate surrounding the reform efforts features more prominently in mainstream outlets then replications themselves may eventually be seen as being worthy of publication there, as well. Unfortunately, some high-profile journals routinely reject replication studies for lack of novelty; and thus, an important component of this approach will be to lobby editors, publication boards, and societies to allow for such contributions to be published. One might worry that allowing replication studies into top journals that had previously been known for novel research findings will dilute the pages of those journals with a flood of studies that simply seek to verify those results. Future meta-scientific research can track the number of replication studies that are actually submitted and published at journals that allow them, while also tracking the effects on the credibility and replicability of the results that are published at those journals.

R3. Conclusion: There should be no special rules for replication studies

The title of our target article, "Making replication mainstream," was chosen to reflect our beliefs about the role that direct replications should play in science. Although we advocate direct replications, we acknowledge that replications are (1) only one tool among many to improve science, (2) not necessarily the most important reform that scientists who are concerned about methodological reform should adopt, or (3) even a practice that all scientists must necessarily prioritize in their own research efforts. Instead, we argued that direct replication should become a more normal, more *mainstream* part of the scientific process. An important part of our goal of making replication mainstream is to ensure that replication studies are not held to different standards than other forms of research.

A number of the commentaries could be interpreted as proposing special rules for conducting and evaluating the results of replication studies. In our target paper we discussed a broad range of statistical tests that could be used to evaluate replication results, but some commentators suggest that replication research needs to go even further in the number of tests conducted and the sophistication of those tests. Although we are of course in favor of reconsidering the appropriateness of any default analytic approaches, we believe that this goal in no way applies

https://doi.org/10.1017/S0140525X17001972

specifically to replication studies. In fact, one of the things replication efforts have shown is that analytic approaches and reporting standards may need to be improved across the board. Several commentaries seem to be aligned with this view (e.g., **Giner-Sorolla et al.**; **Nuijten et al.**; **Schimmack**, **Spellman & Kahneman**; and **Simons et al.**)

Commentators such as **Wegener & Fabrigar** suggested that replication research should adhere to the standards that hold for original research, standards that include rigorous pretesting and the inclusions of supplemental tests to ensure that the manipulations and measures worked as intended. They suggest that replications rarely meet these standards. However, they provide no evidence that replication studies routinely fail to meet these standards or, more importantly, that original research itself is actually held to the high standard that they describe for replications. Indeed, our experience is that the practice of conducting replication attempts frequently reveals methodological problems in the original studies that would have gone unnoticed without the attention to detail that conducting direct replications requires (e.g., Donnellan et al. 2015). It would be informative to conduct a meta-scientific study in which original and replication studies were scored for dimensions of rigor of the sort identified by Wegener and Fabriagar.

As consumers of research, we, of course, start with our subjective impressions regarding the typical quality of original and replication research; and our personal impression (which is also not informed by systematic data) is that replication studies are already more likely to include these features than the typical original study. After all, these studies have the advantage of relying on existing protocols, and thus, replication researchers have far less flexibility when it comes to data analysis as the original study provides numerous constraints. Many replications have far larger samples sizes and generate effect size estimates that are seemingly more plausible than original studies. When original studies find significant results, the fact that the study did not include critical methodological features that would have been useful to explain a nonsignificant result is noticed by few. Concerns about the magnitude of the effect size estimate are more often raised when the estimate is tiny (as is often the case with null results from replication studies) than when it is large (as is often the case with primary studies based on modest sample sizes). Considerable attention is devoted to explaining away small effect size estimates by appeal to hidden moderators, whereas less attention is paid to explaining why a large effect is plausible given the outcome variable in question or strength of the experimental manipulation. Regardless of who is right about the prevalence of the features of high-quality research, we agree that they are desirable and as a field we should push for their inclusion in all research, replication or otherwise. We worry, however, that their absence is often used, in an ad hoc fashion, as a way to dismiss failed replications of original studies that used the exact same methodology. In extreme cases, critics may even ignore the strengths of the replication studies when attempting to privilege the original result.

Ioannidis, identifying some of the features we have identified, pushes this argument further than even we do and suggests that replications often have more utility than original studies because biases are more common in original research. Many commentators provided special rules for how to identify studies that should be replicated. For example, **Witte & Zenker** stress that it is important to evaluate the potential for generating theoretical knowledge before launching a replication project. Several commentaries address this concern by providing concrete solutions. **Hardwicke, Tessler, Peloquin, & Frank** and **Coles, Tiokhin, Scheel, Isager, & Lakens** both propose translating the question of which replication to run into a formal decision-making process, whereby replications would be deemed worthy to run or not based on the utility we expect to gain from them. Their suggestions essentially amount to considering the costs and benefits of running a particular replication and evaluating the subjective probability we assign the underlying theory, hearkening back to the quintessentially Bayesian ideas previously put forward by the likes of Wald, Lindley, Savage, and others. Their suggestions can aid individual researchers and groups when they go about deciding how to allocate their own time and effort.

These are all interesting and potentially useful perspectives to take moving forward. At the same time, we do not think that special rules for selecting replication studies are needed, or even desirable. Certainly original research studies vary in the contribution they make to science, yet few propose formal mechanisms for deciding which new original studies should be conducted. Much original research builds on, or is even critical of, prior theory and research (as should be the case in a cumulative science). Idiosyncratic interests and methodological expertise guide the original research questions that people pursue. This should be true for replication research, as well. People conduct replications for many reasons: because they want to master the methods in an original study, because they want to build on the original finding, and yes, even because they doubt the validity of the original work. But this is true regardless of whether the follow-up study that a person conducts is a replication or an entirely new study building on prior work.

As we noted in the target article, although the ability to replicate a research finding is a foundational principle of the scientific method, the role of direct replication in the social and behavioral sciences is surprisingly controversial. The goal of our article was to identify and address the major reasons why this controversy exists and to suggest that science would benefit from making replication more mainstream. The commentaries on this article strengthen our belief that an increased focus on replication will benefit science; at the same time these commentaries pushed us to think more about the reasons why controversy about replication exists. We hope that the resulting debate will encourage all scientists to think carefully about the role that direct replication should play in building a cumulative body of knowledge. Once again, we thank these commentators for their insightful comments and we look forward to seeing these ideas evolve as social and behavioral sciences engage with a broad range of meta-scientific issues in the years to come.

ACKNOWLEDGMENT

Alexander Etz was supported by Grant 1534472 from National Science Foundation's Methods, Measurements, and Statistics panel and by the National Science Foundation Graduate Research Fellowship Program (No. DGE1321846).

NOTE

1. Current address: Department of Psychology, Psychology Building, 316 Physics Road, Room 249A, Michigan State University, East Lansing, MI 48824.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively]

[Author name initials "AG" are same for both [Ana Gantman] and [Andrew Gelman] and the initials "GP" are the same for both [Gabriele Paolacci] and [Gordon Pennycook]. Please confirm the change of author name initials [Ana Gantman as "AGa"], [Andrew Gelman as "AGe"], [Gabriele Paolacci as "GPa"] and [Gordon Pennycook as "GPe"] in reference list just to differentiate the author names.

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P. & Cubelli, R. (2017) Questionable research practices among Italian research psychologists. *PLoS One* 12(3):e0172792. Available at: http://doi.org/10.1371/journal.pone.0172792. [MBN]

Aklin, M. & Urpelainen, J. (2014) Perceptions of scientific dissent undermine public support for environmental policy. *Environmental Science and Policy* 38:173–77. Available at: http://doi.org/10.1016/j.envsci.2013.10.006. [MB]

Alexander, D. M., Jurica, P., Trengove, C., Nikolaev, A. R., Gepshtein, S., Zvyagintsev, M., Mathiak, K., Schulze-Bonhage, A., Reuscher, J., Ball, T. & van Leeuwen, C. (2013) Traveling waves and trial averaging: The nature of single-trial and averaged brain responses in large-scale cortical signals. *NeuroImage* 73:95–112. Available at: https://doi.org/10.1016/j.neuroimage.2013.01.016. [DMA]

Alexander, D. M., Trengove, C. & van Leeuwen, C. (2015) Donders is dead: Cortical traveling waves and the limits of mental chronometry in cognitive neuroscience. *Cognitive Processing* 16(4):365–75. Available at: https://doi.org/10.1007/s10339-015-0662-4. [DMA]

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C., Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., Di Domenico, A., Drummond, A., Echterhoff, G., Edlund, J. E., Eggleston, C. M., Fairfield, B., Franco, G., Gabbert, F., Gamblin, B. W., Garry, M., Gentry, R., Gilbert, E. A., Greenberg, D. L., Halberstadt, J., Hall, L., Hancock, P. J. B., Hirsch, D., Holt, G., Jackson, J. C., Jong, J., Kehn, A., Koch, C., Kopietz, R., Körner, U., Kunar, M. A., Lai, C. K., Langton, S. R. H., Leite, F. P., Mammarella, N., Marsh, J. E., McConnaughy, K. A., McCoy, S., McIntyre, A. H., Meissner, C. A., Michael, R. B., Mitchell, A. A., Mugayar-Baldocchi, M., Musselman, R., Ng, C., Nichols, A. L., Nunez, N. L., Palmer, M. A., Pappagianopoulos, J. E., Petro, M. S., Poirier, C. R., Portch, E., Rainsford, M., Rancourt, A., Romig, C., Rubínová, E., Sanson, M., Satchell, L., Sauer, J. D., Schweitzer, K., Shaheed, J., Skelton, F., Sullivan, G. A., Susa, K. J., Swanner, J. K., Thompson, W. B., Todaro, R., Ulatowska, J., Valentine, T., Verkoijen, P. P. J. L., Vranka, M., Wade, K. A., Was, C. A., Weatherford, D., Wiseman, K., Zaksaite, T., Zuj, D. V. & Zwaan, R. A. (2014) Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science* 9:556–78. [aRAZ]

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., van der Hulst, M., Jonas, K. J., Lai, C. K., Levitan, C. A., Miller, J. K., Moore, K. S., Meixner, J. M., Munafò, M. R., Neijenhuijs, K. I., Nilsonne, G., Nosek, B. A., Plessow, F., Prenoveau, J. M., Ricker, A. A., Schmidt, K., Spies, J. R., Steiger, S., Strohminger, N., Sullivan, G. B., van Aert, R. C. M., van Assen, M. A. L. M., Vanpaemel, W., Vianello, M., Voracek, M. & Zuni, K. (2016) Response to comment on "estimating the reproducibility of psychological science." *Science* 351(6277):1037. [aRAZ, TEH]

Anderson, S. F., Kelly, K. & Maxwell, S. E. (2017) Sample-size planning for more accurate statistical power: A method of adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science* 28(11):1547–62. [PDLH]

Arechar, A. A., Kraft-Todd, G. T. & Rand, D. G. (2017) Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association* 3(1):1–11. [GPa]

Baker, M. (2015, August 27) Over half of psychology studies fail reproducibility test. *Nature News*. Available at: https://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248. [FS]

Baker, M. (2016) Is there a reproducibility crisis? *Nature* 533:452–54. [aRAZ]

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012) The rules of the game called psychological science. *Perspectives on Psychological Science* 7(6):543–54. Available at: http://doi.org/10.1177/1745691612459060. [EHW, MBN]

Bakker, M. & Wicherts, J. M. (2011) The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods* 43(3):666–78. Available at: http://doi.org/10.3758/s13428-011-0089-5. [MBN]

Bakker, M. & Wicherts, J. M. (2014) Outlier removal and the relation with reporting errors and quality of research. *PLoS One* 9(7):e103360. Available at: http://doi.org/10.1371/journal.pone.0103360. [MBN]

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P. & Vandekerckhove, J. (2018) Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America* 115(11):2607–12. Available at: http:/doi.org/10.1073/pnas.1708285114. [rRAZ, TEH]

Barrera, M., Jr. & Rosen, G. M. (1977) Detrimental effects of a self-reward contracting program on subjects' involvement in self-administered desensitization. *Journal of Consulting and Clinical Psychology* 45:1180–81. [SOL]

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

Baumeister, R. F. (2016) Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology* 66:153–58. Available at: http://doi.org/10.1016/j.jesp.2016.02.003. [aRAZ]

Baumeister, R. F., Bratslavsky, E., Muraven, M. & Tice, D. M. (1998) Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74(5):1252–65. [AK]

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

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

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C. Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Ho, T.-H., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafò, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Shönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J. & Johnson, V. E. (2017) Redefine statistical significance. *Nature Human Behaviour* 2:6–10. Available at: http://doi.org/10.1038/s41562-017-0189-z. [JPAI]

Berk, R. A., Campbell, A., Klap, R. & Western, B. (1992) The deterrent effect of arrest in incidents of domestic violence: A Bayesian analysis of four field experiments. *American Sociological Review* 57:698–708. [RJM]

Berman, J. S. & Reich, C. M. (2010) Investigator allegiance and the evaluation of psychotherapy outcome research. *European Journal of Psychotherapy and Counselling* 12:11–21. [WT]

Blakeley, B., McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2018) Abandon statistical significance. Preprint. Available at: https://arxiv.org/abs/1709.07588. [FS]

Blakey, S. M. & Abramowitz, J. S. (2016) The effects of safety behaviors during exposure therapy for anxiety: Critical analysis from an inhibitory learning perspective. *Clinical Psychology Review* 49:1–15. [SOL]

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S. & Forstmann, B. U. (2015) A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* 66:115–33. Available at: https://doi.org/10.1016/j.cortex.2014.11.019. [KG]

Bohannon, J. (2014) Replication effort provokes praise – and 'bullying' charges. *Science* 344:788–89. [GPe, WT]

Bonett, D. G. (2012) Replication-extension studies. *Current Directions in Psychological Science* 21:409–12. [rRAZ]

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009) *Introduction to meta-analysis*. Wiley. [TEH]

Bowen, A. & Casadevall, A. (2015) Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proceedings of the National Academy of Sciences of the United States of America* 112:11335–40. [JPAI]

Box, G. E. P. (1979) Robustness in the strategy of scientific model building. In: *Robustness in statistics*, ed. R. L. Launer & G. N. Wilkinson, pp. 201–36. Academic. [rRAZ]

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* 50:217–24. [aRAZ, HI]

Brown, N. J. L. & Heathers, J. A. J. (2017) The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science* 8(4):363–69. Available at: http://doi.org/10.1177/1948550616673876. [MBN]

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):365–76. [aRAZ]

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmedj, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. & Wu, H. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–36. [aRAZ]

Campbell, D. T. (1969) Reforms as experiments. *American Psychologist* 24(4):409. [AGa]

Campbell, D. T. (1991) Methods for the experimenting society. *Evaluation Practice* 12(3):223–60. [AGa]

Carlin, J. B. (2016) Is reform possible without a paradigm shift? *The American Statistician,* Supplemental material to the ASA statement on *p*-values and statistical significance 10. [JLT]

Casey, L., Chandler, J., Levine, A. S., Proctor, A. & Strolovitch, D. Z. (2017, April–June) Intertemporal differences among MTurk worker demographics. *SAGE Open*, 1–15. doi: 10.1177/2158244017712774. [GPa]

Cesario, J. (2014) Priming, replication, and the hardest science. *Perspectives on Psychological Science* 9:40–48. [aRAZ]

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

Chambers, C. D. (2013) Registered reports: A new publishing initiative at *Cortex*. *Cortex* 49(3):609–10. Available at: http://doi.org/10.1016/j.cortex.2012.12.016. [ARK]

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

Chambless, D. L. & Ollendick, T. H. (2001) Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology* 52:685–716. [SOL]

Chandler, J., Mueller, P. & Paolacci, G. (2014) Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods* 46(1):112–30. [GPa]

Chandler, J., Paolacci, G., Peer, E., Mueller, P. & Ratliff, K. A. (2015) Using nonnaive participants can reduce effect sizes. *Psychological Science* 26(7):1131–39. [GPa]

Chandler, J. & Shapiro, D. (2016) Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology* 12:53–81. [GPa]

Chartier, C. R. (2017) The psychological science accelerator: A distributed laboratory network. Blog post. Available at: https://christopherchartier.com/2017/09/21/the-psychological-science-accelerator-a-distributed-laboratory-network. [WT]

Chavalarias, D., Wallach, J., Li, A. & Ioannidis, J. P. A. (2016) Evolution of reporting of p-values in the biomedical literature, 1990–2015. *Journal of the American Medical Association* 315(11):1141–48. [JPAI]

Chawla, D. S. (2017, April 7) Online platform aims to facilitate replication studies. *The Scientist*. Available at: https://www.the-scientist.com/?articles.view/articleNo/49161/title/Online-Platform-Aims-to-Facilitate-Replication-Studies/. [MAG]

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., Bowen, J. D., Bredbow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fenrnández-Rouco, N., Fuglestad, P. T., Goldberg, R. M., Golom, F. D., Gündoğdu-Aktürk, E., Hoplock, L. B., Houdek, P., Kane, H. S., Kim, J. S., Kraus, S., Leone, C. T., Li, N. P., Logan, J. M., Millman, R. D., Morry, M. M., Pink, J. C., Ritchey, T., Root Luna, L. M., Sinclair, H. C., Stinson, D. A., Sucharyna, T. A., Tidwell, N. D., Uysal, A., Vranka, M., Winczewski, L. A. & Yong, J. C. (2016) Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro & Hannon (2002). *Perspectives on Psychological Science* 11(5):750–64. [aRAZ]

Cohen, J. (1990) Things I have learned (so far). *American Psychologist* 45(12):1304. [aRAZ]

Cohen, J. (1994) The earth is round ($p < .05$). *American Psychologist* 49:997–1003. [TC]

Collaborative Replications and Education Project (2018) Current study list and selection methods. Available at: https://osf.io/flaue/wiki/home/. [RG-S]

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., Mcllor, W. & Oxman, A. D. (1993) Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA* 269(21):2749–53. [aRAZ]

Cook, T. D., Campbell, D. T. & Peracchio, L. (1990) Quasi-experimentation. In: *Handbook of industrial and organizational psychology, vol. 1, 2nd edition*, ed. M. D. Dunnette & L. M. Hough, pp. 491–576. Consulting Psychologists. [SOL]

Cortina, J. M., Aguinis, H. & DeShon, R. P. (2017) Twilight of dawn or of evening? A century of research methods in the Journal of Applied Psychology. *Journal of Applied Psychology* 102(3):274–90. [RJM]

Coyne, J. C. (2016) Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology* 4:28. Available at: http://doi.org/10.1186/s40359-016-0134-3. [aRAZ, SOL]

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. Available at: http://doi.org/10.1016/j.jesp.2015.10.002. [aRAZ, AMT, US]

Cronbach, L. J. & Meehl, P. E. (1955) Construct validity in psychological tests. *Psychological Bulletin* 52:281–302. [SOL]

Deutsche Forschungsgemeinschaft (2017) *Die Replizierbarkeit von Forschungsergebnissen (The replicability of research findings)*. Available at: www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_en.pdf. [FS]

DeVoe, S. E. & House, J. (2016). Replications with MTurkers who are naïve versus experienced with academic studies: A comment on Connors, Khamitov, Moroz, Campbell, and Henderson (2015). *Journal of Experimental Social Psychology* 67:65–67. [GPa]

Difallah, D., Filatova, E. & Ipeirotis, P. (2018) Demographics and dynamics of mechanical Turk workers. In: *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining, Marina Del Rey, CA, USA February 5–9, 2018*, pp. 135–143. Available at: https://dl.acm.org/citation.cfm?doid=3159652.3159661. [GPa]

Donnellan, M. B., Lucas, R. E. & Cesario, J. (2015) On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012): Study 1. *Emotion* 15(1):109–19. [rRAZ]

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek B. A. & Johannesson, M. (2015) Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences of the United States of America* 112:15343–47. [WT]

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim L. & Tetlock, P. E. (2015) Political diversity will improve social psychological science. *Behavioral and Brain Sciences* 38:e130. Available at: http://doi.org/10.1017/S0140525X14000430. [BE]

Dubé, C., Rotello, C. M. & Heit, E. (2010) Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review* 117:831–63. [EH]

Dudo, A., Dunwoody, S. & Scheufele, D. A. (2011) The emergence of nano news: Tracking thematic trends and changes in US newspaper coverage of nanotechnology. *Journalism and Mass Communication Quarterly* 88:55–75. Available at: http://doi.org/10.1177/107769901108800104. [MB]

Duhem, P. (1954) *The aim and structure of physical theory.* Princeton University Press. [AOH]

Dunlap, K. (1926) The experimental methods of psychology. In: *Psychologies of 1925*, ed. C. Murchison, pp. 331–53. Clark University Press. [aRAZ, WT]

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. & McIntosh, C., eds. (in press) *Information and accountability: A new method for cumulative learning*. Cambridge University Press. [AGa]

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., Chung, L. Y., Hermann, A. D., Joy-Gaba, J. A., Line, M. J., Rule, N. O., Sacco, D. F., Vaughn, L. A. & Nosek, B. A. (2017) Observe, hypothesize, test, repeat: Luttrell, Petty & Xu (2017) demonstrate good science. *Journal of Experimental Social Psychology* 69:184–86. [DTW]

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bofiglio, D. B. V., Boucher, L., Brown, E. R., Budima, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirschenbaum, J., Klein, R. A., Lucas, R. E., Lustgraff, C. J. N., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siesman, C., Sobocko, K., Sternglanz, R. W., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J. & Nosek, B. A. (2016a) Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* 67:68–82. Available at: http://doi.org/10.1016/j.jesp.2015.10.012. [DTW, RG-S, SS]

Ebersole, C. R., Axt, J. R. & Nosek, B. A. (2016b) Scientists' reputations are based on getting it right, not being right. *PLoS Biology* 14(5):e1002460. Available at: https://doi.org/10.1371/journal.pbio.1002460. [aRAZ, GPe]

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J. & Ioannidis, J. P. (2014) Reanalysis of randomized clinical trial data. *Journal of the American Medical Association* 312(10):1024–32. Available at: http://doi.org/10.1001/jama.2014.9646. [MBN]

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W.,

Poirier, C. & Prenoveau, J. M. (2016) Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science* 11:158–71. [arRAZ]

Epskamp, S. & Nuijten, M. B. (2016) statcheck: Extract statistics from articles and recompute p-values. Available at: https://cran.r-project.org/web/packages/statcheck/ (R package version 1.2.2). [MBN]

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. [aRAZ]

Estes, W. K. (1956) The problem of inference from curves based on group data. *Psychological Bulletin* 53(2):134–40. [DMA]

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

Etz, A., Haaf, J. M., Rouder, J. N. & Vandekerckhove, J. (in press) Bayesian inference and testing any hypothesis you can specify. Preprint. Available at: https://psyarxiv.com/wmf3r/. [rRAZ]

Etz, A. & Vandekerckhove, J. (2016) A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE* 11(2):e0149794. Available at: http://doi.org/10.1371/journal.pone.0149794. [aRAZ]

Etz, A. & Wagenmakers, E. J. (2017) JBS Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science* 32(2):313–29. [aRAZ]

Evans, J. St. B. T., Barston, J. L. & Pollard, P. (1983) On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition* 11:295–306. [EH]

Everett, J. A. & Earp B. D. (2015) A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology* 6:1–4. [WT]

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

Fanelli, D. (2011) Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3):891–904. Available at: http://doi.org/10.1007/s11192-011-0494-7. [ARK]

Fanelli, D., Costas, R. & Ioannidis, J. P. A. (2017) A meta-assessment of bias in science. *Proceedings of the National Academy of Sciences of the United States of America* 114:3714–19. [JPAI]

Fayant, M. P., Sigall, H., Lemonnier, A., Retsin, E. & Alexopoulos, T. (2017) On the limitations of manipulation checks: An obstacle toward cumulative science. *International Review of Social Psychology* 30(1):125–30. Available at: https://doi.org/10.5334/irsp.102. [EHW]

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(1):120–28. Available at: http://doi.org/10.1037/a0024445. [aRAZ, ARK]

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):555–61. [aRAZ]

Fetterman, A. K. & Sassenberg, K. (2015) The reputational consequences of failed replications and wrongness admission among scientists. *PLoS One* 10(12):e0143723. Available at: https://doi.org/10.1371/journal.pone.0143723. [aRAZ, GPe]

Fiedler, K. & Schwarz N. (2015) Questionable research practices revisited. *Social Psychological and Personality Science* 7:45–52. Available at: http://doi.org/10.1177/1948550615612150. [aRAZ]

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* 108:275–97. Available at: http://doi.org/10.1037/pspi0000007. [aRAZ]

Fiske, S. T. (1992) Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology* 63 (6):877. [AGa]

Fletcher, P. C. & Grafton, S. T. (2013) Repeat after me: Replication in clinical neuroimaging is critical. *NeuroImage: Clinical* 2:247–48. Available at: https://doi.org/10.1016/j.nicl.2013.01.007. [KG]

Foss, D. J. & Blank, M. A. (1980) Identifying the speech codes. *Cognitive Psychology* 12(1):1–31. Available at: https://doi.org/10.1016/0010-0285(80)90002-X. [MAG]

Foss, D. J. & Gernsbacher, M. A. (1983) Cracking the dual code: Toward a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior* 22:609–32. Available at: https://doi.org/10.1016/S0022-5371(83)90365-1. [MAG]

Franco, A., Malhotra, N. & Simonovits, G. (2014) Publication bias in the social sciences: Unlocking the file drawer. *Science* 345(6203):1502–505. [aRAZ]

Frankfurt, H. (2005) *On bullshit*. Princeton University Press. Available at: http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=5452992. [GPe]

Freudenburg, W. R., Gramling, R. & Davidson, D. J. (2008) Scientific certainty argumentation methods (SCAMs): Science and the politics of doubt. *Sociological Inquiry* 78:2–38. Available at: http://doi.org/10.1111/j.1475-682X.2008.00219.x. [MB]

Garnham, A., Traxler, M., Oakhill, J. & Gernsbacher, M. A. (1996) The locus of implicit causality effects in comprehension. *Journal of Memory and Language* 35:517–43. Available at: https://doi.org/10.1006/jmla.1996.0028. [MAG]

Gelman, A. (2013) I'm negative on the expression "false positives." Blog post. Available at: http://andrewgelman.com/2013/11/07/nix-expression-false-positives/. [ACe]

Gelman, A. (2015) The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management* 41 (2):632–43. [JLT]

Gelman, A. (2016a) The problems with p-values are not just with p-values. *The American Statistician*, Supplemental material to the ASA statement on p-values and statistical significance 10. [JLT]

Gelman, A. (2016b) The time-reversal heuristic – A new way to think about a published finding that is followed up by a large, preregistered replication (in context of Amy Cuddy's claims about power pose). Blog post. Available at: http://andrewgelman.com/2016/01/26/more-power-posing/. [AGe, AMT]

Gelman, A. & Carlin, J. B. (2014) Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9:641–51. [AGe]

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. [aRAZ]

Gernsbacher, M. A. (1989) Mechanisms that improve referential access. *Cognition* 32:99–156. Available at: https://doi.org/10.1016/0010-0277(89)90001-2. [MAG]

Gernsbacher, M. A. & Hargreaves, D. J. (1988) Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language* 27:699–717. [MAG]

Gernsbacher, M. A., Keysar, B., Robertson, R. R. W. & Werner, N. K. (2001a) The role of suppression and enhancement in understanding metaphors. *Journal of Memory and Language* 45:433–50. Available at: https://doi.org/10.1006/jmla.2000.2782. [MAG]

Gernsbacher, M. A., Robertson, R. R. W. & Werner, N. K. (2001b) The costs and benefits of meaning. In: *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity*, ed. D. S. Gorfein, pp. 119–37. APA. Available at: https://doi.org/10.1037/10459-007. [MAG]

Ghelfi, E., Christopherson, C. D., Fischer, M. A., Legate, N., Lenne, R., Urry, H., Wagemans, F. M. A., Wiggins, B., Barrett, T., Glass, M., Guberman, J., Hunt, J., Issa, N., Paulk, A., Peck, T., Perkinson, J., Sheelar, K., Theado, R. & Turpin, R. (in preparation) The influence of gustatory disgust on moral judgement: A pre-registered multi-lab replication. [HI]

Gilbert, D. T., King, G., Pettigrew, S, & Wilson, T. D. (2016) Comment on "estimating the reproducibility of psychological science." *Science* 351(6277):1037. Available at: http://doi.org/10.1126/science.aad7243. [aRAZ, MB, RG-S, TEH]

Giner-Sorolla, R. (2016) Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology* 65:1–6. [RG-S]

Giofrè, D., Cumming, G., Fresc, L., Boedker, I. & Tressoldi, P. (2017) The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLoS One* 12(4):e0175583. Available at: http://doi.org/10.1371/journal.pone.0175583. [MBN]

Glick, A. R. (2015) The role of serotonin in impulsive aggression, suicide, and homicide in adolescents and adults: A literature review. *International Journal of Adolescent Medicine and Health* 27:143–50. [SOL]

Goldin-Meadow, S. (2016) Preregistration, replication, and nonexperimental studies. *Association for Psychological Science Observer* 29(8):2. [NAC]

Goodman, J. K. & Paolacci, G. (2017) Crowdsourcing consumer research. *Journal of Consumer Research* 44(1):196–210. [GPa]

Gorgolewski, K., Nichols, T., Kennedy, D. N., Poline, J.-B. & Poldrack, R. A. (2017a) Promoting replications through positive incentives. *Figshare*. Available at: https://doi.org/10.6084/m9.figshare.5278327.v1. [KG]

Gorgolewski, K., Nichols, T., Kennedy, D. N., Poline, J.-B. & Poldrack, R. A. (2017b) Replication award creation kit. *Figshare*. Available at: https://doi.org/10.6084/m9.figshare.5567083.v1. [KG]

Grahe, J., Brandt, M., IJzerman, H. & Cohoon, J. (2014, February 28) Replication education. *Association for Psychological Science: Observer*. Available at: https://www.psychologicalscience.org/observer/replication-education. [MAG]

Grahe, J. E., Brandt, M. J., IJzerman, H., Cohoon, J., Peng, C., Detweiler-Bedell, B. & Weisberg, Y. (2015) Collaborative Replications and Education Project (CREP). Available at: http://osf.io/wfc6u. [HI]

Gray, K. (2017) How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science* 12(5):731–41. [TC]

Greenberg, J., Solomon, S., Pyszczynski, T. & Steinberg, L. (1988) A reaction to Greenwald, Pratkanis, Leippe, and Baumgardner (1986): Under what conditions does research obstruct theory progress? *Psychological Review* 95:566–71. [HI]

Greenwald, A. G. (1975) Significance, nonsignificance, and interpretation of an ESP experiment. *Journal of Experimental Social Psychology* 11:180–91. [aRAZ]

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R. & Baumgardner, M. H. (1986) Under what conditions does theory obstruct research progress? *Psychological Review* 93:216. [HI]

Grice, J., Barrett, P., Cota, L., Felix, C., Taylor, Z., Garner, S., Medellin, E. & Vest, A. (2017) Four bad habits of modern psychologists. *Behavioral Sciences* 7:53. [DRL]

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., Elson, M., Evans, J. R., Fay, B. A., Fennis, B. M, Finley, A., Francis, Z., Heise, E., Hoemann, H., Inzlicht, M., Koole, S. L., Koppel, L., Kroese, F., Lange, F., Lau, K., Lynch, B. P., Martijn, C., Merckelbach, H., Mills, N. V., Michirev, A., Miyake, A., Mosser, A. E., Muise, M., Muller, D., Muzi, M., Nalis, D., Nurwanti, R., Otgaar, H, Philipp, M. C., Primoceri, P., Rentzsch, K., Ringos, L., Schlinkert, C., Schmeichel, B. J., Schoch, S. F., Schrama, M., Schütz, A., Stamos, A., Tinghög, G., Ullrich, J., vanDellen, M., Wimbarti, S., Wolff, W., Yusainy, C., Zerhouni, O. & Zwienenberg, M. (2016) A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science* 11(4):546–73. [arRAZ, AK, EHW]

Hawthorne, J. (2014) Bayesian confirmation theory. In: *The Bloomsbury Companion to the Philosophy of Science,* ed. S. French & J. Saatsi, p. 197. Bloomsbury Academic. [AOH]

Heit, E., Hahn, U. & Feeney, A. (2005) Defending diversity. In: *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*, ed. W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman & P. Wolff, pp. 87–99. American Psychological Association. [EH]

Heit, E. & Rotello, C. M. (2014) Traditional difference-score analyses of reasoning are flawed. *Cognition* 131:75–91. [EH]

Henrich, J., Heine, S. J. & Norenzayan, A. (2010) Most people are not WEIRD. *Nature* 466(7302):29. [DJS]

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. [aRAZ]

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

Hovland, C. I. & Weiss. W. (1951) The influence of source credibility on communication effectiveness. *Public Opinion Quarterly* 15:635–50. [REP]

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. [aRAZ]

Hultsch, D. F. & Hickey, T. (1978) External validity in the study of human development: Theoretical and methodological issues. *Human Development* 21(2):76–91. Available at: https://doi.org/10.1159/000271576. [DMA]

IntHout, J., Ioannidis, J. P. & Borm, G. (2016) Obtaining evidence by a single well-powered trial or by several modestly powered trials. *Statistical Methods in Medical Research* 25:538–52. [JPAI]

Ioannidis, J. P. (2005) Why most published research findings are false. *PLoS Medicine* 2(8):e124. [aRAZ, JPAI]

Ioannidis, J. P. (2008) Why most discovered true associations are inflated. *Epidemiology* 19:640–48. [JPAI]

Ioannidis, J. P. (2013a) Implausible results in human nutrition research. *British Medical Journal* 347:f6698. [JPAI]

Ioannidis, J. P. (2013b) Discovery can be a nuisance, replication is science, implementation matters. *Frontiers in Genetics* 4:33. [JPAI]

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzburg, M., Page, G. P., Petretto, E. & van Noort, V. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics* 41(2):149–55. [MBN]

Ioannidis, J. P. & Trikalinos, T. A. (2007) An exploratory test for an excess of significant findings. *Clinical Trials* 4:245–53. [TEH]

James, W. (1907) Pragmatism's conception of truth. *The Journal of Philosophy, Psychology and Scientific Methods* 4(6):141–55. [AGa]

Jeffreys, H. (1961) *Theory of probability.* Oxford University Press. [aRAZ]

John, L. K., Loewenstein, G. & Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5):524–32. [aRAZ, MBN]

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

Kahneman, D. (2014) A new etiquette for replication. *Social Psychology* 45(4):310–11. [BAS, DJS, WT]

Kanai, R., Bahrami, B., Roylance, R. & Rees, G. (2012) Online social network size is reflected in human brain structure. *Proceedings of the Royal Society B Biological Sciences* 279(1732):1327–34. Available at: https://doi.org/10.1098/rspb.2011.1959. [KG]

Kane, M. T. (2001) Current concerns in validity theory. *Journal of Educational Measurement* 38:319–42. [SOL]

Kerr, N. L. (1998) HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2:196–217. Available at: http://doi.org/10.1207/s15327957pspr0203_4. [aRAZ]

Kevic, K., Murphy, B., Williams, L. & Beckmann, J. (2017) Characterizing experimentation in continuous deployment: A case study on bing. In: *Proceedings of the 39th International Conference on Software Engineering: Software Engineering in Practice Track*, pp. 123–32. IEEE Press. [AGa]

Kidwell, M. C. Lazarevic, L. B. Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S. & Nosek, B. A. (2016) Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology* 14(5):e1002456. Available at: http://doi.org/10.1371/journal.pbio.1002456. [MBN]

Klein, J. R. & Roodman, A. (2005) Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Physics* 55:141–63. [RJM]

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Hunstinger, J. R., IJerzman, H., John, M.-S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A. & Nosek, B. A. (2014a) Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology* 45(3):142–52. Available at: http://doi.org/10.1027/1864-9335/a000178. [aRAZ, PDLH, SS, JLT]

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJerzman, H., John, M.-S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A. & Nosek, B. A. (2014b). Data from investigating variation in replicability: A "Many Labs" Replication Project. *Journal of Open Psychology Data* 2(1):e4. [RG-S]

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJerzman, H., John, M.-S., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A. & Nosek, B. A. (2014c). Theory building through replication: Response to commentaries on the "Many Labs" replication project. *Social Psychology* 45(4):299–311. [TEH]

Koehler, D. J. (2016) Can journalistic "false balance" distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied* 22 (1):24–38. Available at: http://doi.org/10.1037/xap0000073. [MB]

Krupnikov, Y. & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science* 1(1), 59–80. [GPa]

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. [aRAZ]

Kunert, R. (2016) Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin and Review* 23(5):1631–38. [aRAZ]

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

Lakatos, I. (1978) *The methodology of scientific research programs, vol. I.* Cambridge University Press. [EHW]

Lakens, D. (2013) Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4:863. Available at: https://doi.org/10.3389/fpsyg.2013.00863. [DMA]

Lakens, D. (2016) The replication value: What should be replicated? Blog post. Available at: http://daniellakens.blogspot.co.uk/2016/01/the-replication-value-what-should-be.html. [RG-S]

Lakens, D. (2017) Five reasons blog posts are of higher scientific quality than journal articles. Blog post. Available at: http://daniellakens.blogspot.de/2017/04/five-reasons-blog-posts-are-of-higher.html. [BE]

Lawrence, P. A. (2003) The politics of publication. *Nature* 422:259–61. Available at: http://doi.org/10.1038/422259a. [ARK]

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A. & Hoyle, R. H. (2017) Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43 (6):793–813. Available at: https://doi.org/10.1177/0146167217697695. [GPe]

LeBel, E. P., Berger, D., Campbell, L. & Loving, T. J. (2017) Falsifiability is not optional. *Journal of Personality and Social Psychology* 113:254–61. [aRAZ]

LeBel, E. P. & Peters, K. R. (2011) Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology* 15(4):371–79. [RG-S]

Lee, M. D. & Wagenmakers, E. J. (2005) Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review* 112:662–68. [DRL]

Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuitjen, M. B., and Goodman, S. N. (2017) Five ways to fix statistics. *Nature* 551(7682):557–59. [JLT]

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F. & Grahe, J. (2018) Self-esteem, self-disclosure, self-expression, and connection on Facebook: A collaborative replication meta-analysis. *Psi Chi Journal of Psychological Research* 23:98–109. [HI]

Levelt Committee, Noort Committee & Drent Committee (2012, November 28) Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. Retrieved 21 August 2017 from www.tilburguniversity.edu/upload/3ff904d7-547b-40ae85febea38e05a34a_Final%20report%20Flawed%20Science.pdf. [aRAZ]

Lewin, K. (1943/1997). Psychological ecology. In G. W. Lewin & D. Cartwright (Ed.), *Resolving social conflicts & field theory in social science* (pp. 289–300). American Psychological Association. [AGa]

Lewis, M. L. & Frank, M. C. (2016) Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b) *Journal of Experimental Psychology: General* 145:e72–80. [TEH]

Lilienfeld, S. O. & Pinto, M. D. (2015) Risky tests of etiological models in psychopathology research: The need for meta-methodology. *Psychological Inquiry* 26:253–58. [SOL]

Lindsay, D. S., Simons, D. J. & Lilienfeld, S. O. (2016) Research preregistration 101. *Association for Psychological Science Observer* 29:14–17. [SOL]

Lipsey, M. W. & Wilson, D. B. (1993) The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist* 48:1181–209. [RJM]

Little, D. R., Altieri, N., Fific, M. & Yang, C-T. (2017) *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms.* Academic. [DRL]

Lord, F. M. & Novick, M. R. (1968) *Statistical theories of mental test scores*. Addison-Wesley. [EHW]

Lucas, R. E. & Donnellan, M. B. (2013) Improving the replicability and reproducibility of research published in the *Journal of Research in Personality*. *Journal of Research in Personality* 4(47):453–54. [rRAZ]

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

Luttrell, A., Petty, R. E. & Xu, M. (2017) Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology* 69:178–83. [DTW]

Ly, A., Etz, A., Marsman, M. & Wagenmakers, E. J. (2017) Replication Bayes factors from evidence updating. *PsyArXiv preprints*. Available at: https://psyarxiv.com/u8m2s/. [aRAZ]

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. [aRAZ]

Lykken, D. T. (1968) Statistical significance in psychological research. *Psychological Bulletin* 70:151–59. [aRAZ, SOL]

MacCoun, R. J. & Perlmutter, S. (2015) Hide results to seek the truth. *Nature* 526:187–89. [RJM]

MacCoun, R. J. & Perlmutter, S. (2017) Blind analysis as a correction for confirmatory bias in physics and in psychology In: *Psychological science under scrutiny: Recent challenges and proposed solutions*, ed. S. O. Lilienfeld & I. Waldman, pp. 297–322. Wiley. [RJM]

MacKay, D. J. (1992) Information-based objective functions for active data selection. *Neural Computation* 4(4):590–604. [TEH]

Maher, B. & Anfres, M. S. (2016) Young scientists under pressure: What the data show. *Nature* 538:444–45. Available at: http://doi.org/10.1038/538444a [ARK]

Makel, M. C., Plucker, J. A. & Hegarty, B. (2012) Replications in psychology research: How often do they occur? *Perspectives on Psychological Science* 7 (6):537–42. [aRAZ, PDLH]

Manicas, P. T. & Secord, P. F. (1983) Implication for psychology of the new philosophy of science. *American Psychologist* 38(4):399–413. Available at: https://doi.org/10.1037/0003-066X.38.4.399. [DMA]

Many junior scientists need to take a hard look at their job prospects. Editorial. (2017) *Nature* 550(7677):429. Available at: http://doi.org/10.1038/550429a. [ARK]

Martin, G. N. & Clarke, M. (2017) Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology* 8(523):1–6. [PDLH]

Martinez, J. E., Funk, F. & Todorov, A. (2018). Quantifying idiosyncratic and shared contributions to stimulus evaluations. Available at: http://psyarxiv.com/6vr8z. [AGa]

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):e1–15. Available at: http://doi.org/10.1037/xge0000038. [aRAZ]

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* 70:487–98. Available at: http://dx.doi.org/10.1037/a0039400. [aRAZ]

McShane, B. B. & Bockenholt, U. (2017) Single paper meta-analysis: Benefits for study summary, theory-testing, and replicability. *Journal of Consumer Research* 43(6):1048–63. [JLT]

McShane, B. B. & Bockenholt, U. (2018) Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika* 83(1):255–271. [JLT]

McShane, B. B., Bockenholt, U. & Hansen, K. T. (2016) Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science* 11(5):730–49. [JLT]

McShane, B. B. & Gal, D. (2016) Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science* 62(6):1707–18. [JLT]

McShane, B. B. & Gal, D. (2017) Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association* 112(519):885–95. [JLT]

McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2017) *Abandon statistical significance.* Technical report, Northwestern University. Available at: https://arxiv.org/abs/1709.07588. [AGe, JLT]

Meehl, P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* 34(2):103–15. [DMA, DRL]

Meehl, P. E. (1990a) Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports* 66(1):195–244. Available at: https://doi.org/10.2466/pr0.1990.66.1.195. [DMA, TC, DRL]

Meehl, P. E. (1990b) Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry* 1:108–41. Available at: http://doi.org/10.1207/s15327965pli0102_1. [aRAZ, TC, SS]

Mellers, B., Hertwig, R. & Kahneman, D. (2001) Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science* 12:269–75. [aRAZ]

Merton, R. K. (1942) The normative structure of science. In: *The sociology of science: Theoretical and empirical investigations*, ed. R. K. Merton, pp. 267–80. University of Chicago Press. [SS]

Mihura, J. L., Meyer, G. J., Dumitrascu, N. & Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psychological Bulletin* 139, 548–605. [SOL]

Mill, J. S. (1882/2014) *A system of logic, 8th edition*. Harper and Brothers. (Original work published 1882.) Available at: https://ebooks.adelaide.edu.au/m/mill/john_stuart/system_of_logic/index.html [SS]

Morey, R. D. & Lakens, D. (2016) Why most of psychology is statistically unfalsifiable. Available at: https://github.com/richarddmorey/psychology_resolution/blob/master/paper/response.pdf [aRAZ]

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K., Yantis, C. A. & Skitka, L. J. (2017) The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology* 113(1):34. [TC]

Mullen, B., Muellerleile, P. & Bryant, B. (2001) Cumulative meta-analysis: A consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin* 27(11):1450–62. [TEH]

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

National Cancer Institute–National Human Genome Research Institute (NCI-NHGRI) Working Group on Replication in Association Studies (2007) Replicating genotype-phenotype associations. *Nature* 447:655–60. [aRAZ]

Nelson, L. D., Simmons, J. P. & Simonsohn, U. (2018) Psychology's renaissance. *Annual Review of Psychology* 69:511–34. Available at: http://doi.org/10.1146/annurev-psych-122216-011836. [BE, BAS]

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R. & Yarkoni, T. (2015) Promoting an open research culture. *Science* 348:1422–25. [BAS]

Nosek, B. A. & Bar-Anan, Y. (2012) Scientific utopia: I. Opening scientific communication. *Psychological Inquiry* 23:217–43. Available at: http://doi.org/10.1080/1047840X.2012.692215. [BE]

Nosek, B. A. & Errington, T. M. (2017) Making sense of replications. *eLife* 6:e23383. [arRAZ]

Nosek, B. A., Spies, J. R. & Motyl, M. (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7(6):615–31. Available at: http://doi.org/10.1177/1745691612459058. [ARK, GPe]

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., Van Assen, M. A. L. M. & Wicherts, J. M. (2017) Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology* 3(1):1–22. Available at: http://doi.org/10.1525/collabra.102. [MBN]

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S. & Wicherts, J. M. (2016) The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods* 48(4):1205–26. Available at: http://doi.org/10.3758/s13428-015-0664-2. [EH, MBN]

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716. Available at: http://doi.org/10.1126/science.aac4716. [arRAZ, MB, BE, RG-S, EH, PDLH, AK, JLT, US, WT, EHW]

Paluck, E. L. & Shafir, E. (2017) The psychology of construal in the design of field experiments. *Handbook of Economic Field Experiments* 1:245–68. [AGa]

Pashler, H. & Harris, C. (2012) Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science* 7:531–36. [aRAZ]

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* 7:528–30. [aRAZ]

Peer, E., Vosgerau, J. & Acquisti, A. (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46 (4):1023–31. [GPa]

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J. & Fugelsang, J. A. (2015) On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making* 10(6):549–63. [GPe]

Petrocelli, J., Clarkson, J., Whitmire, M. & Moon, P. (2012) When $ab \neq c - c'$: Published errors in the reports of single-mediator models. *Behavior Research Methods* 45(2):595–601. Available at: http://doi.org/10.3758/s13428-012-0262-5 [MBN]

Petty, R. E. (2015) The replication crisis: Social psychology versus other sciences. Paper presented at the annual meeting of the Society of Experimental Social Psychology, Denver, CO. [REP]

Petty, R. E. & Cacioppo, J. T. (2016) Methodological choices have predictable consequences in replicating studies on motivation to think: Commentary on Ebersole *et al*. (2016). *Journal of Experimental Social Psychology* 67:86–87. [DTW]

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R. & Frank, M. C. (2015) A second look at automatic theory of mind. *Psychological Science* 26 (9):1353–67. [TEH]

Pitt, J. C. (1990) The myth of science education. *Studies in Philosophy and Education* 10:7–17. Available at: http://doi.org/10.1007/BF00367684 [MB]

Platt, J. R. (1964) Strong inference. *Science* 146(3642):347–53. [TEH]

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J. B., Vul, E. & Yarkoni, T. (2017) Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* 18(2):115–26. Available at: http://doi.org/10.1038/nrn.2016.167 [ARK]

Popper, K. (1959) *The logic of scientific discovery*. Routledge. [SS]

Popper, K. R. (1959) *Logic of scientific discovery*. Basic. [WT]

Popper, K. R. (1959/2002) *The logic of scientific discovery*, translation of Logik der Forschung. Routledge. [aRAZ]

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A. & Greene, J. D. (2014) Social heuristics shape intuitive cooperation. *Nature Communications* 5:3677. [GP]

Rewarding negative results keeps science on track. Editorial. (n.d.). *Nature* 551:414. Available at: http://www.nature.com/articles/d41586-017-07325-2. [KG]

Rieth, C. A., Piantadosi, S. T., Smith, K. A. & Vul, E. (2013) Put your money where your mouth is: Incentivizing the truth by making nonreplicability costly. *European Journal of Personality* 27:120–44. [AOH]

Robertson, C. T. & Kesselheim, A. S. (2016) *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. Academic. [RJM]

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. [aRAZ]

Rosen, G. M. (1993). Self-help or hype? Comments on psychology's failure to advance self-care. *Professional Psychology: Research and Practice* 24:340–45. [SOL]

Rosenthal, R. (1979) The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86(3):638–41. [aRAZ]

Rotello, C. M., Heit, E. & Dubé, C. (2015) When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin and Review* 22:944–54. [aRAZ, EH]

Rotello, C. M., Masson, M. E. J. & Verde, M. F. (2008) Type I error rates and power analyses for single-point sensitivity measures. *Perception and Psychophysics* 70:389–401. [EH]

Rothstein, H. R. & Bushman, B. J. (2012) Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods* 17:129–36. [aRAZ]

Royal Netherlands Academy of Arts and Sciences (2018) *Replication studies. Improving reproducibility in the empirical sciences*. KNAW. [MBN]

Salmon, W. C. (1984) *Scientific explanation and the causal structure of the world*. Princeton University Press. [EH]

Schimmack, U. (2012) The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods* 17:551–56. [US]

Schimmack, U. (2014) The test of insufficient variance (TIVA): A new tool for the detection of questionable research practices. Working paper. Available at: https://replicationindex.wordpress.com/2014/12/30/the-test-of-insufficient-variance-tiva-a-new-tool-for-the-detection-of-questionable-research-practices/. [US]

Schimmack, U. (2017) 'Before you know it' by John A. Bargh: A quantitative book review. Available at: https://replicationindex.wordpress.com/2017/11/28/before-you-know-it-by-john-a-bargh-a-quantitative-book-review/. [US]

Schimmack, U. & Brunner, J. (submittrd for publication) Z-Curve: A method for estimating replicability based on test statistics in original studies. Submitted for Publication. [US]

Schimmack, U., Heene, M. & Kesavan, K. (2017) Reconstruction of a train wreck: How priming research went off the rails. Blog post. Available at: https://replicationindex.wordpress.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/. [US]

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. Available at: http://dx.doi.org/10.1037/arc0000029. [aRAZ]

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. [aRAZ, AGe]

Schnall, S. (2014a) An experience with a registered replication project. Blog post. Available at: http://www.psychol.cam.ac.uk/cece/blog#anchor-2. [WT]

Schnall, S. (2014b) Further thoughts on replications, ceiling effects and bullying. Blog post. Available at: http://www.psychol.cam.ac.uk/cece/blog. [WT]

Schnall, S. (2014c) Social media and the crowd-sourcing of social psychology. Blog post. Available at: http://www.psychol.cam.ac.uk/cece/blog. [WT]

Schönbrodt, F. D. (2018). *p-checker: One-for-all p-value analyzer*. Available at: http://shinyapps.org/apps/p-checker/. [MBN]

Schooler, J. (2011) Unpublished results hide the decline effect. *Nature* 470:37. [WT]

Schooler, J. (2014) Metascience could rescue the 'replication crisis'. *Nature* 515:9. [WT]

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., Ly, A., Marsman, M., Murase, T., Qureshi, I., Schaerer, M., Thornley, N., Tworek, C. M., Wagenmakers, E.-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M. A., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K. Inbar, Y., Jiang, L., Kellogg, T., Kennedy, D. M., Legate, N., Luoma, T. P., Maibuecher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A.-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei, C., Wetter, E., Wilson, C. T., Darrous, S.-C. & Uhlmann, E. L. (2016) The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology* 66:55–67. [aRAZ, WT]

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., Tamman, A. J. & Puhlmann, L. (2015) Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General* 144(6):e142–58. [aRAZ]

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. , Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T. J., Hederos, K., Heene, M., Hofelich Mohr, A. J., Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D. M., Lei, R., Lindsay, T. A., Liverani, S., Madan, C. R., Molden, D., Molleman, E., Morey, R. D., Mulder, L. B., Nijstad, B. R., Pope, N. G., Pope, B., Prenoveau, J. M.,

Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F. D., Sherman, M. F., Sommer, S. A., Sotak, K., Spain, S., Spörlein, C., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S. & Nosek, B. A. (2017) Many analysts, one dataset: Making transparent how variations in analytical choices affect results. PsyArXiv Preprint. Available at: https://psyarxiv.com/qkwst/. [aRAZ]

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* 22:1359–66. Available at: http://doi.org/10.1177/0956797611417632. [aRAZ, BE, MBN, BAS]

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2018) False-positive citations. *Perspectives on Psychological Science* 13(2):255–59. [BE]

Simmons, J. & Simonsohn, U. (2015) Power posing: Reassessing the evidence behind the most popular TED talk. Blog post. Available at: http://datacolada.org/37. [AGe]

Simons, D. J., Holcombe, A. O. & Spellman, B. A. (2014) An introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science* 9(5):552–55. [BAS, JLT]

Simons, D. J., Shoda, Y. & Lindsay, D. S. (2017) Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science* 12:1123–28. Available at: http://doi.org/10.1177/1745691617708630. [aRAZ, AOH, HI, DJS]

Simonsohn, U. (2015) Small telescopes: Detectability and the evaluation of replication results. *Psychological Science* 26:559–69. [aRAZ, GPa]

Simonsohn, U. (2016, March 3) [47] Evaluating replications: 40% full≠60% empty. Available at: https://web.archive.org/web/20170709184952/http://datacolada.org/47. [aRAZ]

Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014) P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General* 143(2):534–47. [TEH]

Simpson, E. H. (1951) The Interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 13(2):238–41. [JPdR]

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

Smart, R. G. (1964) The importance of negative results in psychological research. *Canadian Psychologist* 5:225–32. [aRAZ]

Smith, G. T. McCarthy, D. M. & Anderson, K. G. (2000) On the sins of short-form development. *Psychological Assessment* 12:102–11. [SOL]

Smith, P. L. & Little, D. R. (2018) Small is beautiful: In defence of the small-*N* design. *Psychonomic Bulletin & Review*. Available at: https://doi.org/10.3758/s13423-018-1451-8. [DRL]

Spellman, B. A. (2015) A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science* 10:886–99. [aRAZ, BAS]

Sripada, C., Kessler, D. & Jonides, J. (2014) Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science* 25:1227–34. [AK]

Srivastava, S. S. (2011, December 31) Groundbreaking or definitive? Journals need to pick one. Blog post. Available at: https://spsptalks.wordpress.com/2011/12/31/groundbreaking-or-definitive-journals-need-to-pick-one/. [SS]

Srivastava, S. (2012, September 17) A Pottery Barn rule for scientific journals. Blog post. Available at: https://hardsci.wordpress.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/. [rRAZ]

Ståhl, T., Zaal, M. P. & Skitka, L. J. (2016) Moralized rationality: Relying on logic and evidence in the formation and evaluation of belief can be seen as a moral issue. *PLoS One* 11(11):e0166332. Available at: https://doi.org/10.1371/journal.pone.0166332. [GPe]

Stanley, T. D., Carter, E. C. & Doucouliagos, H. (November 2017) What meta-analyses reveal about the replicability of psychological research. Deakin Laboratory for the Meta-Analysis of Research, Working Paper . [JLT]

Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016) Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* 11 (5):702–12. Available at: http://doi.org/10.1177/1745691616658637. [THE, MBN]

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(285):30–34. Available at: http://doi.org/10.2307/2282137. [aRAZ, US]

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* 49:108–12. [aRAZ]

Sternberg, S. (1969) The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica* 30:276–315. [DRL]

Stewart, N., Chandler, J. & Paolacci, G. (2017) Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences* 21(10):736–48. [GPa]

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G. & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making* 10(5):479–91. [GPa]

Stirman, S. W., Gamarra, J. M., Bartlett, B. A., Calloway, A. & Gutner, C. A. (2017) Empirical examinations of modifications and adaptations to evidence based psychotherapies: Methodologies, impact, and future directions. *Clinical Psychology: Science and Practice* 24:396–420. [SOL]

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. & Taufer, M. (2016) Enhancing reproducibility for computational methods. *Science* 354(6317):1240–41. [TEH]

Strack, F. (2017) From data to truth in psychological science. A personal perspective. *Frontiers in Psychology* 8:702. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5432643/. [FS]

Strevens, M. (2001) The Bayesian treatment of auxiliary hypotheses. *The British Journal for the Philosophy of Science* 52(3):515–37. [AOH, rRAZ]

Strevens, M. (2006) The bayesian approach to the philosophy of science: In: *Encyclopedia of philosophy*, ed. D. M. Borchert, pp. 495–502. Macmillan Reference. [TEH]

Strevens, M. (2017) Notes on Bayesian confirmation theory [paper]. Available at: http://www.nyu.edu/classes/strevens/BCT/BCT.pdf. [AOH]

Stroebe, W. & Strack, F. (2014) The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science* 9:59–71. Available at: http://doi.org/10.1177/1745691613514450. [aRAZ, FS, AMT, DTW]

Szucs, D. & Ioannidis, J. P. (2017a) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology* 15(3):e2000797. Available at: http//doi.org/10.1371/journal.pbio.2000797. [ARK]

Szucs, D. & Ioannidis, J. P. A. (2017b) When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience* 11:390. Available at: https://doi.org/10.3389/fnhum.2017.00390. [DMA]

Tackett, J. L., Brandes, C. M. & Reardon, K. W. (in press) Leveraging the Open Science Framework in clinical psychological assessment research. *Psychological Assessment* [JLT]

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L, Krueger, R. F, Miller, J. D., Oltmans, T. F. & Shrout, P. E. (2017a) It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science* 12(5):742–56. [JLT, SOL]

Tackett, J. L., McShane, B. B., Bockenholt, U. & Gelman, A. (2017b) Large scale replication projects in contemporary psychological research. Technical report, Northwestern University. Available at: arXiv:1710.06031. [JLT]

Thomson, K. S. & Oppenheimer, D. M. (2016) Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making* 11(1):99–113. [GPa]

Tierney, W., Schweinsberg, M., Jordan, J., Kennedy, D. M., Qureshi, I., Sommer, A., Thornley, N., Madan, N., Vianello, M., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y, Ly, M. M., Murase, T., Schaerer, M., Tworek, C. M., Wagenmakers, E-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M. A., Cordon, D., Cushman, F., Ditto, P. H., Amell, A., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Legate, N., Luoma, T. P., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packhan, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A.-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei, C., Wetter, E., Wilson, C. T., Darroux, S.-C. & Uhlmann, E. L. (2016) Data from a pre-publication independent replication initiative examining ten moral judgement effects. *Nature Scientific Data* 3:160082. Available at: http://doi.org/10.1038/sdata.2016.82. [WT]

Tracy, J. L. & Beall, A. T. (2014) The impact of weather on women's tendency to wear red or pink when at high risk for conception. *PLoS One* 9(2):e88852. [AGe]

Traxler, M. J. & Gernsbacher, M. A. (1992) Improving written communication through minimal feedback. *Language and Cognitive Processes* 7:1–22. Available at: https://doi.org/10.1080/01690969208409378. [MAG]

Traxler, M. J. & Gernsbacher, M. A. (1993) Improving written communication through perspective-taking. *Language and Cognitive Processes* 8:311–34. Available at: https://doi.org/10.1080/01690969308406958. [MAG]

van Aert, R. C. & van Assen, M. A. (2017) Bayesian evaluation of effect size after replicating an original study. *PLoS One* 12(4):e0175302. [aRAZ]

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 of the United States of America* 113(23):6454–59. [aRAZ]

van Erp, S., Verhagen, A. J., Grasman, R. P. P. P. & Wagenmakers, E.-J. (2017) Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data* 5 (1):4. DOI: http://doi.org/10.5334/jopd.33. [JLT]

Vanpaemel, W., Vermorgen, M., Deriemaecker, L. & Storms, G. (2015) Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra* 1(1):1–5. Available at: http://doi.org/10.1525/collabra.13. [MBN]

Veldkamp, C. L. S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A. L. M. & Wicherts, J. M. (2014) Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One* 9(12):e114876. Available at: http://doi.org/10.1371/journal.pone.0114876. [MBN]

Verhagen, A. J. & Wagenmakers, E.-J. (2014) Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General* 143:1457–75. [aRAZ]

Vohs, K. D. (2018) A pre-registered depletion replication project: The paradigmatic replication approach. Presented at the Symposium at the 2018 Society of Personality and Social Psychology Annual Convention, Atlanta, GA. [BAS]

Vul, E., Harris, C., Winkielman, P. & Pashler, H. (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* 4:274–90. Available at: http://doi.org/10.1111/j.1745-6924.2009.01125.x. [BE]

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K, Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupianez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Ozdoğru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E. & Zwaan, R. A. (2016a) Registered replication report: Strack, Martin & Stepper (1988). *Perspectives on Psychological Science* 11:917–28. [arRAZ]

Wagenmakers, E.-J., Verhagen, A. J. & Ly, A. (2016b) How to quantify the evidence for the absence of a correlation. *Behavior Research Methods* 48:413–26. [aRAZ]

Wagge, J., Johnson, K., Meltzer, A., Baciu, C., Banas, K., Nadler, J. T., IJzerman, H. & Grahe, J. E. (in preparation). Elliott *et al.*'s (2011) "Red, rank, and romance" effect: A meta-analysis of CREP replications. [HI]

Wald, A. (1947) *Sequential analysis*. Wiley. [EHW]

Wald, A. (1950) *Statistical decision functions*. Wiley. [NAC]

Wallot, S. & Kelty-Stephen, D. G. (2018) Interaction-dominant causation in mind and brain, and its implication for questions of generalization and replication. *Minds and Machines* 28(2):353–74. Available at: https://doi.org/10.1007/s11023-017-9455-0. [DMA]

Washburn, A. N., Hanson, B. E., Motyl, M., Skitka, L., Yantis, C., Wong, K., Sun, J., Prims, J., Mueller, A. B., Melton, Z. J. & Carsel, T. S. (2018) Why do some psychology researchers resist using proposed reforms to research practices? A description of researchers' rationales. *Advances in Methods and Practices in Psychological Science*. Published online March 7, 2018. Available at: https://doi.org/10.1177/2515245918757427. [TC]

Wicherts, J. M., Bakker, M. & Molenaar, D. (2011) Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One* 6(11):e26828. Available at: http://doi.org/10.1371/journal.pone.0026828. [MBN]

Wicherts, J. M., Borsboom, D., Kats, J. & Molenaar, D. (2006) The poor availability of psychological research data for reanalysis. *American Psychologist* 61:726–28. Available at: http://doi.org/10.1037/0003-066X.61.7.726. [MBN]

Widaman, K. (2015) Confirmatory theory testing: Moving beyond NHST. The score. Newsletter. Available at: http://www.apadivisions.org/division-5/publications/score/2015/01/issue.pdf. [DMA]

Witte, E. H. & Melville, P. (1982) Experimentelle Kleingruppenforschung: Methodologische Anmerkungen und eine empirische Studie. [Experimental small group research: Methodological remarks and an empirical study.] *Zeitschrift für Sozialpsychologie* 13:109–24. [EHW]

Witte, E. H. & Zenker, F. (2016a) Reconstructing recent work on macro-social stress as a research program. *Basic and Applied Social Psychology* 38(6):301–307. [EHW]

Witte, E. H. & Zenker, F. (2016b) Beyond schools – reply to Marsman, Ly & Wagenmakers. *Basic and Applied Social Psychology* 38(6):313–17. [EHW]

Witte, E. H. & Zenker, F. (2017a) Extending a multilab preregistered replication of the ego-depletion effect to a research program. *Basic and Applied Social Psychology* 39(1):74–80. [EHW]

Witte, E. H. & Zenker, F. (2017b) From discovery to justification. Outline of an ideal research program in empirical psychology. *Frontiers in Psychology* 8:1847. Available at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01847/full. [EHW]

Wood, J. M., Garb, H. N., Nezworski, M. T., Lilienfeld, S. O. & Duke, M. C. (2015) A second look at the validity of widely used Rorschach indices: Comment on Mihura, Meyer, Dumitrascu, and Bombel (2013). *Psychological Bulletin* 141:236–49. [SOL]

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):369–90. [aRAZ]

Yarkoni, T. & Westfall, J. (2017) Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 1–23. [DRL]

Zwaan, R. A. (2017, May 8) Concurrent replication. Blog post. Available at: https://rolfzwaan.blogspot.nl/2017/05/concurrent-replication.html. [rRAZ, MAG]

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 and Review*. Available at: http://doi.org/10.3758/s13423-017-1348-y. [aRAZ, GPa]