Original Manuscript

A Validity-Based Framework for Understanding Replication in Psychology

Personality and Social Psychology Review
1-29

DOI: 10.1177/1088868320931366
pspr.sagepub.com


Leandre R. Fabrigar[1], Duane T. Wegener[2], and Richard E. Petty[2]

Abstract

In recent years, psychology has wrestled with the broader implications of disappointing rates of replication of previously demonstrated effects. This article proposes that many aspects of this pattern of results can be understood within the classic framework of four proposed forms of validity: statistical conclusion validity, internal validity, construct validity, and external validity. The article explains the conceptual logic for how differences in each type of validity across an original study and a subsequent replication attempt can lead to replication "failure." Existing themes in the replication literature related to each type of validity are also highlighted. Furthermore, empirical evidence is considered for the role of each type of validity in non-replication. The article concludes with a discussion of broader implications of this classic validity framework for improving replication rates in psychological research.

Keywords

research methods, replication, statistical conclusion validity, construct validity, internal validity, external validity

Introduction

There is little dispute that psychology, and perhaps personality and social psychology in particular, is experiencing a crisis of confidence regarding its methodology. Many of its traditional research practices have come under critical scrutiny. Indeed, the existence of a number of well-known phenomena in personality and social psychology such as social priming (e.g., Harris et al., 2013; Shanks et al., 2013) and ego-depletion (e.g., Hagger et al., 2016) have been called into question. This crisis in the discipline is also reflected in the vast methodological literature that has accumulated in recent years regarding potentially problematic research practices. Numerous journals have devoted special issues or special sections to the topic (e.g., *Perspectives on Psychological Science* in 2011, 2012, & 2014; *Journal of Mathematical Psychology* in 2013; *Journal of Experimental Social Psychology* in 2015 & 2016). Likewise, a number of edited and authored books have appeared recently with titles indicating serious concerns regarding psychological research such as *Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions* (Lilienfeld & Waldman, 2017), *The Seven Deadly Sins in Psychology: A Manifesto for Reforming the Culture of Scientific Practice* (Chambers, 2017), and *Psychology in Crisis* (Hughes, 2018). Concerns reflected by such publications have also led to the creation of new organizations whose goal is to reform research practices (e.g., Center for Open Science, Society for the Improvement of Psychological Science) and garnered attention in articles and op-eds in prominent media outlets such as the *New York Times* (e.g., Barrett, 2015; Carey, 2018).

Although the causes of this crisis regarding methods in psychology are no doubt multiple, nothing has brought more attention to the issue than the results of several large-scale replication attempts (e.g., Ebersole et al., 2016; Klein et al., 2014, 2018; Open Science Collaboration, 2015). Notably, these efforts have often reported disappointingly low levels of successful replication of previously documented effects (sometimes 50% or less). These replication rates have been far from ideal and less than might have been expected. However, the reasons for low replication rates and their broader implications for the field have been a matter of intense debate.

Some scholars have seen low replication rates as indicative of fundamental flaws in the traditional way psychological research has been conducted. According to these psychologists, attempts to replicate previously published psychological effects often fail because the originally published effects were spurious and resulted from problematic methodological practices (e.g., L. K. John et al., 2012;

[1] Queen's University, Kingston, Ontario, Canada
[2] The Ohio State University, Columbus, USA

Corresponding Author:
Leandre R. Fabrigar, Department of Psychology, Queen's University, Kingston, Ontario, Canada K7L 3N6.
Email: fabrigar@queensu.ca

Simmons et al., 2011). Such a view is exemplified by statements suggesting that under certain assumptions, at least, a majority of published results in psychology could be false (e.g., Pashler & Harris, 2012), a claim that has been made more explicitly for medical research (Ioannidis, 2005). Similarly, news stories on the crisis have noted that some scholars have argued that traditional research practices are so flawed that it is necessary "to burn things to the ground" (Bartlett, 2018). From the standpoint of such critics, current replication efforts represent more informative explorations of psychological phenomena than the original studies.

In contrast, other scholars have not interpreted low rates of replication as suggesting a hopelessly flawed discipline. Rather, advocates of this viewpoint note that failure to obtain previously demonstrated effects could have resulted from the complexity of the original phenomena under investigation and/or potentially problematic methods used in the replication efforts (e.g., Gilbert et al., 2016; Stroebe & Strack, 2014; Van Bavel et al., 2016). These commentators have argued that replication failures can result when researchers inadequately take into account such complexities, leading to erroneous interpretations of their results. Such views have led some to regard the conclusions and actions taken by some critics as misguided and perhaps even destructive (Bartlett, 2018). From this point of view, some recent replication efforts are not necessarily more informative and could even be less informative than the original studies they aim to replicate.

Overview

The goal of the present article is not to argue for one side or another. Rather, our goal is to suggest a conceptual framework for organizing and understanding the many themes that have emerged. We believe that viewing replication issues through the lens of an organized conceptual framework can contribute to the debate in at least three ways. First, the conceptual lens can help to highlight similarities and differences among the various explanations for and solutions to disappointing replication rates that have been advanced in the literature. Second, the conceptual lens can help to identify some of the unstated and/or untested assumptions underlying these explanations and proposed solutions. Finally, the lens can focus attention on new or neglected explanations and potential solutions.

A Validity-Based Conceptual Framework

The central premise of the present article is that, when considering why a replication study has failed to reproduce the findings of an earlier study, it is useful to consider this question in light of the now classic research validity typology originally proposed by Cook and Campbell (1979). Building on earlier work by Campbell and Stanley (1966), Cook and Campbell proposed that any study could be evaluated in terms of four fundamental types of validity: statistical conclusion validity, internal validity, construct validity, and external validity (see also Shadish et al., 2002). This validity typology continues to be prominently featured in many contemporary discussions of research methods (e.g., Brewer & Crano, 2014; Crano et al., 2015; Kenny, 2019; E. R. Smith, 2014; West et al., 2014). Although the four validity types have long been regarded as useful in evaluating original studies, we argue that they can also be helpful in designing replication studies and interpreting their results. More specifically, we postulate that anytime a replication study has failed to obtain the same results as the original study, this discrepancy is likely a function of differences between the studies in one or more of these four types of validity. To more concretely illustrate this assertion, we consider each of the four types of validity in turn and discuss the logic for how each validity might play a role in any given failure to replicate a previously demonstrated finding. For purposes of our discussion, we focus on a scenario in which an original study provides evidence supportive of an effect and a replication study has failed to produce significant evidence of that effect. However, the logic we present can also be readily applied to understanding other patterns of discrepancy between original studies and their replications.[1] For purposes of simplicity, we illustrate the role of these four types of validity in the context of the experimental designs that consume the bulk of replication efforts in psychology, though most of the observations are also applicable to nonexperimental studies. We next describe each of the four types of validity and their applicability to the replication crisis.[2]

Statistical Conclusion Validity

Statistical conclusion validity refers to the accuracy of a conclusion regarding a relation between or among variables of interest. The conclusion is accurate if claiming a particular relation exists when there really is such a relation in the population or claiming no relation exists when no relation exists in the population (Cook & Campbell, 1979). The term "relation" is used broadly to refer to a wide range of statistical indices (e.g., measures of association, tests of differences in means) and can be applied to simple bivariate relations as well as more complex relations such as interaction effects. In most discussions, violations of statistical conclusion validity can take one of the two forms: Type I error (i.e., concluding that a relation exists when there is no relation) or Type II error (i.e., concluding that no relation exists when, in fact, a relation is present). When considering a situation in which original researchers have concluded there is a relation and replication researchers conclude that there is no relation, a statistical conclusion validity perspective suggests two possibilities. First, the original study might have produced a Type I error and the replication study correctly failed to find evidence of a relation.

Alternatively, the original study might have correctly suggested the existence of a relation and the replication study produced a Type II error.

Our survey of the contemporary replication literature reveals that concerns regarding statistical conclusion validity have strongly shaped the views of many psychologists in this debate. Indeed, among those who argue that low replication rates reflect fundamental problems in psychology, a common inference has been that if a replication study has adequate statistical power, failure to replicate is a function of the original study having produced a Type I error. That is, the original study is presumed to be lower in statistical conclusion validity than the replication study. This belief that poor statistical conclusion validity producing a Type I error in original studies (e.g., due to low statistical power) plays a central role in low replication rates can be illustrated by numerous statements in the recent replication literature such as those below:

> Among all proposed reforms to make our research more reproducible, I am most interested in increasing the statistical power of our published studies. (Vazire, 2016, p. 4)

> So, in low-powered studies, significant results are more likely to be false-positive errors, and p-values close to .05 are particularly untrustworthy, because they are unlikely to reach the alpha level required in an underpowered study to achieve an acceptable false discovery rate . . . The downsides of false positives are well known. Creating interventions and further studies based on false positives is a misuse of resources. (Giner-Sorolla et al., 2019, p. 18)

> Low power is a problem in practice because of the normative publishing standards for producing novel, significant, clean results and the ubiquity of null hypothesis significance testing as the means of evaluating the truth of research findings. As we have shown, these factors result in biases that are exacerbated by low power. Ultimately, these biases reduce the reproducibility of neuroscience findings and negatively affect the validity of the accumulated findings. (Button et al., 2013, p. 373)

Thus, one overarching theme of the contemporary replication literature has been that numerous failures to replicate original findings suggest in part that Type I errors are pervasive in the psychological literature and that a critical challenge for the field is to identify why Type I errors are so common and to develop strategies for minimizing them. To date, a number of potential answers have been offered to these challenges, and the various answers constitute many of the most prominent themes in the contemporary replication literature.

For example, as the quotes above indicate, a number of methodologists have noted the prevalence of published original studies with low power. They postulate that because (a) such studies are more prone to producing extreme effect size estimates, (b) it is possible to conduct more of these small-sample studies than highly powered large-sample studies, and (c) there is a bias against publishing null results (Greenwald, 1975), then low power might have played a major role in the publication of false positive effects (e.g., Button & Munafò 2017). This view has led some commentators to advocate stricter standards for statistical power in psychological research as a means of enhancing replicability (e.g., Vazire, 2016). Others have argued that the traditional alpha of .05 provides inadequate protection against Type I error and should be replaced with more stringent alpha levels in original research (e.g., .005; Benjamin et al., 2017; Greenwald et al., 1996). Part of this argument is that such changes would enhance the replicability of published results (because the published effects would be more likely to be real). Finally, some methodologists have focused on the notion that conducting a large number of small (low power) studies—some of which produce extreme effect sizes that are published—can lead to exaggerated statistical evidence for an effect in the literature. Because of this, some researchers have focused on developing indices aimed at detecting when selective reporting might be occurring in an effort to prevent the publication of exaggerated statistical evidence (e.g., Francis, 2012; Schimmack, 2012; Simonsohn et al., 2014).

Yet another rationale given for prevalent Type I errors in the published literature is the use of *questionable research practices* (QRPs; for example, L. K. John et al., 2012; Simmons et al., 2011) by the authors of original studies. QRPs cover a wide range of behaviors that overlap with the themes already discussed, but QRPs also include additional practices. Most involve ways in which data are collected, analyzed, or reported that can lead (particularly in conjunction with one another) to inflated Type I error rates. Advocates of this rationale have argued that replicability of findings could be greatly enhanced by better educating researchers as to the problematic nature of QRPs and requiring greater transparency in the description of research practices.

Finally, some have argued that false positives are common in psychology because of inherent limitations in the statistical approach that most social scientists use. These critics argue for adopting alternative approaches to traditional null hypothesis significance testing (NHST). Some have even suggested abandoning traditional NHST and instead reporting effect sizes and their corresponding confidence intervals (e.g., Cumming, 2014; Schmidt, 1996). Others have promoted the use of Bayesian statistics (e.g., Hoijtink et al., 2019; Wagenmakers et al., 2017).

None of these explanations or proposed solutions have been without controversy. The soundness of many as explanations and/or solutions has been challenged (e.g., Fiedler et al., 2012; Finkel et al., 2015; Stroebe, 2016). Moreover, not all of the explanations/solutions are necessarily consistent with one another. The goal of this review is not to evaluate the strengths and weaknesses of each viewpoint focused on statistical conclusion validity, but we note that all of these themes share the underlying assumption that

statistical conclusion validity is an important part of the replication crisis.

Before concluding our discussion of statistical conclusion validity, it is worth noting that the traditional Type I/Type II null hypothesis testing (NHST) framework is not the only possible way of talking about statistical conclusion validity. This framing was how Cook and Campbell (1979) conceptualized this form of validity and it remains the dominant perspective among replication researchers and methodologists who point to false finding rates or Type I error as a major problem in psychological research. However, alternatives to traditional NHST have been proposed. For example, Bayesian statistics have been advocated and some might regard Bayesian approaches as stepping outside the notion of true versus false findings. To us, the Bayesian approach (at least as it is most commonly implemented in practice) does not so much move beyond such assumptions as test and describe the findings differently than in traditional NHST.

However, there is another long-held notion that there are no true null hypotheses (e.g., Bakan, 1966; Cohen, 1990; Meehl, 1978; see Morrison & Henkel, 1970) which would require reevaluation of the notion of Type I error. An alternative to testing a null hypothesis with a nondirectional alternative is not to specify a null at all but to conduct a symmetric pair of one-tailed tests, each with $p(\text{error}) = \alpha/2$ (see Jones & Tukey, 2000). The outcome would be to act as if the population effect lies in one direction, the opposite direction, or the sign of the difference is not yet determined. Such conclusions largely match typical conclusions based on NHST but without testing a null hypothesis per se. In this approach, there is no such thing as Type I error. Type II error could still be a failure to make a directional claim when there is a difference in the population, but the only claim made in error would be one of the incorrect direction (Type III; Shaffer, 2002). We agree that the point null might never be completely true. If so, "false" findings might generally be considered claims of a direction when the sign of the direction should still be in doubt. Alternatively, researchers might want to include in Type I errors population values that fall close enough to zero to fall within a "loose" null hypothesis (though there are rarely clear criteria for what would be close enough; see Bakan, 1966, for discussion). Because the notion of Type I error continues to play a key role in recent methodological discussions, we retain that language throughout the current article, but we also believe it reasonable to note that the notion of Type I error per se might require adjustment acknowledging the plausible and widespread belief that no population difference between psychologically relevant conditions will truly be strictly zero.[3]

Regardless of whether one adopts a traditional NHST view of statistical conclusion validity or an alternative perspective, the threats to statistical conclusion validity are still largely the same (e.g., the distorting effects of QRPs). Moreover, as indicated earlier, we argue that statistical conclusion validity is only one of the four categories of explanations for non-replication. Thus, it is useful to turn our attention to each of the other three types of validity and their relevance for the replication crisis.

Internal Validity

In many cases, psychologists are not simply interested in establishing whether there is a relation between or among variables. Their theorizing allows them to conceptually designate one or more variables as an independent variable (IV; or predictor variable) and one or more variables as a dependent variable (DV). That is, the researcher wishes to postulate a causal relation (i.e., the IV as operationalized produces the observed changes in the DV as operationalized). *Internal validity* refers to the extent to which a relation among variables can be interpreted as causal in nature (i.e., whether the IV as manipulated plausibly caused observed differences in the DV; Cook & Campbell, 1979). In the simplest two-condition experiment in which an experimental treatment is compared with a control condition (i.e., absence of treatment), the question is whether the difference observed between the treatment and control condition on the DV was produced by the presence of the treatment. In nonexperimental studies, internal validity is comparatively low and it is generally quite difficult to reach firm causal conclusions. In the case of experiments involving random assignment to conditions, internal validity is higher and the basis for causal inferences is much stronger.

Even in randomized experiments with appropriate control groups, however, threats to internal validity can arise. For example, any factor that introduces a post hoc violation of random assignment, such as differential attrition in experimental conditions, can compromise internal validity. If a large number of research participants fail to complete the experiment in one condition and nearly all participants complete the experiment in another condition, this could violate random assignment if the participants who drop out differ on some relevant psychological dimension from those who complete the study. Indeed, if dropout rates are nontrivial in magnitude but at comparable levels in both conditions, this attrition could still threaten internal validity if the psychological factors producing dropouts in the two conditions are different.

For purposes of the present discussion, such a threat to internal validity could be problematic for two reasons. First, if the violation of random assignment introduces a preexisting difference in the groups that is in some way related to the DV, it could result in the emergence of a spurious effect. That is, the IV might appear to be exerting an impact on the DV when it has no actual effect. Second, if a violation of random assignment introduces a preexisting difference that is related to the DV in a manner that is opposite the effect of the IV, it could result in the emergence of a spurious null effect. That is, the IV could be exerting an impact on the DV, but this effect could be masked by a countervailing relation

of the preexisting difference in groups with the DV. Thus, from the standpoint of internal validity, there are two possible explanations for why an original study might have demonstrated an effect and a replication study might have failed to obtain this effect. First, the original study might have been high in internal validity, but the replication study introduced a threat to internal validity that masked the effect of the IV on the DV. Second, the original study might have suffered from a threat to internal validity that produced a spurious effect, whereas the replication study was higher in internal validity and thus no spurious effect emerged.

Although there is a conceptual basis for internal validity playing a role in failures to replicate, internal validity differences have not been identified as key to the contemporary replication literature. None of the major themes in this literature have been closely tied to issues of internal validity. Indeed, at first glance, it might be difficult to imagine how at a practical level differences in internal validity between an original study and a replication study might arise. However, such a situation might be more plausible than is initially apparent. For instance, any time a new study involves a change in the nature of the population, setting, or recruitment method that could affect the motivation or ability of participants to complete the study, this could result in differences in participant attrition across studies.

Notably, data collections in online environments such as Amazon's Mechanical Turk (MTurk) have become common in psychology, whereas 15 years ago data collections in such settings were extremely rare (e.g., see Anderson et al., 2019). Thus, there may be many cases of replication in which the original study was run in a laboratory whereas the replication study was run online. Online studies routinely produce much higher attrition rates than laboratory studies, and it can be difficult to discern attrition rates in popular online platforms such as MTurk. Indeed, researchers collecting online data seldom report attrition rates and how this problem was managed, and direct examinations have suggested that high attrition rates can sometimes result in online experiments producing distorted results (see Zhou & Fishbach, 2016).

To illustrate the potential distortion, Zhou and Fishbach (2016) conducted two online studies to show how attrition could produce seemingly implausible findings. In one study, participants were randomly assigned to list four happy events from the past 12 months or twelve happy events from the past 12 months. They then rated the difficulty of the memory task. This ease of retrieval experiment (Schwarz et al., 1991) produced the surprising finding that participants found the four-event task more difficult than the 12-event task. However, the attrition rate was 69% in the 12-event condition and 26% in the four-event condition, thereby violating random assignment. Thus, the result was likely driven by people who found the 12-event task particularly difficult simply dropping out of the study. In a second study, they randomly assigned participants to describe how applying eyeliner versus shaving cream made them feel. Participants then reported their weight. Surprisingly, the manipulation appeared to influence participants' weight—participants were lighter in the eyeliner condition than the shaving cream condition. Interestingly, this study produced similar levels of attrition (32% in the eyeliner condition and 24% in the shaving cream condition), but attrition in the two conditions was different for men and women, thereby leading to proportionally more women in the eyeliner than the shaving cream condition (a violation of random assignment). Thus, self-reported weight was likely lower in the eyeliner condition because more women were in this condition and their average weight was lower than that of the men in the study. As these studies help illustrate, in light of the high rates of attrition in online studies and their potentially distorting effects, internal validity (especially differential attrition) might play a more important role in replication failures than has been recognized in the literature.

Construct Validity

In basic research, psychologists are primarily interested in formulating and testing general theories regarding psychological and behavioral phenomena. Therefore, the focus is on testing hypotheses regarding relations between psychological constructs (i.e., the conceptual IV and conceptual DV). In any given study, a researcher chooses particular operationalizations of those constructs (i.e., the operational IV and the operational DV). However, no single operationalization is likely to be a pure reflection of its intended construct; the operationalization will also reflect random influences and unintended systematic influences (i.e., unintended alternative constructs or confounds; Cook & Campbell, 1979). Thus, manipulations and measures can vary in the degree to which they correspond to the constructs of interest. *Construct validity* refers to the extent to which the operationalizations of the independent variables (IVs) and dependent variables (DVs) in a study correspond to their intended conceptual variables (constructs). In the context of the current discussion, low construct validity can create one of the two possible problems.

First, low construct validity can result in studies producing "misleading" null effects. Specifically, even if a researcher's hypothesis regarding a relation between the constructs of interest is correct, a study might fail to provide evidence of the effect if the operationalization of the IV or DV is poorly mapped onto the construct. For example, if either the IV or DV operationalization (or both) is contaminated by a high level of random error, no effect might be observed because there is simply too much "noise" in the data to consistently detect the effect even if such an effect exists in the population (e.g., see Stanley & Spence, 2014). Alternatively, if either operationalization substantially reflects an unintended construct, no effect might emerge if the unintended construct is unrelated to the other construct or if it has contradictory effects to the intended construct. For

example, imagine a study designed to test whether perceived competence makes people more willing to undertake difficult tasks. The study manipulates perceptions of competence by having participants do extremely well or only moderately well on a vocabulary test and then presents participants with a choice of attempting different verbal ability tests of varying difficulty. In this case, the intent of the manipulation is to influence perceptions of verbal competence. However, imagine this manipulation has an impact on participants' mood in addition to or instead of perceived verbal competence (i.e., success produces positive mood). The manipulation might not influence task choice if perceptions of competence have not been influenced but mood has been influenced and mood is unrelated to the difficulty of the tasks people choose. Alternatively, perhaps both perceived task competence and mood were influenced, but they have opposite effects on task choice (e.g., if people in good moods avoid difficult tasks to decrease the likelihood of failure that would ruin one's mood; Wegener & Petty, 1994). If so, the manipulation might have no overall effect on task choice because the effects of perceived competence and mood cancel each other out.

In the various scenarios regarding random or systematic error, the obtained null effect is not strictly "spurious" in that the IV as operationalized truly has no effect on the DV as operationalized. However, the researcher's interpretation of this null effect could be in error. That is, a researcher might conclude that there is no relation between the *conceptual variables* (constructs) of interest, when in fact it could simply be because the operationalization(s) have failed to properly represent those constructs (in this case, the IV).

A second possible problem emerging from poor construct validity is that it can lead to the emergence of "misleading" non-null effects. That is, if either operationalization is reliable but captures an unintended construct and the unintended construct happens to be related to other constructs represented in the study, an effect can emerge, but the meaning of that effect is obscured. Once again, the effect is not spurious in that the IV as operationalized truly has had an impact on the DV as operationalized. However, the interpretation of this effect is likely to be in error. A researcher might conclude that the effect has emerged because the hypothesized relation between the conceptual variables is correct, when in fact one or both of the constructs of interest have played no role in the emergence of the effect at the operational level. In reality, it could be that no relation exists between the conceptual variables (constructs) of interest.

These two potential problems resulting from poor construct validity suggest two possible explanations for why an original study might demonstrate an effect and a replication study might fail to provide evidence of the effect. First, consider a situation where there is a relation between the constructs of interest in the population. If the operationalizations in the original study have high construct validity whereas one or both operationalizations in the replication study have lower construct validity, this could lead to the emergence of an effect in the original study and a "failure" (null effect) in the replication. This could occur regardless of whether the operationalizations are the same across the original and replication study (but a change in sample, setting, or time has changed the level of construct validity) or are different across studies. Second, consider a situation where no relation exists between the constructs in the population. If either operationalization in the original study happened to reflect one or more unintended constructs related to the constructs of interest, evidence of a (misleading) non-null effect could emerge in the original study. However, if the operations in the replication study have higher construct validity than in the original study, this misleading effect would fail to emerge in the replication.[4]

Although sometimes acknowledged, the potential role of construct validity in replication failures has seldom been a central focus in the contemporary replication literature. Nonetheless, in recent years a number of commentators have noted various conceptual reasons for why changes in populations or contexts across original and replication studies could alter the psychological properties of the operationalizations (even when identical or nearly identical experimental manipulations or measures are used; for example, Fabrigar & Wegener, 2016; Finkel et al., 2017; Gilbert et al., 2016; Petty, 2018; Stroebe & Strack, 2014; Wegener & Fabrigar, 2018). As a simple example, using the same cartoons to induce good mood in two different eras can fail to have the same effect on the underlying construct because of changes in societal tastes. Indeed, the specific manipulations or measures used in original studies are often selected to provide relatively optimal operationalizations of their intended constructs in the specific population and context of the original research rather than to be operationalizations broadly applicable across settings and people. It is not uncommon for researchers to conduct pretesting of experimental manipulations prior to using them in a primary study or to employ manipulation checks within a primary study as a means of assessing the construct validity of manipulated independent variables. However, such practices are far from universal in original or replication research. Systematic reviews have indicated that many psychological researchers fail to undertake adequate construct validation efforts of the dependent measures used in their studies (Flake et al., 2017; Fried & Flake, 2018, 2019). Commentators emphasizing issues of construct validity have suggested that in the presence of such concerns, low replication rates do not necessarily indicate Type I errors in the original studies. However, skeptics of this viewpoint have noted that such arguments have been based on conceptual logic and hypothetical examples rather than actual empirical demonstrations of the role of construct validity problems in non-replication (e.g., see Zwaan et al., 2018a, p. 47).

Construct validity has also played an important role in the on-going debate regarding the appropriate roles of direct

(exact) versus conceptual replication (e.g., Crandall & Sherman, 2016; Fabrigar & Wegener, 2016; Finkel et al., 2017; LeBel et al., 2017; LeBel & Peters, 2011; Nosek et al., 2012; Pashler & Harris, 2012; Petty, 2018; Simons, 2014; Stroebe & Strack, 2014; Zwaan et al., 2018a, 2018b). Two central themes have emerged in this debate. First, there has been spirited discussion regarding the relative value of these two types of replication and the extent to which one should be emphasized at the expense of the other. Second, there has been substantial discussion regarding how distinct these two approaches truly are and what the concept of direct replication actually means in the context of psychological research.

When examining differing viewpoints on these two issues, much of the disagreement seems to originate with differences in the extent to which commentators are focused on studies in which general or more specific constructs are of key interest. For example, in theory testing research, investigators are most interested in abstract constructs like "attractiveness" or "academic performance" that can be operationalized in many different ways. In more applied research, investigators are more interested in much more specific and concrete constructs (e.g., Oprah Winfrey) or particular outcomes (e.g., grade point averages) that are closer to and sometimes nearly identical to the operations used to represent the constructs. When constructs in an original study are viewed at an abstract and general level (e.g., does "frustration" lead to "aggression?"), the fundamental goal of replication is to replicate relations between those general constructs, regardless of whether the same or very different operationalizations are used to represent them. For such researchers, conceptual replication is viewed as having potential benefits. When constructs in an original study are viewed more concretely and perhaps are nearly identical to operations (e.g., will a particular dosage of a drug enhance students' grades over the course of a semester?), it is more reasonable to assume that the best way to achieve construct validity is to use the same operationalizations. In such situations, direct (exact) replication has obvious benefits.

Researchers producing "failed" replications have sometimes briefly acknowledged that changes in the psychological properties of manipulations or measures as a function of context or population might have contributed to the failure (e.g., Ebersole et al., 2016, p. 81; Klein et al., 2018, p. 482; Open Science Collaboration, 2015, p. aaac4716-6). However, construct validity explanations have not been given as the primary reason for non-replication. Rather, replication researchers have typically assumed that the likelihood of construct validity problems should be minimized so long as the replication procedure matches the original procedure as closely as possible (i.e., a direct/exact replication; cf. Gilbert et al., 2016), an assumption that has problems, as noted earlier.

There has often been little attempt on the part of replication researchers to empirically evaluate the extent to which the operations used in direct replications map onto the concepts in original studies, even in cases where the original researchers explicitly accorded substantial attention to construct validity issues. For example, the manipulations and measures used in some original studies that were later replicated initially underwent some form of psychometric evaluation (e.g., see Petty & Cacioppo, 2016). In some cases, manipulations or measures were pretested to ensure they met certain psychometric criteria. Most commonly, the performance of experimental manipulations was evaluated with a manipulation check or the performance of measures was assessed with some form of psychometric analysis (e.g., factor analysis). Replication efforts have often not reported which original studies included such assessments of operationalizations and which did not. Likewise, when such assessments were reported in the original studies, replication reports have infrequently indicated whether psychometric assessments were also replicated either before or in conjunction with the replication study. Finally, when data relevant to assessing the performance of the manipulations or measures in replication studies have been available, replication researchers have often failed to fully report such data and consider their implications when evaluating why the replication failed to reproduce the original findings. Indeed, in the few cases where the psychometric properties of measures have been evaluated in replications, these investigations have indicated that the psychometric properties of the measures have sometimes differed across the original and replication study thereby making interpretation of non-replication difficult (see Fried & Flake, 2018).

As an illustration of some of these observations, it is worth considering the Many Labs 3 replication initiative (Ebersole et al., 2016). In this multilab effort, 10 published studies were replicated. Of the original studies, explicit mention of pretesting of operationalizations occurred for one study (Cacioppo et al., 1983). However, the report of the replication study made no mention of this pretesting procedure in the original study and whether this pretesting protocol was also replicated. In a second study (Tversky & Kahneman, 1973), the authors did not conduct pretesting but based their selection of experimental stimuli on linguistic analyses published 8 years earlier. The replication report made no mention of the basis of the original authors' choice of stimuli and did not indicate whether any attempt was made to ascertain whether more contemporary linguistic analyses could confirm that the stimuli from almost 50 years earlier were still appropriate. In three original studies (Cacioppo et al., 1983; Monin & Miller, 2001; Szymkow et al., 2013), manipulation checks of independent variables were reported. None of the replication reports noted the use of manipulation checks in the original studies nor were any results provided for these manipulation checks in the replication studies. Indeed, there was no mention made of whether the manipulation checks were included in the replication studies. Finally, in the one nonexperimental original study (De Fruyt et al., 2000), the original study used

previously validated self-report scales to assess the association between conscientiousness and persistence. In the replication study, the original 48-item measure of conscientiousness was replaced with a two-item measure and the multi-item self-report measure of persistence was replaced with a measure of the amount of time spent working on an unsolvable anagram task. In the replication report, no discussion was provided of the psychometric properties of the two-item scale relative to the original 48-item scale nor were data reported to evaluate whether performance on the anagram task would be related to scores on the original self-report measure of persistence. Thus, lack of attention to comparable construct validity across original and replication studies is a plausible cause of replication failure.

External Validity

External validity refers to the extent to which the results of an original study can be generalized to other situations and populations (Cook & Campbell, 1979).[5] When evaluating whether the results of a study can be said to generalize, this judgment must be made in the context of the goals of the study. For a purely applied study that examines rather concrete constructs closely tied to specific operations, the relevant generalization might very well concern the extent to which the relation between a specific operationalization of an IV and a specific operationalization of a DV generalizes to a new setting or population. When the goal of a study is basic theory testing, the relevant generalization question likely differs. That is, the primary focus of generalization would usually concern the degree to which the relation between the psychological constructs of interest generalizes rather than the specific operationalizations of these constructs used in the original study, though generalization would include generalizing the impact of the specific operationalizations to the extent that those operationalizations continue to represent the same constructs for the new population or setting.

Thus, in basic research where theory testing is the primary focus, it is important to distinguish between situations in which changes in the setting or participants could have produced different results because of construct versus external validity limitations. If the results of an original study fail to be replicated because the manipulations or measures no longer effectively capture their intended constructs in the new setting or with the new participants, the discrepancy between studies is most appropriately interpreted as a construct validity issue. In contrast, if the operationalizations continue to function similarly in the replication study (i.e., they appropriately represent the original constructs) but the results are different because of a difference in the relation between constructs in the new setting or with the new participants, this discrepancy is better considered as an external validity issue.

Differentiating between these two possibilities is obviously important for interpreting a replication failure and gaining a full understanding of the conceptual phenomenon of interest in theory testing research. However, it can also be of value in applied contexts. Imagine a case in which the effect of an applied intervention is found to have no effect because the intervention as operationalized originally no longer influences the focal construct presumed to influence the DV. Such a failure would suggest that the original specific version of the intervention cannot be used in the new population or setting, but it does not invalidate the broader concept of the intervention. The original study still constitutes a valid "proof of concept," and an alternative version of this intervention might well be effective to the extent that it successfully represents the original construct of interest. However, if the effect of the intervention is not replicated because of a problem with external validity, this suggests that variants of the original intervention that represent the original construct are also unlikely to be successful in the new population or setting. A fundamentally different strategy might be needed.

When considering cases in which an original study and a replication study have produced discrepant results, low external validity can provide a very straightforward explanation. Even if both the original and replication study were high in statistical conclusion validity, internal validity, and construct validity, it is possible that the two studies differ on characteristics of the participants or the context that could alter the relation between the constructs reflected by the IV and DV (i.e., one or more characteristics might moderate the effect of the IV on the DV). If so, one would not expect the effect demonstrated in the original study to generalize to the new study's population or setting.

A number of commentators on replication efforts have suggested that differences in population or context might moderate the emergence of effects and thus account for differences across original and replication studies (e.g., see Barsalou, 2016; Cesario, 2014; Dijksterhuis, 2014; Stroebe & Strack, 2014). For the most part, when discussing moderators, commentators have not distinguished between moderators that might produce their effects as a result of changes in the construct validity of operationalizations (as just described) versus those that reflect changes in the nature of the relation between constructs (external validity). Indeed, moderation can also represent a case of statistical conclusion validity. For example, a study on frustration leading to aggression initially conducted with a male population might fail to replicate in a sample of females (i.e., gender moderates the effect) because (a) the operation used to induce frustration in males fails to produce frustration in females (construct validity problem[6]), (b) although operations represent the constructs well in both populations, frustration fails to produce aggression in females (external validity problem), (c) although the inductions represent the constructs on average in both populations and frustration does produce

aggression to the same extent in both the male and female population, the female sample chosen for the replication does not represent the population of females as well as the male sample chosen represents the population of males leading to a smaller effect size in the female sample and thus requiring a larger sample to detect the effect in the new female sample (statistical validity problem).

In speculations in prior commentaries about the impact of moderators, factors identified as potential explanations for differences across studies are most often attributed to external validity. Yet, as noted above, they might sometimes be more appropriately interpreted as construct or statistical validity concerns. Importantly, commentators who have stressed the potential role of external validity have cautioned that because of such possible moderators, the fact that many replication studies have failed to reproduce the results of original studies does not necessarily indicate that Type I errors are pervasive in the published literature.

The potential role of external validity in replication has been acknowledged to some degree in multi-lab replication efforts when they examined potential moderation by lab or other factors (e.g., Alogna et al., 2014; Ebersole et al., 2016; Hagger et al., 2016; Klein et al., 2014, 2018). However, some advocates of a prominent role for replication in psychology have expressed discomfort regarding the inferential ambiguity of potential post hoc "hidden moderator" explanations for discrepancies between original and replication studies as illustrated below:

> It is practically a truism that the human behavior observed in psychological studies is contingent on the cultural and personal characteristics of the participants under study and the setting in which they are studied. The depth with which this idea is embedded in present psychological theorizing is illustrated by the appeals to "hidden moderators" as explanations of failures to replicate when there have been no empirical tests of whether such moderators are operative . . . (Klein et al., 2018, p. 482)

> 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. (Zwaan et al., 2018b, p. 6)

Our point here is not to debate the likelihood of the presence of "hidden moderators," but to note some moderators might best be interpreted as factors likely to affect construct or statistical conclusion validity, whereas others are probably more appropriately conceptualized as factors likely to regulate the relations among underlying constructs in the new setting and participants (i.e., factors affecting external validity). Thus, consideration of moderators in replication studies and the specific type of validity concern to which moderation should be attributed might have been more superficial than is desirable.

To date, when testing potential moderators in replication studies, the primary focus has been on examining what we call "generic" moderators (e.g., see Ebersole et al., 2016; Hagger et al., 2016; Klein et al., 2014, 2018). That is, the primary focus has been on moderator effects involving broad characteristics of samples and settings such as country of data collection, language of country, setting (i.e., online vs. laboratory), task order, and time of semester of data collection. Exploration of such moderator effects provides less than optimal tests of the boundary conditions for psychological phenomena because the psychological meaning of the examined moderator is not specified. That is, such moderators are largely atheoretical. For example, consider a characteristic such as time of semester (e.g., see Ebersole et al., 2016). It is difficult to clearly relate this potential moderator to a specific psychological construct (e.g., it might reflect the level of conscientiousness of participants, participant procrastination, participant busyness). Moreover, to the extent that one can relate it to a specific construct, it might best be regarded as a proxy for the construct rather than an optimal operationalization of it (e.g., a formal scale assessing conscientiousness would be better). Perhaps even more problematic, many of these generic moderators such as time of semester might well be capturing multiple constructs, some of which could be exerting contradictory effects (e.g., perhaps time of semester reflects both conscientiousness and the degree to which participants are naïve to psychological methods). Finally, the characteristics themselves have typically not been selected based on theoretical logic or prior empirical evidence but rather logistical convenience (e.g., the ease with which they can be documented or measured in multiple settings) or their presumed applied importance. Were one to carefully examine the literatures associated with any of the psychological effects being examined, it is not clear that most or any of these generic moderators would appear on a list of the most conceptually compelling moderators for any of the specific psychological phenomena being investigated. In light of these conceptual and methodological ambiguities, it is perhaps not surprising that replication efforts have reported relatively few effects of such generic moderators. And as noted above, if such moderation was found, it would be useful to know if the moderation should be attributed to issues involving external, construct, or statistical conclusion validity.

Sometimes, though more rarely, multi-lab replication efforts have examined more specific moderators of particular phenomena (e.g., measures of specific personality traits or individual differences), although even in these cases more text has been devoted to discussing results of generic moderators than specific moderators (e.g., Ebersole et al., 2016; Klein et al., 2018). Conceptually, we suspect that examining specific moderators would often be more promising. In the

case of specific moderators, the intended construct of interest is usually stated and the construct is generally examined because of past theory or research related to a specific phenomenon under investigation. In addition, such moderators usually involve measures specifically developed to assess the intended construct (or were based on a prior measure designed to assess the construct). That being said, an important limitation of most of these investigations has been that logistical convenience (i.e., ease of data collection) has strongly dictated both the measures used and the conceptual moderators selected (for an exception, see Alogna et al., 2014). Because these multi-lab replication efforts have often been restricted to sets of very short studies that can all be administered (often online) within a comparatively short time frame, moderator variables have often been streamlined versions of scales rather than the "gold standard" measures of the intended constructs (often with no accompanying psychometric evaluation of the adequacy of the streamlined measure; cf. Widaman et al., 2011). Shortened versions of scales can be less likely to produce effects than their original longer versions (e.g., Bakker & Lelkes, 2018). Similarly, only the sorts of moderators that can be easily assessed with a few brief self-report items are typically considered for inclusion. More subtle characteristics of context or characteristics of participants that require more intensive measurement protocols have generally been avoided. Finally, because the moderators being tested are almost always measured rather than manipulated, variations in the mean levels of these moderators across labs (or participants) could obviously be confounded with other factors (that could themselves have conflicting effects). Thus, once again, it is not surprising that evidence for moderator effects in the accumulated replication initiatives has been comparatively sparse.

Emphasis on Different Types of Validity in the Replication Literature

Our core argument is that when viewed from the standpoint of Cook and Campbell's typology of validity, any failure to replicate the findings of a prior study could be accounted for by one or a combination of four different sets of explanations related to the four kinds of validities. However, as already noted, there has been substantial asymmetry in attention accorded to each of these explanations. In our reading, all of the primary explanations for poor replication rates and proposed solutions for increasing replication rates (e.g., increased power, prevention of QRPs, more stringent alpha levels, use of Bayesian statistics, and replacing statistical tests with effect sizes and confidence intervals) are based on the assumption that the primary cause of non-replication is low statistical conclusion validity. Construct validity and external validity have been acknowledged as relevant to understanding non-replication, but neither has assumed a central role in replication initiatives or the literature on replication. Internal validity has for the most part been ignored as a possible explanation for non-replication.

Importantly, more formal content analyses of publications in psychology support this narrative account. As an illustration, consider the recently published books on the current crisis that were cited earlier. Of the 18 chapters in Lilienfeld and Waldman (2017), when discussing issues related to one of the four types of validity, 12 focused primarily on issues related to statistical conclusion validity. Only one chapter focused primarily on external validity, one chapter focused primarily on construct validity, and one chapter primarily discussed issues related to internal validity when discussing the four forms of validity (three chapters did not discuss any of the four types of validity). Likewise, in Chambers (2017), statistical issues were the most prominently featured form of validity in three of the eight chapters. None of the other forms of validity was the primary form of validity featured in any single remaining chapter (these chapters dealt with outright fraud and other concerns not directly related to Cook and Campbell's forms of validity). Only Hughes (2018) did not show a clear emphasis on statistical conclusion validity. Hughes (2018) primarily highlighted issues related to statistical conclusion validity in one chapter, construct validity in one chapter, and external validity in one chapter, with another chapter providing comparatively balanced treatment of these three types of validity (three other chapters discussed issues not directly related to the forms of validity).

Broader content analyses suggest similar conclusions. In one recently published paper (Fabrigar et al., 2019), a systematic content review was conducted on books, special issues of journals, and special sections of journals focusing on replication issues (a total 88 journal articles and book chapters published in eight distinct special issues/sections of journals and one book). For each publication, whether it focused at least in part on issues directly related to each of the four types of validity was examined, and, if so, if there was a predominant focus on one of the types of validity. It was found that, when discussing validity issues, 61% of publications focused primarily on statistical conclusion validity, whereas 9% focused primarily on external validity and 8% focused primarily on construct validity. Only 5% focused primarily on internal validity and another 17% did not discuss issues directly related to any of the four types of validity. More comprehensive and detailed content analyses suggested similar conclusions (Vaughan-Johnston et al., 2020).

Likewise, one major way that advocates of reform have promoted changes in psychological research is by arguing for changes in submission guidelines to journals. In response to such recommendations, a number of journals have revised their submission requirements. These changes have strongly emphasized statistical conclusion validity. To illustrate, consider the statement of principles articulated when a new editorial team assumed responsibility for *Social Psychological and Personality Science* (Vazire, 2016). Of the four aims and

their implementation highlighted by the new editor (see p. 4), three focused primarily on issues related to statistical conclusion validity. None of the other three types of validity were highlighted in *any* of the aims and/or their implementation. Once again, broader content analyses of journal guidelines and policies led to similar conclusions (Fabrigar et al., 2019). In an examination of the guidelines of seven major social–personality journals and six major marketing journals, content analysis indicated 45% of guidelines for social-personality psychology journals were directly related to statistical conclusion validity (e.g., a guideline regarding sample size and statistical power). For the remaining three validities, only about 1% to 3% of guidelines addressed one of these forms.[7] Likewise, of the guidelines examined in marketing journals, 26% related to statistical conclusion validity, whereas none of the other three types of validity were represented.

Finally, the primacy of statistical conclusion validity can be seen in published replication efforts themselves. When reporting their findings, replication researchers have routinely provided detailed justification for and empirical evidence of the adequacy of their statistical analyses. Discussion of statistical power and calculations related to statistical power are common. Likewise, full reporting of analytical plans is considered an essential practice. In contrast, as we have noted, detailed discussion of psychometric evaluations of measures and manipulations (construct validity) has been relatively rare. Likewise, we have noted that exploration of the role of external validity has occurred only sporadically and has seldom been a central objective guided by theory and methodological best practices. Indeed, a number of advocates of replication initiatives have suggested that when concerns exist regarding the construct validity or external validity of a replication study, replication researchers should not be required to provide empirical evidence in support of the validity of their methods (e.g., to document that their replication IVs and DVs have the appropriate construct validity; see Klein et al., 2018, p. 482; Wagenmakers et al., 2016, p. 924; Zwaan et al., 2018a, p. 48; Zwaan et al., 2018b, pp. 6–7). Instead, these authors suggest, perhaps the burden of proof should rest with critics of the replication initiative rather than with the authors of the replication effort.[8] In our view, the burden should fall on original and replication researchers alike. For example, original researchers can make clear what kinds of validity checks are optimal for successful replication, and replication researchers can more generally consider the multiple validities of the replication research.

Empirical Demonstrations of Validity Processes in Failures to Replicate

Thus far, our discussion of the role of the four types of validity in non-replication has largely been confined to the conceptual level. It is also useful to consider the extent to which clear empirical evidence exists demonstrating a role for each type of validity in replication efforts. To date, there have been few explicit demonstrations in psychology. Most replication initiatives have focused on simply examining whether effects can be reproduced and not on understanding *when* or *why* some of these effects have failed to emerge. Similarly, critics of replication studies have largely confined their responses to conceptual arguments and only rarely conducted empirical studies to gauge the viability of their criticisms. Thus, the empirical literature on this issue is relatively modest.

A distinction should be made here between research providing evidence for how violating the four validities can distort the research literature versus evidence of a role for the validities in producing replication failures. For example, although there is little evidence thus far for internal validity playing a role in specific cases of non-replication, there is cogent evidence for the more general point that violations of internal validity can seriously distort study findings. For example, the distorting impact of violations such as differential attrition have been clearly documented (e.g., Zhou & Fishbach, 2016). A similar point can also be made regarding the other three types of validity. The distorting effects that each type of validity can produce have been discussed extensively. It is the *application* of these distorting effects to understanding replication failures that has been comparatively rare. Nonetheless, in recent years, concrete cases have begun to emerge in which the role of different types of validity was examined, and these empirical examples suggest that the sorts of processes we just reviewed are more than simply hypothetical possibilities.

Evidence Relating Statistical Conclusion Validity and Replication Failure

Although statistical conclusion validity has long been presumed to play a central role in non-replication in general and has frequently been suggested as a potential cause of non-replication in specific cases, actual empirical evaluations of its role in specific cases are difficult to find. In some cases, replication researchers have chosen not to advance any specific explanation for why some effects have not replicated (e.g., Ebersole et al., 2016; Klein et al., 2014). In other cases (e.g., Calin-Jageman, 2018, p. 256; Eerland et al., 2016, p. 167; Klein et al., 2018, p. 482; Shanks et al., 2013, p. 7), replication researchers have acknowledged the possibility of construct validity or external validity playing a role in non-replication but then have downplayed the likelihood of these explanations and/or suggested that acceptance of such explanations might undermine the importance of the original phenomenon of interest (e.g., suggesting that the original phenomenon might be less interesting if it only emerges in highly restricted contexts). Yet, in many of these cases, replication researchers have speculated that Type I error in the original study (i.e., lack of statistical conclusion

validity) might be a very plausible explanation for the discrepancy (e.g., Calin-Jageman, 2018, p. 256; Eerland et al., 2016, pp. 166–167; Shanks et al., 2013, pp. 7–8).

However, even in these later cases, researchers have not gone beyond noting that the original study appeared to have some properties conducive to the occurrence of Type I error (e.g., a small sample size). Clearly establishing that Type I error occurred in a given case is difficult. For example, although a small sample size does increase the likelihood of obtaining an extreme estimate of effect size, there is no way to be certain in a given case that original evidence for an effect represented an extreme effect size from a distribution centered on zero. Similarly, there might be concerns that tests in a given article might reflect one or more QRPs (such as use of different covariates across studies).[9] However, establishing that such practices did in fact occur, and if so, were responsible for the original significant effect emerging is much more challenging. Thus, constructing a clear empirical case for statistical conclusion validity in a specific case of non-replication is an elusive goal.

Evidence Relating Construct Validity and External Validity to Replication Failure

Until recently, arguments for a potential role of construct validity and external validity in non-replication largely rested on conceptual logic. However, some researchers have begun to provide data in support of explanations based on construct validity and/or external validity. In some cases, these empirically based responses to non-replication have involved conducting a new replication study to demonstrate that the effect of interest can be reproduced if concerns regarding construct validity and/or external validity are appropriately addressed. In other cases, researchers have gone a step further to experimentally test whether particular construct- or external-validity-related methodological features of the replication study were responsible for the failure to reproduce the originally demonstrated effect. This still small, but emerging empirical literature has suggested that it is unwise to dismiss these two explanations for replication failures. To illustrate this point, it is useful to consider two recent and relatively clear cases in which the roles of construct and external validity in non-replication have been examined (for an interesting comparison example, see Calin-Jageman, 2018; Ottati et al., 2015, 2018).

Facial-feedback effects using the "pen-in-mouth" paradigm. As an initial illustration, it is useful to consider the case of the facial-feedback effect. In the original study, participants were instructed to hold a pen in their mouths using their teeth (activating muscles involved in smiling) or using their lips (activating muscles involved in pouting; Strack et al., 1988). Participants performed this task while rating the funniness of cartoons. Participants reported being more amused by the cartoons when "smiling" rather than "pouting." The original pen-in-mouth study obtained widespread attention, even being featured on the cover of *Science* (May 18, 2007).

Noting the prominence of this original study in the psychological literature and the fact that the study had never been directly replicated, Wagenmakers et al. (2016) conducted a multi-lab direct replication. In this initiative, 55 researchers across 17 labs ($N = 1{,}894$) followed a standardized protocol aimed to closely model the original study. They found no evidence of the effect. In considering the failure to reproduce the effect, Wagenmakers et al. (2016) were unable to identify any features of their replication study that might have accounted for the discrepancy in findings. Wagenmakers et al. did, however, note that they could not rule out that some "unexplained factor" might be responsible for the difference in results but also noted the statistically compelling evidence for a null effect and the comparatively homogeneous set of effects obtained across multiple labs.

In considering this failure to replicate, Strack (2016) suggested several possible explanations. Three involved factors related to external validity. First, he noted that because the facial-feedback effect is now commonly taught in psychology courses and the original Strack et al. (1988) study is often mentioned in particular, some participants from psychology subject pools (used in 14 of the 17 replication labs) might have been aware of what was being be tested, thereby altering the results. Strack also noted that no formal funnel interview procedure had been used to screen participants for suspicion and that the three studies not relying on psychology subject pools produced an effect more consistent with the original finding. Second, Strack (2016) questioned whether contemporary participants understood the 1980s-era cartoons used in the original and replication studies.[10] Finally, Strack (2016) noted that the Wagenmakers et al. (2016) procedure included visibly present video cameras recording participants (to ensure that they produced the requested facial expression). Although the added camera was intended to increase validity of the facial expressions in the replication study (i.e., to ensure that participants complied with the instructions), it might also have induced a subjective self-focus that is known to alter reliance on internal cues (e.g., Haas, 1984; Libby & Eibach, 2011; McIsaac & Eich, 2002; Wicklund & Duval, 1971). Because the original study did not include video recording of participants, this difference in procedure might play a role in the different effects.[11]

To date, there have been at least two published empirical tests relevant to evaluating whether the concerns highlighted by Strack (2016) might have played a role in failure to replicate the original facial-feedback effects. In one effort, Marsh et al. (2019) conducted a study to determine whether the original effect could be reproduced using procedures consistent with the recommendations of Strack (2016). Moderately funny contemporary cartoons were used, and the study did not have a video camera present. Participants were psychology students tested 2 weeks prior to coverage of the facial-feedback effect in their introductory course. This study

produced a significant facial-feedback effect with effect sizes that were not significantly weaker than those of the original study. However, this study did not include conditions that aimed to replicate the failure of Wagenmakers et al. (2016) such as testing some students after coverage of the facial-feedback effect or with a video camera, so it is not clear whether the identified moderators influenced whether the effect emerged.

In a more direct empirical response to the Wagenmakers et al.'s (2016) failed replication, Noah et al. (2018) directly evaluated the video camera explanation. This study used a revised set of cartoons pretested to ensure that they were moderately funny. The study was conducted using participants who were not psychology students and thus presumably unlikely to know about the facial-feedback effect. Of most central interest, however, Noah et al. (2018) randomly assigned participants to a version of the study in which a video camera was present or a version in which no video camera was present. Their results replicated the original finding when no camera was present, but failed to produce evidence of an effect when the camera was present. Thus, they argued that the presence of a camera played an important role in accounting for the discrepancy between the original and replication studies. The Noah et al. (2018) study constitutes only a single empirical test of the camera hypothesis and thus conclusions solely based on it should be treated with caution. In addition, though it does suggest that the presence of a camera might be sufficient to eliminate the original effect, it does not directly speak to the extent to which participants' awareness of the facial-feedback effect and the specific cartoons used in the replication might also have contributed to the failure to find an effect in the Wagenmakers et al. (2016) replication. Nonetheless, it is a good example of going beyond speculation about a potential moderator and actually testing its importance.

Finally, evidence potentially relevant to evaluating the role of the presence of a camera in the emergence of the facial-feedback effect was recently reported in a comprehensive meta-analysis of facial-feedback literature conducted by Coles et al. (2019). Overall, they found evidence of a facial-feedback effect. Importantly, a meta-analytic comparison of studies that included visibly present camera recording equipment versus those that did not failed to produce evidence of moderation ($p = .36$). That being said, it is not clear that this comparison clearly refutes the camera explanation. Such meta-analytic comparisons do not involve random assignment and thus unknown confounds could exist in the comparison. Furthermore, the analysis involved studies using a variety of manipulations of facial movements and thus it does not directly speak to the effect of the pen-in-mouth procedure used by Strack et al. (1988). It could be that different manipulations of facial movements are not all equally affected by the same moderators. It is interesting to note that the meta-analytic estimate of the facial-feedback effect in the two conditions fell in the same direction as in Noah et al. (2018; camera visibly present: $d = .17$, $p = .003$; no camera visibly present: $d = .23$, $p = .0000007$).[12] Moreover, the primary discrepancy between this meta-analysis and the Noah et al. experiment was that a significant effect still emerged in the camera present studies in the meta-analysis, whereas the effect was weak and nonsignificant in the camera condition of Noah et al. study. This result is contrary to any claim that the facial-feedback effect is illusory.

Taken as a whole, what can be concluded from the discrepancy in findings between Strack et al. (1988) and Wagenmakers et al. (2016) and the external validity explanations advanced by Strack (2016)? First, there does not appear to be a compelling case that the specific 1980s-era cartoons used in Wagenmakers et al. (2016) are problematic. These cartoons were pretested by Wagenmakers et al. and confirmed to be moderately funny and nothing about their specific content is obviously unique to the 1980s (see Wagenmakers & Gronau, 2018). However, it is not clear that the potential roles of the visibly present camera or participants' prior knowledge of the facial-feedback effect can be casually dismissed. Noah et al. (2018) constitutes a direct experimental test of the camera hypothesis and the results are suggestive, though it is true that the test of the difference in magnitude of the facial-feedback effect across conditions did not reach significance ($p = .051$ one-tailed; see Wagenmakers & Gronau, 2018). Nonetheless, the statistical case for the emergence of an effect in the Noah et al. no-camera condition is "substantial" using typical Bayesian labels (see Schul, Noah, & Mayo, 2018). This finding is further supported by clear emergence of the effect in the comparatively well-powered ($N > 400$; Marsh et al., 2019) replication that did not include a camera and tested the effect in a population unaware of facial-feedback research.

In summary, the original Strack et al. (1988) study, the no-camera condition of Noah et al. (2018), and the Marsh et al. (2019) study have all provided evidence of the pen-in-mouth manipulation producing a facial-feedback effect. The Coles et al. meta-analysis provided more broad support for the facial-feedback hypotheses regardless of camera presence. In contrast, the Wagenmakers et al. (2016) replication effort did not find any effect. The two most obvious methodological differences between the first three studies and the Wagenmakers et al. study are that the first three studies did not have a visibly present camera and were conducted in populations unlikely to have any prior knowledge of facial-feedback research. Thus, it is not clear that one can build a compelling case for a statistical conclusion validity explanation for the discrepancy between Wagenmakers et al. (2016) and the original Strack et al. (1988) study. Rather, one or more factors related to external validity appear more plausible. The Wagenmakers et al. (2016) replication effort tested the effect of the IV in a context, and possibly in a population, to which the original effect might not be expected to generalize.

Importantly, the factors most likely responsible for nonreplication could not have been readily detected in the Wagenmakers et al. (2016) data because these factors were either held constant across the replication data sets (i.e., the presence of the camera) or only varied modestly across labs (i.e., the use of participants from psychology subject pools vs. other sources). In addition, it is not clear that these factors can be regarded as "hidden" moderators. Suggesting that studies be conducted in participant populations unaware of the hypotheses being investigated is certainly not a novel methodological consideration (e.g., Kruglanski, 1975; Weber & Cook, 1972). Likewise, the effect of subjective self-focus on reliance on internal cues is not a novel idea (e.g., Haas, 1984; Libby & Eibach, 2011; McIsaac & Eich, 2002; Wicklund & Duval, 1971), nor is the use of video cameras to create changes in self-focus an innovative methodological development (e.g., Haas, 1984; Insko et al., 1973; Scheier & Carver, 1980; Vallacher, 1978; Wicklund & Duval, 1971). Therefore, although the potential motives behind using original materials or introducing ways to ensure that a protocol is effectively followed are certainly understandable, there may be times when such attempts inadvertently affect other variables that also play key roles in the effects of interest. All that is left to completely close this circle is for the Wagenmakers et al. group or some other independent lab to conduct a study in which they too evaluate whether the presence of a camera, and perhaps awareness of facial-feedback effects are key factors in moderating the facial-feedback effect.

Need for cognition as a moderator of argument quality effects on evaluation. Another useful illustration of an empirical exploration of construct validity and external validity in nonreplication was in response to the "Many Labs 3" project discussed earlier (Ebersole et al., 2016). In the replication effort, research teams from 20 labs attempted to replicate 13 previously reported "high interest value" effects from 10 studies. One of the effects was showing a relation between individual differences in Need for Cognition (NC) and the extent of processing of persuasive messages (Cacioppo et al., 1983). Cacioppo et al. (1983) found that participants' dispositional motivation to engage in effortful cognitive activity (as measured by the NC scale) interacted with the quality of persuasive arguments provided to influence evaluations of the message advocacy. Participants' evaluations were more influenced by the quality of message arguments (strong vs. weak) when they were relatively high rather than low in NC. However, in the replication based on data collected from 20 different sites ($N = 2{,}696$) and one online sample ($N = 737$), Ebersole et al. (2016) found no evidence for a NC $\times$ Argument Quality interaction. This was surprising because this particular interaction had been replicated several times previously and was supported in meta-analytic assessments of the prior studies available (Carpenter, 2015). Ebersole et al. (2016) were unable to generate an explanation for the discrepancy between their findings and the original study, concluding that they did "not have an explanation for why no effect was observed under these circumstances" (p. 81).

In considering this failure to replicate, Petty and Cacioppo (2016) highlighted several methodological features of the replication that might have substantially eroded the construct validity of the independent variables and introduced contextual changes that would be likely to inhibit the emergence of the original effect. First, Ebersole et al. (2016) used a six-item measure of NC whose psychometric properties had never been fully evaluated, rather than the previously validated 34-item scale (Cacioppo & Petty, 1982) used in the original study or a shorter validated 18-item scale used in many other demonstrations of NC effects (Cacioppo et al., 1984). Second, the original study equated pre-message attitudes across high- and low-NC groups by recruiting pairs of participants who differed in NC but not in pre-message attitudes toward the message topic. The replication study did not control for any differences in initial attitudes thereby potentially confounding NC scores with pre-message attitudes. Both of these differences between the original research and the replication study might have decreased the construct validity of the NC measure in the replication.

Another important difference between the original and replication study was that the replication researchers chose to model their strong and weak messages on more moderate (i.e., less strong and less weak) versions of these messages than in the original study. Also, the messages used in the replication study were only about half the length of those in the original study. Shorter message length suggests that the replication messages contained fewer or less fully articulated arguments than in the original study. Both changes likely decreased the construct validity of the argument quality (AQ) manipulation by creating less of a difference in the quality of arguments between the strong and weak messages. If so, this could limit the upper boundary of the magnitude of the AQ effect that could be obtained by high-NC people in the study, thereby reducing the effect size that could be obtained for the NC $\times$ AQ interaction, even if the AQ manipulation would have little or no effect for people relatively low in NC.[13]

The only deviation from the original study reported in Ebersole et al. (2016) was the use of a shorter NC scale. Deviations with respect to the AQ induction and the failure to unconfound NC scores and pre-message attitudes were not noted. In addition to choosing versions of the messages that likely weakened the AQ manipulation, this problem was further compounded by the fact that the replication research did not include the pretesting protocol of the original study. That is, the original study used pretesting to guide the construction of the strong and weak messages (see Petty & Cacioppo, 1986, for additional discussion). In contrast, the replication study arguments were not pretested before the main replication study. Indeed, some data from the replication study suggested that the AQ manipulation was much

weaker than in the original study (see Petty & Cacioppo, 2016).

Yet another important undocumented difference between the replication study and the original study was the instructions that created "baseline" levels of processing motivation. In the replication, participants were explicitly told that the policy advocated in the message was proposed to take place "immediately." No mention of immediate implementation occurred in the original study. This change raises a potential external validity explanation for the difference in results. Specifically, immediate implementation of a policy is often used as a method of inducing high involvement in an issue, which is known to enhance motivation to carefully process persuasive messages (Petty & Cacioppo, 1979, 1990). Existing theory and prior empirical research indicates that variation in dispositional motivation to engage in effortful cognitive activity (as assessed by the NC scale) is most likely to have an effect when there are no clear situational factors present to motivate people to engage in careful processing of a persuasive message (e.g., Calanchini et al., 2016; Priester & Petty, 1995; S. M. Smith & Petty, 1996; Wheeler et al., 2005). By introducing such a factor, the replication study created a context likely to inhibit the emergence of the NC × AQ effect. In contrast, in the original study and most other studies demonstrating NC effects, either situational factors are left ambiguous or they are constructed to create low situational motivation to process the message (e.g., presenting instructions likely to create low involvement) so that high levels of NC have sufficient room to increase processing over the low baseline level of motivation and low-NC individuals are not already motivated to process (e.g., by high personal relevance).

It is difficult to know why these modifications were introduced, though using a shorter NC scale and shorter messages would reduce the time it takes to complete the study, an advantage when attempting to replicate many studies at one time. Yet, the cumulative impact of these changes from the original procedure would be to work against replicating the original results (see Cacioppo et al., 1996; Petty et al., 2009). Luttrell et al. (2017) then conducted what may be the first published study to empirically document that a failed replication could be replicated along with the original study results. That is, these authors directly evaluated the cumulative impact of the changes to the original study introduced by Ebersole et al. (2016) by conducting an experiment in which they randomly assigned participants to receive either the Ebersole et al. experimental materials or materials designed to produce relatively optimal conditions for producing the effect (i.e., a validated NC scale, statistical control for pre-message attitudes, a stronger AQ manipulation using arguments pretested in prior research, and a lower-relevance context with longer messages). Luttrell et al. (2017) successfully replicated the Cacioppo et al. (1983) NC × AQ interaction when using the optimal procedure and also replicated the failure to find an effect when using the Ebersole et al. (2016) protocol. Then in another replication literature first, these findings were further supported by an independent multi-lab replication of Luttrell et al. (2017) conducted by Ebersole et al. (2017). They too failed to find a significant NC × AQ interaction effect with the Ebersole et al. (2016) protocol (replicating their and Luttrell et al.'s replication failure with those materials), but they did obtain a significant interaction effect with the Luttrell et al. (2017) protocol.

In summary, the discrepancy in results between the original Cacioppo et al. (1983) study and the Ebersole et al. (2016) replication study does not appear to have reflected a problem in the statistical conclusion validity of the original study. Rather, the decisive factors appear to be a combination of differences between the studies related to construct validity and external validity, all of which could have been identified a priori as potentially problematic on the basis of the existing NC literature.[14] The replication study failed to produce the expected effect because it used operationalizations of both IVs that were comparatively lower in construct validity and tested the impact of these IVs in a context to which the original effect would not be expected to generalize (i.e., a context in which all participants were motivated to think because of the high involvement instructions used). As with the facial-feedback replication discussed earlier, these factors shown to be playing a role in suppressing the emergence of the effect were not possible to detect in the Many Labs 3 data because those factors were held constant in those studies, though they were notably different from the original study and other demonstrations of NC effects (and in ways that previous data suggest would decrease the chance of replication).

This empirical examination of potential reasons for non-replication is notable for its concrete empirical evidence regarding how to obtain and not obtain the original effect and the fact that the moderation uncovered was supported in a "Many Labs" replication by the investigators who initially failed to replicate the original effect. Ultimately, it is not entirely clear which of the six changes introduced in the replication effort were responsible for the failure to replicate. The use of a shorter NC scale was explicitly justified by referring to a need to reduce the completion time for the study, but it is not clear why the other changes were introduced.[15] However, perhaps the most important question prompted by this case regards the extent to which this example is unique in replication attempts. Is this particular replication study a rare outlier, or a comparatively typical representative of the larger replication literature, or does it fall somewhere in between? It is impossible to answer this question because there are so few studies in which critics of replication efforts attempt to empirically validate their speculations, and to date there is just one example in which the replicators attempted to validate the insights of the critics of their replication effort (Ebersole et al., 2017). However, knowing whether critics of replication efforts are correct in their speculations about why a replication effort failed or not

could provide valuable insights into the broader implications of disappointing replication rates.

Potential Justifications for Emphasizing Statistical Conclusion Validity

As the previous empirical examples illustrate, it seems unwise to dismiss external validity and construct validity as potential explanations for non-replication. There are clearly cases where the available empirical evidence suggests that one or both played a major role. Most of these empirical efforts were not designed to isolate construct or external validity causes. Yet, the evidence in these examples might be more direct and concrete than evidence for statistical conclusion validity as a cause of any specific replication failure. Even so, far more emphasis has been placed on statistical conclusion validity than any of the other types of validity in the replication literature. Is there a compelling rationale for this asymmetry? There are three possible lines of reasoning that might be adopted to build a case for the emphasis on statistical conclusion validity. However, it is not clear that any of the approaches can at present provide an adequate foundation upon which to justify the existing asymmetry.

Conceptual justification. One approach might be to argue for the primacy of statistical conclusion validity on the basis of some theoretical rationale. However, it is difficult to see what that rationale might be. As previously noted, there is a clear logic for how each type of validity could lead to a discrepancy between an original study and subsequent replication. In the methodological literature, all four types of validity have been considered important dimensions for evaluating research, and no clear consensus exists indicating that threats to one type of validity is more prevalent or severe than the others (though in randomized laboratory experiments, internal validity is likely the least prevalent problem). Prevalence or severity might also relate to an empirical argument (to be addressed shortly), but there would seem to be little conceptual basis for arguing that a deficit in statistical conclusion validity with high levels of construct, internal, and external validity would be any more detrimental to replication efforts than deficits in construct, internal, or external validity accompanied by high levels of the other three validities.

As noted previously, even individuals strongly emphasizing statistical conclusion validity in non-replication have often acknowledged, at least in passing, that other types of validity could also play a role. Yet, the emphasis on statistical issues seems to have been implicitly adopted in the absence of a clearly articulated rationale for this emphasis. We are not arguing that anyone has explicitly made the case that statistical conclusion validity is more central to non-replication than the other validities, but the substantially greater attention given to statistical issues portrays such an approach. If a conceptual basis does exist, it has yet to be articulated explicitly.

Empirical justification. A second approach might be to emphasize statistical issues on empirical grounds. For example, it has been noted that many original psychology studies are underpowered (e.g., Button & Munafò, 2017) and published if results are significant (but not if nonsignificant; Greenwald, 1975). Because significant but underpowered studies are more likely than well-powered studies to reflect extreme estimates of the true effect size, original studies with low power might also be more likely to reflect Type I errors (at least given certain assumptions). If so, one would expect low-powered original studies to replicate less often than high-powered original studies. In some multi-lab replication initiatives, original studies with properties that might increase power (e.g., larger effect sizes, smaller standard errors, and narrower confidence intervals) have replicated at a higher rate than original studies with fewer such properties (e.g., Open Science Collaboration, 2015).

However, such comparisons are not as straightforward as they might seem. Many power-related characteristics are likely confounded with other research features. For instance, when judges evaluated the extent to which the effect in an original study from Open Science Collaboration (2015) was likely to be contextually bound (ratings that could reflect external or construct validity issues), these ratings predicted replication success, whereas the predictive efficacy of power-related properties largely disappeared (Van Bavel et al., 2016a, 2016b). Power-related properties might also be related to design features that differ across subdiscipline, which is also confounded with judged contextual boundedness (see Inbar, 2016).

Alternatively, some might invoke mathematical analyses linking lack of power to the prevalence of false positives in the published literature (where a large presence of false positives could produce high proportions of study results that do not replicate; for example, Pashler & Harris, 2012). However, the soundness of the assumptions underlying the conclusions based on these analyses have been questioned (e.g., see Fabrigar et al., 2019; Stroebe, 2016). For example, analyses suggesting that false positives are pervasive in psychology are highly dependent on assuming very high likelihoods that researchers routinely examine hypotheses that are not correct. Conceptual or empirical justifications for assuming such low prior probabilities in psychology have been lacking. Similarly, these analyses assume that researchers commonly report only a single empirical demonstration of a given effect, when in fact many phenomena of interest in psychology have been tested across multiple studies in a single paper. Thus, it is unclear how informative these analyses are for psychological research.

Even if one accepts either of these empirically based lines of reasoning as valid, these lines of reasoning do not speak directly to whether the asymmetry in emphasis is justified.

The belief might indicate that statistical conclusion validity concerns do contribute to non-replication. However, the question is not whether statistical conclusion validity plays a role, as it surely does. The operative question is whether statistical conclusion validity plays a *greater* role than the other types of validity or a role commensurate with the lion's share of attention it gets (to the point of emphasizing it in attempts to "increase replicability"). To date, no comparative exploration of the relative impact of these types of validities on non-replication has been conducted. Thus, no empirical or mathematical case for considering statistical conclusion validity as the strongest (or dominant) contributing factor to failures to replicate has been presented.

Practical justification. Finally, one might attempt to justify an emphasis on statistical issues on practical grounds. Perhaps researchers should focus on statistical conclusion validity concerns because solving such problems is comparatively easy, whereas grappling with problems related to the other three forms of validity is inherently more complex and difficult. We believe there are at least two objections to this line of reasoning.

First, statistical conclusion validity is not especially simple. It is more multidimensional and conceptually rich than is often appreciated. Most notably, addressing statistical conclusion validity should be about much more than just enhancing statistical power. It reflects a variety of considerations including but not limited to the appropriateness of the statistical model being fit to the data, characteristics of the estimation procedure used to calculate model parameters, and the interplay of these factors with properties of the data (which, as discussed shortly, are closely intertwined with the other three forms of validity). Moreover, statistical power itself is more complex than is sometimes acknowledged and goes well beyond simply enhancing the sample size of a given study (e.g., see Kenny & Judd, 2019; Maxwell et al., 2015; Pek & Park, 2019). Indeed, the underlying complexity of statistical conclusion validity may be apparent in the wide range of different potential solutions that have been advanced.

As noted earlier, though based on a common underlying assumption that statistical conclusion validity concerns are central to non-replication, the recommendations offered to enhance replicability are not necessarily consistent with one another. For example, arguing in favor of a more stringent alpha level in statistical tests (e.g., Benjamin et al., 2017; Greenwald et al., 1996) does not really fit with the viewpoint that null hypothesis testing should be abandoned in favor of reporting effect sizes and confidence intervals (e.g., Cumming, 2014; Schmidt, 1996). The complexity of statistical conclusion validity concerns is also reflected in the criticisms that have been raised in response to many of the recommendations to enhance replicability (e.g., see Fabrigar & Wegener, 2016; Fabrigar et al., 2019; Fiedler et al., 2012; Fiedler & Schwarz, 2016; Finkel et al., 2015; Stroebe, 2016). If one were to look for an overarching theme to these criticisms, it would probably be that explanations and solutions based on statistical issues are guilty of oversimplification. Many of these criticisms arise from concerns that advocates of statistics-centric viewpoints have failed to take into account the multidimensional nature of statistical conclusion validity or failed to recognize that the conclusions they have reached do not generalize as broadly as they believe (e.g., questioning extrapolation of conclusions regarding statistical conclusion validity developed in the context of a single study to the context of multiple studies testing the effect of interest).

A second potential objection regarding the practical justification for emphasizing statistical conclusion validity is that, although dealing with threats to internal validity, construct validity, and external validity can be challenging, researchers are not without resources in this regard (e.g., Cook & Campbell, 1979; Crano et al., 2015; Kenny, 2019; Reis & Judd, 2014; Shadish et al., 2002). For instance, an extensive methodological literature has developed regarding methods for detecting and minimizing threats to internal validity (e.g., Brewer & Crano, 2014; E. R. Smith, 2014; West et al., 2014). Likewise, there is a vast literature on developing and validating measures and experimental manipulations of psychological constructs (e.g., Brewer & Crano, 2014; Fabrigar & Wegener, 2014; O. P. John & Benet-Martinez, 2014; E. R. Smith, 2014; Widaman & Grimm, 2014). Thus, researchers are far from helpless in addressing concerns related to internal validity and construct validity in replication efforts.

Indeed, even external validity, a concern that some have suggested has been neglected in psychological research (e.g., Henrich et al., 2010; Sears, 1986), has been more central to psychological research than is often recognized. Specifically, researchers in many areas of psychology have made the testing of moderator effects a central focus of their work (indeed, many theoretical models are primarily about moderators). That is, in many areas of psychology, there has been interest in whether the effects of a given IV are conditional on levels of a second (and sometimes third) IV (Judd et al., 2014; E. R. Smith, 2014). For example, following the initial introduction of many theories or effects in social psychology, the next wave of research has focused on the conditions under which that theory or effect holds. Perhaps the most well-known example in social psychology is the theory of cognitive dissonance (Festinger, 1957) where literally decades of research has focused on a search for moderators (J. Cooper, 2007; Harmon-Jones, 2019). The same is true for most of the prominent theories in the field (e.g., see Van Lange et al., 2012), consistent with McGuire's (1983) contextualist view of theory development. Although not always framed as such, the investigation of moderators is inherently an exploration of the range of external validity of a given phenomenon, though moderators can also clarify understanding of the conceptual variables involved (construct

validity; Judd et al., 2014; Spencer et al., 2005). By more completely specifying when and for whom particular relations exist among constructs of interest, a moderation-based theory provides greater understanding of the phenomenon than a main-effect theory that does not anticipate limits to those effects.

The specification of boundary conditions is a key part of what many psychological theories are designed to address, and documentation of when a given effect is enhanced, attenuated, or even reversed is very much a part of the empirical literature in social and personality psychology (as well as many other areas of psychology). Hence, when considering why a well-documented phenomenon has not been replicated, existing theory and empirical research will often provide valuable guidance regarding external validity explanations. The examples regarding the NC × AQ interaction and the pen-in-mouth facial-feedback effects provide illustrations of this point. In both cases, the existing literature provided a basis for generating plausible external validity (as well as construct validity) explanations. When a given effect has not been extensively explored, the literature might provide less guidance to conduct a fully informed replication effort (Luttrell et al., 2017). However, the process of specifying and testing potential boundary conditions in emerging literatures is very much a part of what psychologists have done and should do.

Implications and Conclusion

In the previous sections, we have suggested that issues surrounding replication can be profitably organized within the framework of Cook and Campbell's (1979) classic validity typology. We have illustrated at the conceptual level, and in some cases at the empirical level, how any failure to replicate a previously reported finding can be a function of study differences with respect to one or more of these four types of validity. In the final section of this article, we highlight some of the broader implications suggested by this framework.

Consequences of the Validity Asymmetry in the Replication Literature

Despite the fact that there are good conceptual and empirical reasons to think that there are four distinct sets of explanations that can account for failures to replicate, the focus in psychology has seemed to be overly narrow with an emphasis on one of these categories over the others. A natural question that arises out of this observation concerns the implications of this asymmetry. Before addressing this, it is important to note that none of our commentary on this issue should be construed to suggest that statistical conclusion validity is not important. It is a *very* important property of research in the Cook and Campbell (1979) typology and there is every reason to think that it does play a critical role in some failures to replicate (most directly in those where an original single study or a replication attempt is underpowered or made use of problematic data analytic practices, but likely in others as well). Thus, the question is not whether statistical conclusion validity is important to consider, but whether an emphasis on it with a relative lack of attention to the other forms of validity is problematic.

One possible response to this question might be that any negative consequences are comparatively modest. Addressing any problem should only help. Solutions being offered for the replication crisis might be incomplete and will not fully address replication problems, but effectively addressing even one source of non-replication represents progress. Unfortunately, such an assessment could be overly optimistic. Even setting aside controversy regarding some of the statistically oriented recommendations, such a view rests on the assumption that the four types of validity are largely orthogonal to one another. That is, it presumes that placing a strong emphasis on one form validity has no effect (or at least no negative effects) on how researchers address the other types of validity. This assumption is likely in error for at least two reasons articulated next.

Distorting methodological decisions and interpretations of findings. If researchers assume that the dominant reason for non-replication is statistical conclusion validity, this could introduce "conceptual blinders" when designing original and replication studies. For example, if researchers believe that other forms of validity play only modest roles in non-replication, they might become overly focused on sample sizes and less careful in ensuring that operationalizations of IVs and DVs meet appropriate psychometric standards (e.g., failing to conduct or replicate pretesting of IVs and DVs, failing to fully report and evaluate results relevant to construct validity such as manipulation checks or factor analyses). Likewise, they might invest less effort ensuring that they have constructed a background context and identified a participant population that create the appropriate conditions for the effect of interest to emerge.

Exactly these sorts of factors might have played a role in the need for cognition and, to some extent, the facial-feedback examples discussed earlier. Even when there might have been reasons to introduce a design change in a replication (e.g., using a shorter scale to save time; introducing a camera to ensure manipulation fidelity), the potential for those changes to influence construct, external, or even internal validity might not be considered if one believes that such validities play only modest roles in replication. This might account for why the exploration of moderators has largely been an ancillary concern in replication efforts to date and why there has been little emphasis on testing potential explanations for non-replication in multi-lab and other replication efforts.

Blinders to the full set of validity concerns might also exert a substantial influence on how findings are interpreted.

If one assumes that statistical conclusion validity is a dominant cause of non-replication, it will be difficult for researchers, even post hoc, to generate alternative explanations to this account. Consider the facial-feedback non-replication described earlier in which none of the 55 researchers involved was able to identify even one methodological feature that might plausibly have accounted for the discrepancy between the non-replication procedure and those of the original study. However, at least two plausible features were identified by other scholars that could be extracted from the methodological and theoretical literatures, and subsequent empirical research suggested that they might well be viable candidates for explaining the failure to replicate. Likewise, in the need for cognition replication failure, none of the 64 replication researchers was able to generate a single explanation for non-replication related to the operationalizations of the IVs and DV or the context in which their effects were tested. However, the existing literature in this area provided a basis for other scholars to identify at least six plausible explanations related to construct and external validity, and subsequent research suggested that some combination of these factors did indeed play a role in the failure to replicate.

Perhaps an even more extreme version of the "conceptual blinder" effect is that if one assumes a dominant role of statistical conclusion validity, one might believe that high statistical conclusion validity can in some way "immunize" one's study against threats to other validities. For example, researchers might be inclined to think that even if their original research efforts or replication attempts use less optimal operationalizations of the constructs of interest or test this effect in a less optimal context than theory would suggest, perhaps these limitations can be offset using a larger sample size. Unfortunately, as Cook and Campbell (1979) noted throughout their discussion and as many other methodologists have explained, such a view is more appealing than sound. Being high on one type of validity generally affords at best modest and sometimes no protection against threats to another type of validity. Consider a situation in which the operationalizations of the IVs and DVs in a replication attempt represent the same constructs as the original study, but with more random error. Enhancing sample size might help. However, the distortions introduced by even modest increases in random measurement error can often outpace the benefits of even substantial increases in sample size (e.g., see Stanley & Spence, 2014). The distorting effects of random error can become even more pronounced and difficult to anticipate when examining complex models such as models with moderator effects and mediational patterns (e.g., see Bollen, 1989; Judd et al., 2014). Thus, if operationalizations are relatively poor, there might be no realistically achievable sample size sufficient to offset this problem.

Increased sample size will afford virtually no protection against other threats to validity. For instance, if construct validity is eroded because key operationalizations reflect somewhat or very different constructs than intended, increasing sample size might have little to no benefit. Systematic error in measurement will not necessarily result in a lower reliability coefficient (and thus might be difficult to detect). Yet, such error in measurement (or in manipulation) can nonetheless erode effects if the unintended constructs are unrelated to the phenomenon of interest or can completely eliminate or reverse effects if the unintended constructs have opposing effects to the intended construct. Likewise, if the phenomenon is examined in a context or population where it should not emerge (low external validity), no increase in sample size will alter this fact. Similarly, if a key IV is associated with a threat to internal validity (e.g., differential attrition), simply running a study with a very large sample will not solve this problem.[16]

Strategies for maximizing one validity can undermine another. Cook and Campbell (1979) also noted a second more direct problem with assuming that different types of validity are orthogonal. There can be trade-offs in maximizing each type of validity. That is, introducing methodological features that enhance one validity can sometimes reduce another validity. Thus, adopting research strategies focused overwhelmingly on statistical conclusion validity might result in an erosion of other validities. Should one or more of these other validities play an important role in replication, the result of attempts to focus on statistical conclusion validity could bring little change or even a decrease in replication rates.

Consider the seemingly straightforward and generally sensible suggestion that conducting original studies with higher power should make them more replicable. Accordingly, many journals have implemented guidelines requiring that researchers address their choice of sample size and related issues of power. As a result, simple power calculations that are largely a function of study sample size have become an easy criterion for initially evaluating the merits of a given study. Indeed, in some cases, failure to meet this criterion has been seen as sufficient to "triage" a manuscript (i.e., reject a manuscript without reviews). The goal of such a guideline is to increase the power of published research and thereby enhance statistical conclusion validity. All else being equal, increased power is, of course, a desirable objective. However, all else is seldom equal and such policies can have unintended consequences.

For example, the most common strategy to argue for adequate power has been to increase sample sizes, sometimes to levels that go well beyond what is feasible using traditional data collection practices (e.g., psychology undergraduates, community samples). Thus, data collection for original research in some areas of psychology has increasingly shifted to comparatively inexpensive online platforms like Amazon's Mechanical Turk. Such a shift is not inherently problematic. These platforms have certain advantages and they no doubt play a useful role in psychological research. However, they can also introduce potential problems (Anderson et al., 2019). For example, although the samples

provided by such platforms are in some respects more diverse than those of traditional department participant pools and this can be an asset, such heterogeneity can also have drawbacks. Sample diversity increases the likelihood that the psychological meaning of a given experimental manipulation or measure might not be the same or as uniform across participants, thereby violating assumptions of psychometric invariance (Fabrigar & Wegener, 2016; Finkel et al., 2017). Thus, in the context of original studies conducted with such samples, a researcher might be forced to use an operationalization that is not particularly optimal for any given subgroup comprising the sample, but that can function modestly and comparably well across different subgroups. This might result in worse construct validity. Of course, using a weaker or more diverse operationalization can also compromise statistical conclusion validity by reducing power to detect an effect and thus (ironically) requiring more participants in the MTurk sample than would have been required in the more homogeneous student sample.

Similarly, because of the reluctance of participants in such platforms to undertake lengthy experimental protocols and because of the economics of paying participants in such settings, time is often at a premium. Thus, measures and manipulations are often streamlined to make them time efficient, but such changes also increase the likelihood of construct validity being eroded. In addition, collecting data in online environments does not permit the same level of control as a laboratory setting. Thus, the impact of distractions and other extraneous factors is likely to be greater, thereby attenuating the efficacy of experimental manipulations and measures (also decreasing construct validity and requiring more participants to be sufficiently powered).

Online environments also impose limitations on the kinds of research paradigms that can be used. This will obviously preclude the use of some operationalizations of IVs and DVs, thereby further limiting construct validity. It will also present restrictions on the sorts of contexts in which certain phenomena can be tested, thus potentially decreasing external validity. Finally, as previously noted, online environments involve contexts in which there will be increased likelihood of some threats to internal validity such as differential attrition across experimental conditions. In summary, a strategy that places a premium on sample size might improve original research in some ways that could make it more replicable. However, these gains might be offset by eroding other types of validity that diminish the value of original studies and could make them even more difficult to replicate.

The sorts of factors we have identified as relevant to construct validity, external validity, and internal validity also have implications for statistical conclusion validity. Power is a function of both sample size and effect size (which is a function of both observed differences across conditions or across a predictor variable and of "error" variance within conditions or across participants falling at the same level of a given predictor). Although there has been a great deal of attention given to increasing sample size to enhance statistical power, as we have already noted, some of these efforts might also decrease effect sizes (either by reducing between-condition differences or by increasing within-condition "error" variance). Thus, using precise measures and manipulations of high construct validity in a homogeneous sample not only has scientific value in its own right but also has the added benefit of likely increasing power because of an increase in the effect size. Similarly, carefully constructing contexts that are conducive to the emergence of an effect (for both construct validity and external validity reasons) is valuable in its own right but is also likely to enhance power via increases in the obtained effect sizes. Even increased internal validity can sometimes enhance effect sizes if the violations of random assignment that are eliminated are exerting contradictory effects to those exerted by the IV.

We are not suggesting that large sample sizes in original studies or replication studies are undesirable or that conducting research in online environments should be discouraged. There are obvious methodological benefits to large samples, and online data collection platforms provide valuable opportunities for researchers. However, each such decision also has costs. Thus, when evaluating psychological research, it is not clear that sample size (or power) should be given special status to the exclusion of other criteria. If it is given such status, the incentives created for maximizing it will likely lead to choices that compromise other criteria (see also Finkel et al., 2017). Thus, when considering an original or replication study, it is certainly reasonable to start with a consideration of statistical conclusion validity. However, that consideration should then be accompanied by consideration of the other types of validity. Underemphasizing the other three types of validity in original or replication studies is likely to result in conclusions that can be every bit as misleading as underemphasizing statistical conclusion validity. Ultimately, the evidentiary value of any study is a function of more than simply its power, and the power of a given study is a function of more than its sample size. Guidelines that fail to take into account both of these points are an insufficient remedy to improve psychological research and might even harm it.

Improving Replication Rates in Psychological Research

The desire to produce "more replicable" psychological research has been an abiding and appropriate concern in recent years. Beyond suggesting that advocated reforms have been too narrow in their focus on statistical conclusion validity, the present framework also implies that reforms have perhaps been too narrow in another way. "Replicability" is not an exclusive property of the original study; it is a joint function of features of both the original and the replication study. Errors in either can lower replication rates. Thus, improvement in replication rates might well require reforms

in how both original studies *and* their replications are conducted and reported.

Recall that one key assumption that has distinguished opposing camps in the replication debate has been the extent to which the existing empirical literature in psychology is seen as fundamentally flawed versus the extent to which current replications initiatives are seen as problematic. The present framework is agnostic in this debate. It suggests that problems in original studies, replication studies, or both could lead to discrepant results. However, it also provides a set of principles that can help to organize discussions about how both original and replication studies can be improved to increase replication rates in psychological research.

Making original studies more replicable. The validity typology proposed by Cook and Campbell (1979) has long informed methodological decisions and post hoc evaluations of psychological research. Yet, it might be that this framework is less prominently featured in methodological training than it once was and that contemporary researchers are less familiar with it than earlier generations. If so, a renewed emphasis on this perspective and more explicit consideration of it during the design, interpretation, and evaluation of original research would be a healthy development in the field. At a more concrete level, the framework suggests several ways in which new research might be improved to facilitate future replications.

Statistical conclusion validity issues. There is no doubt that some problematic statistical practices and views have existed in the field. Addressing these concerns could be beneficial. Evaluating the strengths and limitations of the numerous specific recommendations that have been offered for remedying such concerns goes beyond the scope of the current discussion, but we note that any effective set of reforms will have to address more than just power and in addressing power will have to move beyond functionally equating it with sample size. Like the other forms of validity, statistical conclusion validity is complex and is not well captured by rigid rules of thumb. Furthermore, efforts to enhance statistical conclusion validity will have to be considered within the broader context of also considering other forms of validity.

Construct validity issues. One way the replicability of psychological research can be enhanced is by paying more attention to the construct validity of operationalizations of IVs and DVs. In some topic areas, the construct validity of measures or manipulations has been taken very seriously. Indeed, in some cases, the development of operationalizations of constructs and evaluation of their psychometric properties has been a major research topic in its own right. For example, personality psychologists have long emphasized developing and carefully validating formal scales of personality (e.g., see O. P. John & Srivastava, 1999). Likewise, the development and evaluation of formal methods of assessing attitudes has been a central topic in social psychology dating back to the 1920s (Thurstone, 1928; see Krosnick et al., 2019; Summers, 1971). However, in other areas, measures and manipulations have been constructed in a fairly informal and ad hoc fashion. These operationalizations often involve little or no pretesting in their development and undergo little or no formal evaluation of their psychometric properties (e.g., see Flake et al., 2017). Although such operationalizations might "work well enough" to sometimes produce their predicted effects, operationalizations constructed in this fashion are much more likely to produce unstable results. Unfortunately, even modest increases in the random error of a measure can lead to substantial increases in the variability of effect sizes obtained across studies even when sampling from exactly the same population (Stanley & Spence, 2014). Moreover, such "noisy" operationalizations might be more likely to shift across contexts and populations than more precise operationalizations that have undergone careful psychometric evaluation to minimize random and systematic error. Thus, a greater emphasis on requiring researchers to provide evidence of the construct validity of their operationalizations in original (and replication) research could help reduce discrepancies between their outcomes.

Along these lines, it is critical to better document the conceptual logic that guided the operationalizations in the first place, the specific pretesting protocols used to inform the operationalizations, the psychometric criteria used to determine that the operationalizations were satisfactory, and the results of psychometric assessments. In many cases, more work goes into the development of operationalizations than might be apparent from an article. With limited journal space, authors are often tempted (sometimes at the behest of editors and reviewers) to only briefly mention or completely omit such "ancillary details" in the interest of allowing fuller discussion of their substantive findings. This lack of construct validity details poses a major challenge for replications. Fortunately, with the ability to provide long-term availability of supplementary information online, such details need not be lost. On a related matter, in recent years, there has been an increased emphasis on making study materials and measures available to assist in future replications. Such resources are useful and can provide a helpful starting point for a replication. However, operationalizations are often developed to optimize construct validity within the specific context and population in which they will be used rather than with the intent of making their psychological properties broadly invariant over contexts and populations (e.g., see Fabrigar & Wegener, 2016; Petty, 2018; Stroebe & Strack, 2014). Thus, the exact materials and measures used in a prior study does not necessarily provide a "replication recipe" (Zwaan et al., 2018b) that can be applied without thought and careful evaluation for suitability in a new participant sample or context (Petty, 2018; cf. Brandt et al., 2014). Rather, providing information on the conceptual logic underlying operationalizations and the process by which they

were developed and by which their validity was evaluated would often be the more useful "recipe" for future replications.

External validity issues. Consideration as to the external validity of a given phenomenon is also important for replication. However, this consideration should be approached in a way that is perhaps different than how researchers often think of external validity. External validity is frequently conceptualized in terms of whether a phenomenon is likely to also emerge in an alternative particular context (e.g., an accounting office), particular population (e.g., Filipinos), or combination (e.g., Filipinos working in an accounting office). Obviously, the potential features of context and population are virtually infinite and such aggregate categories might well involve many differences between the original and new setting or population. Specifying the external validity of a given effect at this level is seldom productive. Rather, external validity is best addressed by grounding effects in theory.

Specifying, the underlying processes responsible for an effect and the psychological constructs most likely to regulate the emergence of these processes is at its core a consideration of external validity. Moreover, specifying the boundary conditions in terms of general principles is more tractable and likely to be more broadly useful to future researchers. For example, instead of asking whether a result obtained with college students would also be obtained with factory workers (cf. Sears, 1986), one could ask the more conceptual question of whether an effect obtained with people relatively high in motivation to think (or issue-relevant knowledge or interest in a topic) would be obtained with people relatively low in motivation to think (or issue-relevant knowledge or interest in a topic). Importantly, each of these conceptual variables can be studied within a population of college students or factory workers (see Petty & Cacioppo, 1996). Although the features of contexts and populations that might influence a particular effect are vast in number, the underlying processes by which these features regulate an effect are likely to be much more limited. Strong theory can thus provide researchers with a manageable set of principles that they can apply to the specific population and context in which a future replication will occur.

As suggested earlier, documentation of background context and population considerations is important for future replications. Often researchers have a clear notion of the sorts of conditions most likely to be conducive to their effects and they construct contexts to produce these conditions. Because of journal space constraints and the fact that these features are constants that do not function as either IVs or DVs, researchers might not explicitly mention these background features or, if they are mentioned, they might not be fully explained. More complete documentation of such considerations in supplementary online materials would be a valuable resource for scholars conducting future replications. As for operationalizations, the specific features of the background factors (e.g., the precise lab room size), although potentially helpful, are not necessarily the most essential information. Rather, the underlying logic for these features (e.g., avoidance of feeling of crowding; creating a low level of baseline information processing), the process by which they were constructed, and any assessments of their efficacy would be more valuable.

Internal validity issues. Internal validity issues have probably not been a major source of problems in original laboratory-based experimental studies. Where such problems could have implications for future replications, however, these problems should be documented and the methods by which they were remedied explained. For instance, if manipulations are found to produce (differential) attrition across conditions and protocols were developed to manage such difficulties, these procedures should be documented in the article or supplementary materials.

Making replication studies more capable of replicating. More explicitly formulating replication efforts in terms of the four types of validity could improve replication rates. To date, replication researchers have been reasonably attentive to issues related to statistical conclusion validity, and such practices are a strength of current replication efforts. Multi-lab efforts have made possible testing of psychological effects using much larger samples than would have been feasible following traditional research practices. In addition, the careful application of meta-analytic tools to such data is a positive development (although such analyses are obviously only as good as the data upon which they are based and cannot overcome fundamental methodological flaws; for example, see Bornstein et al., 2009; H. Cooper, 2017; Nelson et al., 2018). However, active pursuit of statistical conclusion validity should be balanced against the need to consider other forms of validity. Obtaining large samples and testing many different effects within a single study session (where the original research did not conduct multiple studies on the same participants), though appealing in some respects, are not virtues if they undermine other forms of validity.

Streamlining measures and manipulations for convenience should generally be avoided unless the resulting psychometric implications are carefully evaluated (e.g., is construct validity comparable?; Fabrigar & Wegener, 2016). Likewise, if an original study included pretesting and validation of its operationalizations, replicating this phase of the research process should be considered as integral to the replication effort as to the main study (Petty, 2018). Finally, when data permit the evaluation of the psychometric properties of the operationalizations used in the replication study, analyses should be undertaken and fully reported (Flake et al., 2017). Moreover, their implications (e.g., whether such analyses suggest decreased construct validity which could in turn account for a failed replication of the key effect of

interest) for interpreting the substantive effect of interest should also be considered. Construct validity information in replication efforts might sometimes be as interesting and informative as the actual tests of the key effect of interest.

External and internal validity concerns should also be more explicitly addressed in replication studies. Replication researchers should carefully consider the contextual factors and characteristics of the population that are known to be relevant to the emergence of the effect of interest. Often, this can only be determined by delving into the literature related to an effect one aims to replicate. The research relevant to a replication often goes beyond the specific theory and data that guided the original research. Therefore, becoming an expert in the literature related to replications is no less necessary than becoming an expert in the literature related to one's original research. Replication researchers should then attempt to create optimal conditions to demonstrate the effect, or at a minimum, conditions at least as good as in the original study. Of course, this is more feasible to the extent that original researchers provide the necessary information. Efforts to match the conditions of an original study should also be carefully documented by replicators. One implication of this consideration is that if multiple effects are going to be tested as a set, only effects that are likely to emerge in the same context and population should be investigated together. Because different psychological effects can reflect different psychological processes, the conditions conducive to one effect might inhibit another effect (e.g., a psychological effect most likely to emerge when people are engaging in relatively little effortful thought would not pair well with a psychological effect more likely to occur when people are being highly thoughtful). This possibility should be considered when designing replication studies. Finally, with respect to internal validity, researchers should be attentive to how changes in population (e.g., student vs. MTurk) or context (e.g., laboratory vs. online) might affect threats to internal validity such as differential attrition and fully document how such concerns have been managed.

As suggested earlier, replication efforts must be informed by more than the original publication. The first or best-known demonstration of an effect may not be the optimal demonstration or the best description of methods for producing the effect. Subsequent research often leads to improved operationalizations and a better understanding of boundary conditions of an effect. Even when the original study is comparatively optimal, many critical methodological features important to demonstrating the effect might not have been understood at the time. Thus, these features might not have been highlighted in the description of the methods in the original work. In light of these facts, it is essential for replication researchers to familiarize themselves with the broader methodological and theoretical literature related to the phenomenon of interest and conduct an "informed replication" (Luttrell et al., 2017). Failure to consider the broader literature might have contributed to some of the deviations from existing theory and established practices that went undocumented and probably unrecognized in some replication studies.

A broader consideration of different forms of validity also highlights the inferential ambiguity of failed replications. A failed replication has very different methodological or theoretical implications depending on which form(s) of validity might be responsible. Simply reporting that an effect was not replicated is of modest value if little insight can be provided as to why this failure occurred (see also Wegener & Fabrigar, 2018) or if the default assumption is that the failed replication is due to Type I error in the original research. Because non-replication is a joint function of both the original and replication study, arguing that the burden of responsibility for explaining non-replication rests with the original researchers seems one-sided. Such a view is even more suspect if replication researchers have not provided evidence that they carefully attended to all four forms of validity, at least to the same level as in the original research. Again, however, the job of replicators is made simpler to the extent that original researchers are also clear about their attention to these validities. Ultimately, the causes of non-replication are shared and thus explanations for why it happens should be a shared responsibility. With this in mind, replication efforts should emphasize not only examining whether an effect replicates but also testing potential explanations for replication failure when they occur rather than simply indicating whether a plausible explanation for replication failure can be generated.

Conclusion

Whether researchers recognize the fact or not, all psychological studies—whether original or replication—involve a balancing of the four types of validities based on the goals of the research. Different goals mandate different choices in emphasizing these forms of validity. Ideally, the balancing of different validity considerations should be explicitly addressed by researchers when conducting original studies. Doing so not only enhances the methodological rigor of the procedures but also the clarity of the logic behind the procedures and the fundamental goals of the research. The same explicit consideration of this balance is also essential when replication studies are conducted. Doing so enhances not only the alignment of the replication study with the methodological logic of the original study but also with its fundamental goals. If either is out of alignment, replication efforts run the risk of leading to more ambiguity than insight.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported by a Social Sciences and Humanities Research Council of Canada Insight Grant (435-2015-0114) to the first author.

ORCID iD


Duane T. Wegener https://orcid.org/0000-0001-7639-4251

Notes

1. There is some ambiguity regarding when two studies can be said to produce consistent versus inconsistent evidence for the existence of an effect. Different conclusions can be reached depending on the criterion one uses to define consistency (e.g., for various discussions, see Braver et al., 2014; Fabrigar & Wegener, 2016; Maxwell et al., 2015; Open Science Collaboration, 2015). For purposes of the present discussion, we assume the relatively clear case of non-replication in which the original study clearly demonstrated an effect and the replication study has produced a highly discrepant result (e.g., a zero or near-zero effect).
2. Our article is not the first to consider replication in the context of the four Cook and Campbell (1979) validities. However, prior discussions (e.g., Finkel et al., 2017) have tended to conflate replication with statistical validity and emphasize that the "other" validities are also important in designing original research (i.e., in designing research one should consider replication/statistical validity along with the other *competing* types of validity). In contrast, we highlight how each of the four types of validity can contribute to the replicability of a finding (i.e., we do not treat replicability as primarily a statistical conclusion validity concern, but argue that each of the four validities has a role to play in conducting a successful replication study).
3. Recent treatments of heterogeneity of effects across even exact replication studies also suggest that it may not be meaningful to conceive of a single "true" effect size in a population (e.g., Kenny & Judd, 2019). This perspective has a number of interesting implications for statistical power, especially suggesting that a set of smaller (moderately powered) studies can represent greater power as a set than a single large (highly powered) study of the same N as the set of smaller studies.
4. It is also possible for an effect in a replication study to falsely appear to provide convergent evidence of an effect. This could occur if the operationalizations in the replication do not map onto the original constructs, but the unintended constructs represented in the replication effort are related to each other in a manner that parallels the constructs of interest.
5. Shadish et al. (2002) extended this general definition to include whether the results would generalize to other operationalizations of the IV and DV. Because these considerations can also involve questions of construct validity (as noted above), we do not address this extension here.
6. Or the operation used to assess aggression fails to tap into this construct for women.
7. The remaining guidelines related to other practices such as word limits, formatting, and so forth.
8. This asymmetry parallels another related asymmetry that has sometimes emerged in the replication literature. That is, claiming a meaningful effect in original research often requires examination of potential alternative explanations for it. In contrast, potential alternative explanations for failures to replicate are rarely addressed even though, very much like in original research, some of the potential causes of replication failure are theoretically uninteresting, whereas others are potentially quite interesting (see Wegener & Fabrigar, 2018).
9. In considering cases of non-replication, we exclude cases where the data themselves are fraudulent.
10. We have conceptualized Strack's second concern in terms of external validity if the facial-feedback effect is presumed to only emerge for certain types of cartoons (i.e., cartoons that are moderately funny, but not for cartoons that are unfunny or confusing). However, this concern could also be framed in terms of construct validity. Specifically, an alternative (perhaps not mutually exclusive) framing might be that the construct validity of the perceived funniness dependent measure is eroded when applied to cartoons that are unfunny or confusing, because people are incapable of meaningfully responding to such a measure when rating cartoons that evoke very low levels of the intended construct.
11. A fourth explanation noted by Strack (2016) involved an apparent statistical anomaly in the meta-analysis for the replication studies across labs. Specifically, he noted that although effect size and sample size should be uncorrelated, the Wagenmakers et al. (2016) studies produced a positive correlation. He noted that this positive correlation is the opposite of what one would expect if "*p*-hacking" has occurred (i.e., engaging in practices in an attempt to exaggerate the statistical evidence for an effect). This correlation would instead be consistent with reverse *p*-hacking (i.e., engaging in practices in an attempt to weaken statistical evidence for an effect). He argued that it was important to resolve this anomaly. If considered in light of the Cook and Campbell (1979) typology, this fourth explanation could be considered a case in which it is being suggested that the replication study might be of lower statistical conclusion validity than the original study.
12. It is also worth noting that the Marsh et al. (2019) test of the facial-feedback effect was not included in this meta-analysis because it was published subsequent to the studies reviewed in the Coles et al. (2019) meta-analysis.
13. Petty and Cacioppo (2016) also noted that shortening the messages from the original could have produced other unintended effects. Specifically, research has shown that those high in need for cognition are less motivated to think about simple than complex messages whereas those low in need for cognition are more motivated to think about simple than complex messages (See et al., 2009). If the short messages struck recipients as simple to process, it could have motivated low-NC participants to think and reduced motivation of high-NC participants to think, thereby reducing the likelihood of obtaining the original

NC × AQ interaction on attitudes. Thus, the use of shorter messages could also have created a context in which the originally demonstrated effect might not be expected to emerge (an external validity difference between the original and replication research).

14. The use of weaker independent variables in the replication than in the original could also compromise statistical conclusion validity, but the large increase in the sample size in the replication should have at least in part compensated for that. Thus, we focus on external and construct validity concerns.
15. The only factor that could be specifically evaluated in terms of its impact on Ebersole et al.'s failure to replicate the original Cacioppo et al. finding was their use of a shorter NC scale. Luttrell et al. compared analyses based on items from the short and long versions of the scale. The key interaction was obtained with both sets of items, though the effect size was larger with the longer scale as might be expected given its higher construct validity and reliability. Even in this case, however, the role of the shorter scale cannot be definitively assessed because responses to the six short-scale items in the context of the longer scale might be different than responses to these same six items in isolation (e.g., see Knowles, 1988).
16. Introducing multiple factors that erode one or more types of validity could be even more consequential than might be expected from a consideration of these factors in isolation. If the adverse effects of these sources of error combine in a multiplicative (interactive) manner rather than an additive fashion, the consequences could be even more problematic and difficult to offset even in cases where enhanced sample size might offer modest protection against all or some of these threats to validity.

References

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

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*, 842-850.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423-437.

Bakker, B. N., & Lelkes, Y. (2018). Selling ourselves short? How abbreviated measures of personality change the way we think about personality and politics. *The Journal of Politics*, *80*, 1311-1325.

Barrett, L. F. (2015, September 1). Psychology is not in crisis. *The New York Times*, A23.

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

Bartlett, T. (2018, September 21). I want to burn things to the ground: Are the foot soldiers behind psychology's replication crisis saving science—Or destroying it. *The Chronicle of Higher Education*, B6-B9.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., … Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6-10.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley.

Bornstein, M., Hedges, L. V., Higgins, P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley.

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

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333-342.

Brewer, M. B., & Crano, W. D. (2014). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 11-26). Cambridge University Press.

Button, K. S., & Munafò, M. R. (2017). Powering reproducible research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny* (pp. 22-33). John Wiley.

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

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116-131.

Cacioppo, J. T., Petty, R. E., Feinstein, J., & Jarvis, B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197-253.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of "need for cognition." *Journal of Personality Assessment*, *48*, 306-307.

Cacioppo, J. T., Petty, R. E., & Morris, K. (1983). Effects of need for cognition on message evaluation, argument recall, and persuasion. *Journal of Personality and Social Psychology*, *45*, 805-818.

Calanchini, J., Moons, W. G., & Mackie, D. M. (2016). Angry expressions induce extensive processing of persuasive appeals. *Journal of Experimental Social Psychology*, *64*, 88-98.

Calin-Jageman, R. J. (2018). Direct replications of Ottati et al. (2015): The earned dogmatism effect occurs only with some manipulations of expertise. *Journal of Experimental Social Psychology*, *78*, 249-258.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Rand McNally.

Carey, B. (2018, July 17). Psychology itself is under scrutiny. *The New York Times*, *D5*.

Carpenter, C. J. (2015). A meta-analysis of the ELM's argument quality × processing type predictions. *Human Communication Research*, *41*, 501-534.

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

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304-1312.

Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, *145*, 610-651.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Rand McNally College.

Cooper, H. (2017). *Research synthesis and meta-analysis: A-step-by-step approach* (5th ed.). SAGE.

Cooper, J. (2007). *Cognitive dissonance: Fifty years of a classic theory*. SAGE.

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.

Crano, W. D., Brewer, M. B., & Lac, A. (2015). *Principles and methods of social research* (3rd ed.). Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

De Fruyt, F., Van De Wiele, L., & Van Heeringen, C. (2000). Cloninger's psychobiological model of temperament and character and the five-factor model of personality. *Personality and Individual Differences*, *29*, 441-452.

Dijksterhuis, A. (2014). Welcome back theory! *Perspectives on Psychological Science*, *9*, 72-75.

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., . . . Nosek, B. A. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty, & Xu (2017) demonstrate good science. *Journal of Experimental Social Psychology*, *69*, 184-186.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68-82.

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

Fabrigar, L. R., & Wegener, D. T. (2014). Exploring causal and noncausal hypotheses in nonexperimental data. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 504-533). Cambridge University Press.

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

Fabrigar, L. R., Wegener, D. T., Vaughan-Johnston, T. I., Wallace, L. E., & Petty, R. E. (2019). Designing and interpreting replication studies in psychological research. In F. Kardes, P. Herr, & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 483-507). Routledge.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α—Error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661-669.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*, 45-52.

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

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*(2), 244-253.

Flake, J. K., & Fried, E. I. (2019, January 17). Measurement schmeasurement: Questionable measurement practices and how to avoid them. https://doi.org/10.31234/osf.io/hs7wm

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370-378.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*, 585-594.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*(3), 29-31.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, 351, Article 1037.

Giner-Sorolla, R., Aberson, C. L., Bostyn, D. H., Carpenter, T., Conrique, B. G., Lewis, Jr. N. A., . . . Soderberg, C. (2019). *Power to detect what? Considerations for planning and evaluating sample size*. Report of the SPSP Power Analysis Working Group. https://osf.io/jnmya/

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

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

Haas, R. G. (1984). Perspective taking and self-awareness: Drawing an E on your forehead. *Journal of Personality and Social Psychology*, *46*, 788-798.

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

Harmon-Jones, E. (Ed.). (2019). *Cognitive dissonance: Reexamining a pivotal theory in psychology* (2nd ed.). American Psychological Association.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLOS ONE*, *8*(*8*), Article e72467.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*, 61-83.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*, 539-556.

Hughes, B. M. (2018). *Psychology in crisis*. Palgrave.

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Science*, *113*, E4933-E4934.

Insko, C. A., Worchel, S., Songer, E., & Arnold, S. E. (1973). Effort, objective self-awareness, choice and dissonance. *Journal of Personality and Social Psychology*, *28*, 262-269.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), Article e124.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Research practices with incentives for truth telling. *Psychological Science*, *23*, 524-532.

John, O. P., & Benet-Martinez, V. (2014). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 473-503). Cambridge University Press.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). Guilford Press.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*(4), 411-414.

Judd, C. M., Yzerbyt, V. Y., & Muller, D. (2014). Mediation and moderation. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 653-676). Cambridge University Press.

Kenny, D. A. (2019). Enhancing validity in psychological research. *American Psychologist*, *74*, 1018-1028.

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*, 578-589.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, Jr. R. B., Bahnik, S., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variability in replication: A Many Labs replication project. *Social Psychology*, *45*, 142-152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variations in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443-490.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*, 312-320.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2019). The measurement of attitudes. In D. Albarracin & B. T. Johnson (Eds.), *The handbook of attitudes, Volume 1: Basic principles* (2nd ed., pp. 45-105). Routledge.

Kruglanski, A. W. (1975). The human subject in the psychology experiment: Fact and artifact. *Advances in Experimental Social Psychology*, *8*, 101-147.

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

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's evidence of Psi as a case study of deficiencies in modal research practices. *Review of General Psychology*, *15*, 371-379.

Libby, L. K., & Eibach, R. P. (2011). Visual perspective in mental imagery: A representational tool than functions in judgment, emotion, and self-insight. *Advances in Experimental Social Psychology*, *44*, 185-245.

Lilienfeld, S. O., & Waldman, I. D. (Eds.). (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. John Wiley.

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*, 178-183.

Marsh, A. A., Rhoads, S. A., & Ryan, R. M. (2019). A multi-semester classroom demonstration yields evidence in support of the facial feedback effect. *Emotion*, *19*, 1500-1504.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is Psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487-498.

McGuire, W. J. (1983). A contextualist theory of knowledge: Its implications for innovation and reform in psychological research. *Advances in Experimental Social Psychology*, *16*, 1-47.

McIsaac, H. K., & Eich, E. (2002). Vantage point in episodic memory. *Psychonomic Bulletin and Review*, *9*, 146-150.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806-834.

Monin, B., & Miller, D. T. (2001). Moral credentials and expression of prejudice. *Journal of Personality and Social Psychology*, *81*, 33-43.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Aldine.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511-534.

Noah, Y., Schul, T., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*, 657-664.

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*, 615-631.

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

Ottati, V., Wilson, C., Osteen, C., & Distefano, Y. (2018). Experimental demonstrations of the earned dogmatism effect using a variety of optimal manipulations: Commentary and response to Calin-Jageman (2018). *Journal of Experimental Social Psychology*, *78*, 240-248.

Ottati Price, E. D., Wilson, C., & Sumaktoyo, N. (2015). When self-perceptions of expertise increase closed-minded cognition: The earned dogmatism effect. *Journal of Experimental Social Psychology*, *61*, 131-138.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 631-636.

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, *24*, 590-605.

Petty, R. E. (2018). The importance of exact conceptual replications. *Behavioral and Brain Sciences*, *41*, Article e146.

Petty, R. E., Briñol, P., Loersch, C., & McCaslin, M. J. (2009). The need for cognition. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 318-329). Guilford Press.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*, 1915-1926.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer-Verlag.

Petty, R. E., & Cacioppo, J. T. (1990). Involvement and persuasion: Tradition versus integration. *Psychological Bulletin*, *107*, 367-374.

Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, *33*, 1-8.

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.

Priester, J. M., & Petty, R. E. (1995). Source attributions and persuasion: Perceived honesty as a determinant of message scrutiny. *Personality and Social Psychology Bulletin*, *21*, 637-654.

Reis, H. T., & Judd, C. M. (Eds.). (2014). *Handbook of research methods in social and personality psychology* (2nd ed.). Cambridge University Press.

Scheier, M. F., & Carver, C. S. (1980). Private and public self-attention, resistance to change, and dissonance reduction. *Journal of Personality and Social Psychology*, *39*, 390-405.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551-566.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. *Psychological Methods*, *1*, 115-129.

Schul, Y., Noah, T., & Mayo, R. (2018, May 10). *Response by Yaacov Schul, Tom Noah, and Ruth Mayo*. https://www.baye sianspectacles.org/musings-on-preregistration/

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*, 195-202.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515-530.

See, Y. H. M., Petty, R. E., & Evans, L. M. (2009). The impact of perceived message complexity and need for cognition on information processing and attitudes. *Journal of Research in Personality*, *43*, 880-889.

Shadish, W., Cook, T., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Houghton Mifflin Harcourt.

Shaffer, J. P. (2002). Multiplicity, directional (Type III) errors, and the null hypothesis. *Psychological Methods*, *7*(3), 356-369.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., . . . Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLOS ONE*, *8*, Article e56515.

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*(11), 1359-1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76-80.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534-547.

Smith, E. R. (2014). Research design. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 27-48). Cambridge University Press.

Smith, S. M., & Petty, R. E. (1996). Message framing and persuasion: A message processing analysis. *Personality and Social Psychology Bulletin*, *22*, 257-268.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845-851.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*, 305-318.

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*, 929-930.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768-777.

Stroebe, W. (2016). Are most published psychological findings false? *Journal of Experimental Social Psychology*, *66*, 134-144.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59-71.

Summers, G. F. (Ed.). (1971). *Attitude measurement*. Rand McNally.

Szymkow, A., Chandler, J., Ijzerman, H., Parzuchowski, M., & Wojciszke, B. (2013). Warmer hearts, warmer rooms: How positive communal traits increase estimates of ambient temperature. *Social Psychology*, *44*, 167-176.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207-232.

Vallacher, R. R. (1978). Objective self awareness and the perception of others. *Personality and Social Psychology Bulletin*, *4*, 63-67.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016a). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454-6459.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016b). Reply to Inbar: Contextual sensitivity helps explain the reproducibility gap between social and cognitive psychology. *Proceedings of the National Academy of Sciences*, *113*(34), E4935-E4936.

Van Lange, P. A. M., Kruglanski, A., & Higgins, E. T. (Eds.). (2012). *Handbook of theories of social psychology* (Vols 1 and 2). SAGE.

Vaughan-Johnston, T. I., Matthews, M., Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). *A content analysis of the replication literature in personality and social psychology and beyond: Conceptual perspectives and current practices* [Manuscript in preparation].

Vazire, S. (2016). Editorial. *Social Psychological and Personality Science*, *7*, 3-7.

Wagenmakers, E. J., Beck, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, and Stepper (1988). *Perspectives on Psychological Science*, *11*, 917-928.

Wagenmakers, E.-J., & Gronau, Q. (2018, May 10). *Musings on preregistration: The case of the facial-feedback effect*. https://www.bayesianspectacles.org/musings-on-preregistration/

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123-138). John Wiley.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*, 273-295.

Wegener, D. T., & Fabrigar, L. R. (2018). Holding replication studies to mainstream standards of evidence. *Behavioral and Brain Sciences*, *41*, Article e155.

Wegener, D. T., & Petty, R. E. (1994). Mood management across affective states: The hedonic contingency hypothesis. *Journal of Personality and Social Psychology*, *66*, 1034-1048.

West, S. G., Cham, H., & Liu, Y. (2014). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 49-80). Cambridge University Press.

Wheeler, S. C., Petty, R. E., & Bizer, G. Y. (2005). Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. *Journal of Consumer Research*, *31*, 787-797.

Wicklund, R. A., & Duval, S. (1971). Opinion change and performance facilitation as a result of objective self-awareness. *Journal of Experimental Social Psychology*, *7*, 319-342.

Widaman, K. F., & Grimm, K. J. (2014). Advanced psychometrics: Confirmatory factor analysis, item response theory, and the study of measurement invariance. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 534-570). Cambridge University Press.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, M. B. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis: An introduction for psychologists* (pp. 39-61). American Psychological Association.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*, 493-504.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018a). Improving social and behavioral science by making replication mainstream: A response to commentaries. *Behavioral and Brain Sciences*, *41*, Article e157.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018b). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, Article e120.