DESIGNING AND INTERPRETING REPLICATION STUDIES IN PSYCHOLOGICAL RESEARCH

Leandre R. Fabrigar, Duane T. Wegener, Thomas I. Vaughan-Johnston, Laura E. Wallace, and Richard E. Petty

We live in an era of angst about how research in psychology is conducted, analyzed, and interpreted. To suggest otherwise ignores significant scholarship in every type of academic outlet, from conference presentations to empirical journals to theoretical journals, book chapters, and edited or authored books. Indeed, the discussion has extended beyond traditional academic outlets to reach websites, newspapers, blogs, podcasts, and more. The many discussions have not represented idle chatter. Journals have taken steps to change submission guidelines. New journals have been created to address methodological practices and advocated improvements. New research funding efforts and data collection efforts have focused on enhancing the role of replication studies per se, and funding agencies have increased attention to factors they believe will increase replicability of results.

The accumulating literature on replication is vast. New publications appear every month, and the debate has, at times, been intense. Passionate scholars have been quoted as saying everything from, it is time to “burn things to the ground” to calling some reformers “methodological terrorists” (Bartlett, 2018). Recommendations for methodological improvement have been numerous and advocated with great, perhaps sometimes outsized, confidence. As the data from “pure replication” studies have come in (which often show a replication rate of 50% or less), many hypotheses have been offered to account for observed (lack of) replication in individual studies and for rates of replication across research areas, design factors, and samples. Those hypotheses range from dismissal of the original findings as “false” (representing one end of the continuum that places greater trust in the replication study) to criticism of the replication effort as not having been conducted appropriately (representing the other end of the continuum that places greater trust in the original research). Emerging from those efforts, many suggestions have been made for how to improve replication rates.

Some of the suggested methodological changes have been based on statistical principles or mathematical modeling, others on simulations, and yet others have followed observations of replication patterns in a particular set of studies. Perhaps as a result of the varied bases for these suggestions, however, some of the recommendations actually conflict with one another. Because of this, one theme we emphasize in the current review is the trade-offs one must consider when designing research (whether a replication study or original research; see also Finkel, Eastwick, & Reis, 2017). Also, many of the hypotheses made to explain replication rates or suggestions made to increase replication rates rest on strong, but often unstated and untested, assumptions. Therefore, another goal for the current review is to make some of these assumptions explicit and to discuss how realistic the
assumptions are, as well as how the assumptions fit together when trying to maximize the validity of replication research.

To date, there has been little attempt to conceptually organize the literature on replication. Most of the discussions have focused on one or a very limited number of principles, with few attempts to relate the various suggestions to one another or to arrange them into a coherent approach to replication. Therefore, the key goal for the current chapter is to provide a broader framework in which to view the replication literature. More specifically, we will use Cook and Campbell’s (1979) classic validity typology to organize this literature (see also Shadish, Cook, & Campbell, 2002). The core notions of statistical conclusion validity, internal validity, construct validity, and external validity have been used for many years to evaluate original research. In our view, these same methodological principles provide a coherent framework to conceptualize replication efforts and identify dimensions along which suggested reforms might be evaluated.

We believe that the validities framework can serve to highlight the often unstated (and/or untested) assumptions currently prominent in the replication literature. This framework can also identify new, or highlight largely neglected, explanations for failures to replicate. In so doing, the validities approach can also assist in understanding and predicting when and why failures to replicate are likely to occur. Such understanding should also suggest strategies for designing replication studies and follow-up research when replication studies fail to produce the expected results. Perhaps even more importantly, reasons for failures to replicate may be related to one another. Therefore, a thorough understanding of the relevant issues should help to highlight likely relations among the causes of failures to replicate. We believe that that validities approach can help to identify those relations as well as the trade-offs that should be considered in designing and interpreting replication research.

Conceptual Overview

Purpose of the Research

Any evaluation of research must start with the purpose of the research, such as whether the research is best considered basic or applied. In most basic research in consumer psychology, the primary goal is to test hypotheses regarding relations among psychological constructs. For example, consider the research conducted by Isen, Shalker, Clark, and Karp (1978, Study 1). The primary research question for that study was whether positive mood would lead to more favorable perceptions of consumer products compared with perceptions of consumers in a more neutral mood. In the study, research participants at a shopping mall were either given a small gift (i.e., notepads to females or nail clippers to males) or were not. Potential participants were then approached by another experimenter conducting a “consumer survey” and were asked to evaluate the present performance and service record of a product they own (i.e., either an automobile or a television set). Results showed that receipt of a gift increased ratings of product performance and reliability, regardless of the sex of the research participant.

The researchers did not conduct the study because they were interested in the specific gifts that were given. Rather, as reported by the researchers, those gifts were chosen because they produced a good mood in the recipient (and different gifts were given to male vs. female potential participants because prior research had shown that the notepad gift was only effective with females, not males). These particular gifts were not at all the only ways to create the desired mood states, and, across a set of studies, the researchers might often use different methods, such as videos or written or audio material (or even imagination on the part of the participants) to bring about similar mood states. For example, in another consumer study investigating the impact of mood on evaluation, students saw an advertisement for a new brand of pen that was embedded in either an episode of a popular situation
comedy TV show or a neutral documentary (Petty, Schumann, Richman, & Strathman, 1993). Students who saw the ad in the context of the comedy provided more favorable ratings of the pen on a series of semantic differential scales. Again, the researchers were not interested in the particular television shows used in the research. The programs were only selected because pretesting showed they induced the appropriate moods in the participant sample.

Similarly, the indexed perceptions of product performance and reliability or the semantic differential scales were not the only possible ways to index favorability of product views in these studies. For example, other studies could test conceptually similar ideas by using different measures of overall attitudes toward the products, or intentions to buy new target products described in an advertisement, or perhaps even ratings of particular product features or actual purchases of the products. Furthermore, in addition to these direct or explicit measures, it would have been possible to gauge evaluation using more indirect or implicit assessments. That is, any type of measure that could be mapped onto the degree of positivity toward the product might have been used.

For these studies and any other basic consumer research, the specific manipulations and measures are often arbitrary in that different, and often many, particular operations might serve as instantiations of the conceptual variables of interest. The hypotheses in such settings are about the constructs influenced by the manipulations or assessed by the measures, not about the particular manipulations or measures themselves. Another important aspect of such settings is that the hypotheses generally specify a direction of effect rather than a particular effect size. For example, many theories of mood and judgment would suggest that perceptions of encountered targets are typically assimilated to the valence of the mood state of the perceiver (e.g., Forgas, 1994; Isen et al., 1978; Schwarz & Clore, 1983; Wegener & Petty, 2001). But none of these theories specifies particular effect sizes for the predicted effects of mood on judgments. That is, under conditions where assimilation effects are expected, these theories predict positive relations between mood and judgment (e.g., an $r = .4$, with positive direction associated with more favorable ratings in positive than in neutral or negative mood). But none of these theories is formulated such that one effect size would support the theory (e.g., $r = .6$), but another significant effect in the same direction would conflict with the theory (e.g., $r = .4$). In this sense, then, the focus in basic research is not on effect size per se (though effect sizes are often reported as one way in which to aggregate effects across studies).

In contrast, in applied studies, researchers can be quite concerned about the particular operationalizations used in a study and about the particular effect size they produce (Petty & Cacioppo, 1996). For example, imagine an applied researcher is testing the effectiveness of two possible comedians (A and B) to use in a campaign. If Comedian A is more effective but costs more money than B, the company might decide that Comedian A’s expenses are only worth it if actual sales reach a particular level that is higher than B’s. In such cases, the researcher cares about the particular operations tested, both the particular independent variable (i.e., comedian A versus B) and the particular dependent measure (amount of sales). As we will discuss later in the chapter, we believe that some misunderstandings related to replication reflect a lack of distinguishing between basic research focused on relations between constructs and applied research focused on particular effect sizes of specific operationalizations.

As suggested earlier, we argue that research can be evaluated with respect to the four types of validity outlined by Cook and Campbell (1979), regardless of whether that research is original research or replication research. Traditionally, the validities have been taught as foundational principles by which original research might be understood. Yet, discussions of replication have typically either ignored the traditional validities altogether or have focused on singular issues, such as statistical power, to the exclusion of other validity-related factors. In contrast, we organize our discussion of replication by referencing the different types of validity (statistical conclusion validity, internal
validity, construct validity, and external validity). By doing so, we also discuss how these principles relate to one another, as well as the trade-offs inherent in maximizing one type of validity at the possible expense of other types of validity. A key premise underlying our discussion is that inconsistent results across original and replication studies often reflect differences across the studies in one or more of these four types of validity.

Defining Replication

Before we begin our discussions of the four validities, it is important to note that, in the contemporary replication literature, researchers have become increasingly sensitive to the fact that there is actually a fair amount of ambiguity in regards to when two studies can be said to have produced consistent versus inconsistent evidence for the existence of an effect (e.g., see Braver, Thoemmes, & Rosenthal, 2014; Fabrigar & Wegener, 2016; Maxwell, Lau, & Howard, 2015). The traditional and still most common definition of replication is whether or not the study achieved statistical significance of an effect (i.e., p < .05) falling in the same direction as in the original research (cf. Fabrigar & Wegener, 2016; Open Science Collaboration, 2015; Simonsohn, 2015). Others have occasionally discussed replication by referring to obtaining an effect in the replication study that falls within the confidence interval of the original study, or an effect in the replication study that falls in the same direction as in the original research (e.g., Open Science Collaboration, 2015). Another approach to replication comes out of a meta-analytic approach to evaluating evolving support for an effect (e.g., Braver et al., 2014; Stanley & Spence, 2014). That is, one way to address the issue of replication is to consider whether evidence from the replication study strengthens or weakens the meta-analytic case for the effect when combined with previous research (see Fabrigar & Wegener, 2016). In this approach, a study strengthening the case might be viewed as “replicating,” or at least “not failing to replicate,” whereas a study weakening the meta-analytic case might be viewed as a “failure to replicate.” This type of approach could also be extended to Bayesian approaches, where replication data that strengthen the prior for the next analysis could be viewed as “replicating” the previous results, whereas replication data that weaken the prior for a following analysis could be viewed as “failing to replicate” the previous results. For the current discussion, rather than dealing with marginal cases in which one might disagree about whether or not a given study counts as a successful or failed replication, we will assume clear cases in which a given study either supports an original result (e.g., with significant results in the same direction as the original) or opposes it (e.g., with near zero or even opposite results to those of the original study).

Statistical Conclusion Validity

Defining Statistical Conclusion Validity

Statistical conclusion validity refers to the extent to which a conclusion regarding a relation among variables is accurate (i.e., claiming there is an effect when there really is an effect, or concluding that there is no effect when there really is not; Cook & Campbell, 1979). The term relation is used broadly to refer to a wide range of statistical indices (e.g., a correlation coefficient, a test of differences between means across experimental conditions) and can involve assessments of simple bivariate relations as well as more complex relations (e.g., an interaction involving two independent variables and the dependent variable). As is discussed in any introductory statistics textbook, there are two fundamental types of error in statistical conclusion validity: Type I error (i.e., falsely concluding that a relation exists when there is no relation) and Type II error (i.e., falsely concluding that no relation exists when in fact a relation is present).
Statistical Conclusion Validity and Failures to Replicate

In situations in which a replication study leads to a different conclusion than the original study regarding the existence of a relation, one potential explanation is that the two studies differ in their levels of statistical conclusion validity. In the contemporary replication literature, the most common form of mismatch has been for the original study to suggest the existence of a relation and the replication study to fail to produce evidence for such a relation, though the opposite is logically possible. In the case of failing to replicate an initial effect, this pattern could be explained by the original study having produced a Type I error and the replication study having correctly failed to find evidence of a relation. Alternatively, the inconsistency in results could be because the original study correctly suggested the existence of a relation, and the replication study produced a Type II error.

Statistical Conclusion Validity in the Replication Literature

Assuming that a replication study, whether conceptual or exact, has failed to produce evidence of a previously demonstrated effect, the critical question then becomes why such a failure has occurred. In the contemporary replication literature, assuming the replication study has adequate statistical power (a determination that is more complicated than is often appreciated; see Maxwell et al., 2015; Pek & Park, 2018), an extremely common answer to this question has been to postulate that the original demonstration of the relation was a Type I error. That is, a common interpretation has been that the discrepancy in results is a function of the original study having been lower in statistical conclusion validity than the replication study. Indeed, arguably the most central theme of the contemporary replication literature has been that much of the difficulty in replicating prior research has been a function of the manner in which data have been collected, analyzed, and reported having undermined the statistical conclusion validity of original psychological research.

Flawed Statistical Practices and Reporting

One prominent illustration of the centrality of statistical conclusion validity in the replication literature is the increased sensitivity to concerns regarding statistical power. Traditionally, statistical power has been a concern with respect to Type II error (e.g., Cohen, 1988). However, recently methodologists have noted the prevalence of studies with low power in the published literature and highlighted this problem as a potential cause for the publication of many false positive effects (e.g., Button & Munafo, 2017). Briefly stated, because studies with small sample sizes are likely to provide poorer estimates of population effect sizes, studies with low power will be more likely to produce outlier estimates (e.g., estimates that deviate substantially from the true population value in either direction). In addition, researchers who routinely conduct many studies with small samples rather than a few with large samples would have more opportunity to uncover significant effects by chance alone. In the context of a single study in which the researcher can argue that virtually any significant pattern of an effect is conceptually meaningful, this could lead to increased false positives in the literature that are accepted as valid effects.

Another example of the assumption that statistical conclusion validity plays an important role in low replication rates is the current debate regarding the appropriate alpha for statistical tests. Some have suggested that the traditional alpha of .05 is too lenient a standard. This concern has led to calls for more stringent alpha levels in original research (e.g., .005) as a means of producing more replicable findings (e.g., Benjamin et al., 2017; Greenland, Gonzalez, Harris, & Guthrie, 1996).

Still others have suggested that even assuming that data are properly collected and analyzed for individual studies, the statistical evidence for a given relation will be greatly exaggerated if researchers
only report those studies that produce a statistically significant effect in a given direction. These commentators argue that selective reporting of studies has substantially enhanced the likelihood of publishing Type I errors and, thus, been an important contributing factor to poor replication rates. Much of the emphasis in this literature has been focused on developing indices that can detect if the results of a set of studies are “too good to be true” and thus are likely to reflect biased reporting of studies (e.g., Francis, 2012; Schimmack, 2012; Simonsohn, Nelson, & Simmons, 2014).

**Questionable Research Practices**

Similarly, a great deal of attention has been accorded to “questionable research practices” (QPRPs; e.g., John, Lowenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). QPRPs cover a wide range of practices that overlap somewhat with the themes already discussed, but also include additional practices. Among the behaviors identified are reporting the results of only some dependent variables, reporting the results of only some conditions, and reporting the results of only studies that produced statistically significant demonstrations of the relation of interest. Other practices include data-driven deletion of observations in data sets and data-driven decisions regarding when to continue or stop the collection of data.

As these examples illustrate, many QPRPs are practices that are considered problematic because of their potential to seriously undermine the statistical conclusion validity of a study by inflating Type I error rates. Indeed, of the 10 QPRPs investigated by John et al. (2012), at least 7 can be clearly regarded as primarily concerns related to statistical conclusion validity. Moreover, although each can be considered potentially problematic on its own, they can pose especially serious threats to statistical conclusion validity when performed in conjunction with one another. Simmons et al. (2011) demonstrated in simulation data that, in the context of a single study in which the researcher can argue that virtually any pattern of an effect is conceptually meaningful, performing four such practices in conjunction with one another can lead to Type I errors occurring in nearly 61% of cases.4

**Alternatives to Traditional Statistical Approaches**

Finally, some have argued that poor replication rates are a function of inherent limitations in the statistical approach that most social scientists use when analyzing data. These critics have argued for adopting alternative approaches to traditional null hypothesis testing. For example, some have advocated that researchers should abandon traditional statistical tests and instead focus on reporting effect sizes and their corresponding confidence intervals (e.g., Cumming, 2014; Schmidt, 1996). Others have promoted the use of Bayesian statistics (e.g., Wagenmakers et al., 2017). Advocacy of both of these alternatives is predicated on the assumption that problems related to statistical conclusion validity are a major limitation of contemporary research and their resolution is critical to resolving concerns regarding replication rates. As noted next, we argue that the contemporary debate about replicability is insufficiently attentive to the other kinds of validity issue that could contribute to a failure to replicate.

**Internal Validity**

**Defining Internal Validity**

*Internal validity* refers to the extent to which a relation among variables can be interpreted as causal in nature (Cook & Campbell, 1979). Typically, in basic research in which a researcher can conceptually designate one or more variables as an independent variable (IV; or predictor variable) and another
variable (or set of variables) as a dependent variable (DV), the researcher is postulating a causal relation (i.e., the IV as operationalized produces the observed changes in the DV as operationalized). However, the extent to which this interpretation can be confidently accepted will depend on the degree to which design features of the study, and to a lesser extent the manner in which the data are analyzed, permit the researcher to rule out alternative interpretations of the relation.

Traditionally, when a relation between an independent variable and a dependent variable is observed, there are two alternative explanations to a causal effect of the IV on the DV. First, the causal direction of the relation could be the reverse (i.e., the DV could cause the IV). This, of course, is not plausible in an experiment where the IV precedes the DV and in which each person’s assignment to levels of the IV is determined by the researcher via random assignment. Second, in a nonrandomized study, the chief threat to internal validity is that there is some cause of the DV other than the IV. Cook and Campbell provide a catalog of such threats (e.g., history, regression).

**Internal Validity and Failures to Replicate**

As we have discussed, when a replication study leads to a different conclusion than the original study, one potential set of explanations for this discrepancy involves concerns about statistical conclusion validity. However, conceptually, statistical issues are only one of a subset of possible explanations. A second possible set of explanations is that the two studies differ in their degree of internal validity. The manner in which internal validity can potentially play a role in failures to replicate can, of course, be quite different, depending on whether the studies in question are experimental or nonexperimental in nature. For purposes of brevity, we will confine our current discussion to cases involving experimental studies where issues of internal validity are presumed to be a high priority.5

In situations in which the original study indicates the existence of a relation and the replication study fails to produce evidence for a relation, there are two possible ways in which differences in internal validity could create such a discrepancy between studies. First, it could be that one of the several threats to internal validity outlined by Cook and Campbell was responsible for the DV rather than the IV. For example, if the IV in the original study led to differential attrition (participant deletion) across experimental conditions, then the relation demonstrated by the original study should not be attributed to the IV. Presuming that the same differential attrition problem is not present in the replication study, this spurious effect that emerged in the original study would not be expected to emerge in the replication. Thus, the discrepancy in results could be a function of the original study having been low in internal validity and the replication study having been higher in internal validity.

A second explanation could be that the original study had no internal validity problem, in that the IV actually does have a causal impact on the DV. However, if in a later replication study the IV leads to differential attrition across experimental conditions, this could produce a null effect or even a reverse effect. In this case, the discrepancy in results between studies would be a function of the original study having been higher in internal validity than the replication study.

**Internal Validity in the Replication Literature**

In the traditional methodological literature, internal validity has been a central concern. Indeed, a sizeable literature has accumulated regarding different design features that can help to enhance internal validity (e.g., see Brewer & Crano, 2014; Smith, 2014; West, Cham, & Liu, 2014). Despite its importance in the more general methodological literature, issues related to internal validity have played little role in the ongoing debate regarding replication. None of the major themes that have emerged in this literature has been closely tied to issues of internal validity. That being said, there is
a clear conceptual basis to expect that issues of internal validity could contribute to some failures to replicate previously demonstrated findings.

For example, changes in the context in which the experiments were run could sometimes lead to differential attrition rates. For instance, the original Isen et al. (1978) experiment was run in a shopping mall and involved a second experimenter approaching participants after the mood manipulation to secure their participation to take part in the study. If the mood manipulation had an impact on willingness to participate in the study, this could introduce a violation of random assignment, thereby creating a potential alternative cause to the IV. If this study were later replicated in a laboratory setting with participants in a standard subject pool, differential attrition might be less likely, as nearly all participants would be expected to participate. Or, more commonly, if an initial laboratory study is replicated in an online sample environment such as MTurk, the independent variables from the original study might induce differential attrition that can be difficult to detect (Zhou & Fishbach, 2016).  

Construct Validity

Defining Construct Validity

As outlined previously, basic consumer psychology researchers are primarily interested in relations between psychological constructs rather than relations between particular operationalizations of those constructs. However, in order to study constructs, researchers must use particular manipulations or measures of those constructs. These manipulations or measures can vary in terms of how closely they correspond to the construct of interest. Construct validity refers to the extent to which the operationalizations of the IVs and DVs in a study correspond to the constructs that a researcher is trying to study.

Construct Validity and Failures to Replicate

Differences in construct validity across studies constitute yet another set of explanations for why a replication study might fail to reproduce an effect demonstrated in an earlier study. In fact, this might be the most underappreciated reason. When researchers use materials in a replication study that do not capture the constructs of interest as well as in an original study, failures to replicate should generally become more likely. These failures to replicate would reflect that the researcher did not successfully manipulate and/or measure the constructs of interest; it would not necessarily indicate a lack of relation between the constructs themselves. Indeed, even if a replication attempt demonstrated the same pattern of results as the original study, if the construct validity of the IV and/or DV is lower than in the original research, the researcher would want to be very careful in interpreting such findings, as they might reflect relations between constructs other than those of interest to the researcher. Importantly, threats to construct validity can occur either when using the same materials as those used in an original study or when using different materials. Though it might often be reasonable to start with the original materials and use them in a replication study, the meaning of the particular operationalizations (i.e., the relation between the operationalization and the construct) could differ across populations, times, and situations (cf. Stroebe & Strack, 2014). For example, there is no guarantee that the particular gifts given by Isen et al. (1978) would create positive mood in 2018, and, if the requisite mood states are not created, then the replication study is not actually testing the hypothesis that guided the original research. Instead, the replication could be viewed as rather irrelevant to tests of the hypothesis (just as a test of brand new materials that have no relation to the constructs of interest would be viewed as irrelevant). Or, the gifts could produce the same moods, but not to the same degree as in the original study, reducing the likelihood of obtaining the same effect.
One possible example of this type of situation relates to studies that have associated weather with life satisfaction. In one of the earliest such studies, Schwarz and Clore (1983) phoned research participants on one of the first sunny days in the spring or on a rainy day. When not specifically reminded of the weather, research participants reported greater life satisfaction on the sunny days than on the rainy days. According to Schwarz and Clore (1983) this was expected to occur because people should be in a better mood on sunny than rainy days (at least in this location and at this time of year). Indeed, the other study in the paper did not address weather at all. Rather, it used a more direct manipulation of mood (i.e., recall and description of an event that made the person feel “really good” or “really bad”). Moreover, in both studies, manipulation checks verified mood differences between the conditions. However, a number of replication efforts using larger and more representative samples have shown much smaller and sometimes nonsignificant effects of weather on life satisfaction (e.g., Feddersen, Metcalfe, & Wooden, 2016; Lucas & Lawless, 2013). Yet, “failure” studies have generally focused on weather per se, rather than measures of mood (e.g., Lucas & Lawless, 2013). If examined in contexts where the weather differences are not associated with mood differences, however, the study is not examining the same conceptual hypothesis that guided the Schwarz and Clore (1983) research. Consistent with a role for mood, some of the larger-scale studies have found weather effects on life satisfaction in the same places where weather influenced net affect variables (but not when weather was not associated with affect; e.g., Connolly, 2013). Variation across studies in whether the weather variables are or are not related to the psychological construct of mood would seem to be crucial in determining the relevance of a given study to the hypothesis tested by Schwarz and Clore (1983).

As demonstrated by these examples, when two studies differ in construct validity, even if they use the very same materials or the exact same measures presumed to capture the variables of interest, one would not necessarily expect to find the same effects in the replication as in the original. Any discrepant results could stem from the differences in the psychological constructs that are involved, thereby muddying the implications for the relations among the constructs of interest. Thus, replications that fail owing to problems with construct validity should not necessarily be considered as evidence against a conceptual hypothesis of interest. Of course, if the initial researchers were interested in the particular operations (weather) rather than the underlying psychological construct (mood), then it would be appropriate to say that a study failing to find a sunny day effect was a failure to replicate whether or not it varied mood, because the original hypothesis was about the operation (weather) and not the psychological construct (mood).

The Role of Construct Validity in the Replication Literature

Interestingly, construct validity has received relatively little attention in recent discussions of replication. That is, in contrast to the prominence of statistical issues in the discussion, comments related to construct validity have been fewer and farther between (for exceptions, see Fabrigar & Wegener, 2016; Finkel et al., 2017; Stroebe & Strack, 2014). In the context of replication, perhaps the most explicit discussion of construct validity to date has revolved around the issue of psychometric invariance (Fabrigar & Wegener, 2016), but construct validity has also appeared in discussions of so-called exact versus conceptual replications (Crandall & Sherman, 2016).

Psychometric Invariance

Fabrigar and Wegener (2016) coined the term psychometric invariance to refer to instances in which two studies have independent and dependent variable operationalizations that do not differ in their psychometric properties. Put another way, the operationalizations in each experiment are equally
good at representing the same underlying constructs (see also, Wegener & Fabrigar, 2018). When attaining psychometric invariance is the goal in a replication effort, it can be said that the researchers are not conducting an exact operational (or direct) replication but an exact conceptual replication (Petty, 2018). Although construct validity is included under the conceptual umbrella of psychometric invariance, psychometric invariance is intended to be a broad concept that encompasses a larger set of psychometric properties including, but not limited to, reliability, convergent validity, discriminant validity, and predictive validity.

Whether the researchers state their assumptions or not, conduct of a non-experimental study generally involves assumptions that the measures of the independent (predictor) and dependent variables each capture an underlying construct of interest, and that the independent variable has a direct (causal) influence on the dependent variable. If a researcher were to replicate this study, the replication would be based on an assumption that the measures capture the same constructs in the replication study as in the original study before it could even test whether the relation between the independent and dependent variables is the same across studies. Therefore, if the factor structure of the scales were different across the original study and the replication study, it would be unclear whether any lack of replication of the IV→DV relation occurred because of an unreliability of the original finding or because the scales reflected different constructs in each of the studies. Indeed, within the measurement literature, researchers have illustrated the importance of factorial invariance. For example, researchers caution against comparing two groups that show differences in the factor structure because the measure for each group may not capture the same construct (see Millsap & Meredith, 2007; Widaman & Grimm, 2014; Widaman & Reise, 1997). Indeed, even if the effect replicated, if the scales had different structures, it would be unclear whether the replication did in fact reflect the same relation between constructs.

When conducting an experiment, in addition to measurement assumptions regarding dependent measures, the researcher assumes that the manipulation of the IV reflects the underlying construct of interest. When comparing an original and a replication study, just as in the measurement example, factorial invariance across studies would need to hold for the scale of the dependent measure. In the case of the manipulated IV, however, psychometric invariance takes a slightly different form. The manipulation is intended to create a difference on some construct of interest (predictive or convergent validity) but is intended not to create differences in other variables that could affect the dependent variable of interest (discriminant validity). Typically, strong manipulations would have a large effect on the intended construct, but no, or a relatively small, effect on unintended constructs.

For example, in the Isen et al. (1978) mood study described earlier, one would have wanted the mood manipulation to affect feelings but not to influence other variables that might affect perceived qualities of the target products, such as activation of beliefs that would be applicable to the target judgments (which may be one reason Isen et al., 1978, used target products that were quite distant conceptually from the types of gift given to create positive mood). If a replication study gave participants gifts that did not affect recipients’ moods, or if the beliefs about the gifts related directly to qualities of the target product, relevance of the replication study to the original hypothesis would be called into question.

Fabrigar and Wegener (2016) argued that direct comparisons between an original and a replication study assume psychometric invariance (or they should assume this). In terms of the current construct validity discussion, the assumed invariance would include that the manipulation of the IV (in an experiment) influences the same construct (and only the same construct) as in the original research, and that the measures (of the IV in a measurement study, and of the DV in both measurement studies and experiments) tap the same constructs across studies. If there are differences in any of these properties of the IV or DV, any failure of replication could be because
different constructs are involved rather than constituting evidence against the IV→DV relation identified in the original research.

In addition, psychometric invariance assumptions also relate to the strength of the impact of the manipulation on the construct representing the IV. For example, as noted earlier, it may be the case that a particular manipulation of a construct still influences that construct in a replication study, but it does so to a lesser extent than in the original research. Although the replication study is using a manipulation that influences the same construct as in the original, if it creates less of a change in the IV-related construct, one would naturally expect less of an impact on the DV. If one thinks of the weaker manipulation as creating less variation in the IV-related construct, one might think of the two studies as differing in the level of construct validity, but such a pattern could also be thought of as related to statistical conclusion validity. Related to statistical concerns, the replication study could produce a smaller effect of the IV and fail to reach significance simply because the impact of the manipulation on the construct of interest was too small. This would mean that the replication study with no changes would require a larger sample size to detect the effect and would reveal a smaller effect size compared with the original study. But neither the smaller effect nor the larger necessary sample would reflect anything about the relation between psychological constructs other than that the manipulation of the IV was less effective in the replication than in the original research.

To date, although there has been a general recognition that failures to replicate could in part be a result of violations of psychometric invariance, there has been little attempt to systematically evaluate psychometric invariance in published replications (e.g., Ebersole et al., 2016; Klein et al. 2014; Open Science, 2015; Wagenmakers et al., 2016). However, subsequent empirical investigations of some of these replication efforts have revealed that some failures to replicate were in large part a function of the operationalizations used in the replication studies more poorly capturing the constructs of interest (e.g., see Ebersole et al., 2017; Luttrel, Petry, & Xu, 2017).

**Direct versus Conceptual Replications**

Direct or exact replications typically refer to instances in which researchers replicate a study using materials that are identical or as close as possible to the original study. Conversely, conceptual replications typically refer to replications using alternative operationalizations of the same psychological constructs. Although most researchers agree on the value of replication, there has been considerable debate about the appropriate roles of direct/exact versus conceptual replications.

Although not always explicit in this discussion, construct validity is an important part of this debate. Any particular instantiation (manipulation or measure) of a construct is likely to have “irrelevancies” (i.e., the influence of constructs other than the intended construct; Cook & Campbell, 1979). Thus, advocates of conceptual replications have argued that, in order to ensure that one’s effect is driven by the construct of interest rather than irrelevancies (confounds) in a particular operationalization, it is useful to demonstrate that a given effect occurs across different types of manipulations and measures (e.g., Brewer & Crano, 2014; Schmidt, 2009; Stroebe & Strack, 2014). Successful conceptual replications therefore give researchers more certainty that an effect actually reflects relations between the constructs of interest.

Despite the benefits to construct validity, critics of conceptual replications have argued that failures to replicate with conceptual replications are open to so many interpretations that they are uninformative (e.g., LeBel & Peters, 2011; Nosek, Spies, & Motyl, 2012; Pashler & Harris, 2012; Simons, 2014). This problem is an extension of the problem that there are generally more potential explanations for any null effect than there are for a significant effect. Critics of conceptual replication have
typically preferred exact replications, in which researchers try to use materials that are the same or as close as possible to those in the original study. This approach is exemplified by the Reproducibility Project (Open Science Collaboration, 2012). A defining feature of this project was “use of original materials, if they are available” (p. 658). Supporters of exact replications have argued that, because exact replications follow the original procedure as precisely as possible, they are more informative about the reliability of the effect. What exactly “the effect” is, however, can sometimes be less than clear. For example, as described earlier, “the effect” is often treated as the observed relation between the particular manipulation or measure that take the place of the predictor or IV and the measure that is treated as the outcome or DV (e.g., an observed association between weather and life satisfaction). However, in the case of basic research at least, the primary hypothesis that guided the original research was not about the particular instantiation per se (e.g., the weather), it was about the psychological constructs that the instantiation(s) captured (i.e., a person’s mood).

Some commentators have argued that conducting a true exact replication would be nearly impossible, as it would require the same subjects at the same time in the same context (e.g., Fabrigar & Wegener, 2016; Petty & Cacioppo, 2016; Stroebe & Strack, 2014). Each of these is unlikely to occur, so even exact replications are subject to psychometric invariance concerns. For example, conducting a replication of the Isen et al. (1978) study today using the same materials as the original study would not ensure that the two studies had psychometric invariance. Similarly, if a researcher had conducted an exact replication in the years following the Isen et al. (1978) research but had done so in a different country or even just a different part of the US, psychometric invariance might well have been violated. Because of this, even exact replication failures are open to multiple interpretations: They might indicate that the original findings are not reliable, but they might also indicate that the operationalizations are less reflective of the constructs of interest in the replication compared with the original study (in addition to many other reasons for null effects that might apply equally well to both exact and conceptual replications). Thus, although there have been various calls for original researchers to provide “replication recipes” that detail the specifics of the operations used (Zwaan, Etz, Lucas, & Donnellan, 2017), in basic research it can sometimes be even more useful to provide a recipe that details the specifics of the psychological constructs involved (Petty, 2018).

An additional feature of many of the calls for exact replications involves that they should be conducted by an independent lab. For example, Roberts (2015, n.p.) stated that, “The gold standard for our science is a pre-registered direct replication by an independent lab.” Indeed, reliable effects should not depend on who conducts the research or where it is conducted, if the appropriate measures and/or manipulations are used. The logic of wanting independent replications seems to rest on the idea that this would remove bias or dishonesty on the part of the replicator. Yet, it remains unclear whether independence of the replicator has resulted in less biased tests. For example, because an independent replicator will necessarily be running a study in a new context, it is possible that the emphasis on using the same materials (that were developed and pretested for use in the original sample and context) could increase the likelihood of a construct validity threat. Indeed, large-scale replication projects such as the Many Labs effort (e.g., Ebersole et al., 2016) might be particularly vulnerable to psychometric invariance problems, as researchers should not expect particular operationalizations developed in one particular sample and context to have the same psychological meaning and effects across different samples and contexts. In many cases, it could be that development of measures and manipulations appropriate to the sample and setting would provide the more informative replication data.

Finally, though not described as a construct validity concern, in their “unified framework to quantify the credibility of scientific findings,” LeBel et al. (2018, p. 394) noted that replicators should consider the evidence for how likely “auxiliary hypotheses” are to be true. Under the umbrella of
“auxiliary hypotheses,” they included measurement instruments “operating correctly.” One key aspect of measurement instruments operating correctly would reflect assessments of construct validity (though, from a psychometric invariance perspective, such issues would extend to manipulations as well as measures).

## External Validity

### Defining External Validity

*External validity* is the extent to which the results of a study can be generalized to other situations and other people (Cook & Campbell, 1979). In considering whether the results of a study can be said to generalize, it is important to keep in mind the fundamental goals of the study. In the case of an applied study, the extent to which the relation between a specific operationalization of an IV and a specific operationalization of a DV generalizes to other settings and people might well be of interest. However, in the context of basic research, the interest is usually in the extent to which the relation between the psychological constructs of interest generalizes rather than the specific operationalizations of these constructs. Indeed, as we have already noted, operationalizations are often developed to maximize construct validity within a particular population and/or setting and would not be expected or necessarily intended to broadly generalize (e.g., see also Fabrigar & Wegener, 2016; Stroebe & Strack, 2014).

For example, as we have noted with our example of the Isen et al. (1978) experiment, one might not expect the specific gifts used in that experiment to induce positive mood in all populations (nor would a sunny day be as happiness-inducing in Southern California as it would in the more overcast Midwest). Were that gift manipulation to fail to show an effect on subsequent product evaluations in a different population, this lack of an effect might not necessarily reflect a lack of external validity in the theoretical sense (i.e., it might not imply that the fundamental nature of the relation between the construct of positive mood and product evaluations differ across these two populations). Rather, it might well be that the relation between the constructs of positive mood and evaluation is invariant, and that the mood manipulation is simply a poorer operationalization of mood (i.e., lower in construct validity) in one population/setting than the other. As such, it can sometimes be difficult to distinguish between violations of psychometric invariance (a construct validity issue) and differences in studies due to external validity. That is, although a replication study result might make it clear that the original operationalization did not generalize to a new sample (a seeming external validity issue), the study may not have shown that the psychological construct did not generalize because that construct may not have been present in the new sample.

## External Validity and Failures to Replicate

When a replication study fails to produce evidence for an effect that was demonstrated in the original study, external validity can provide another very straightforward explanation for the discrepancy. Specifically, even assuming that both the original study and the replication study were high in statistical conclusion validity, internal validity, and construct validity, it is possible that the two studies differed on characteristics of the participants or the context. If any of these individual differences or contextual factors had an impact on the relation between the IV and DV (i.e., moderated the effect of the IV on the DV), one would not expect the replication study to produce the same effect as the original study. That is, one would not expect the effect demonstrated in the original study to generalize to the new study. Stated another way, one might say that the original study lacks external validity with respect to the new population and/or context.
External Validity in the Replication Literature

Within the traditional methodological literature, issues of external validity have long been discussed. Indeed, commentators have often expressed concerns regarding the lack of diversity in research participant samples and its potential distorting effects on the external validity of many findings in psychological research (e.g., Henrich, Heine, & Norenzayan, 2010; Sears, 1986). Likewise, methodologists have also long debated the merits of different data collection contexts (e.g., laboratory versus naturalistic settings) for exploring psychological phenomena and the impact that such settings might have on external validity (e.g., see Brewer & Crano, 2014).

Some of these discussions might lead readers to think that psychologists have placed comparatively little emphasis on external validity. However, when viewed from a different perspective, one might argue that external validity has been central to many areas of psychology. Specifically, over the past several decades, researchers have shown great interest in testing moderator effects (e.g., Hayes, 2013; Judd, Yzerbyt, & Muller, 2014). That is, researchers have been interested in whether the effects of a given IV are conditional on levels of a second (and sometimes third) IV. Explorations of such moderator effects have often been a key component of theory testing. Although not usually conceptualized as such, the investigation of moderators is inherently an exploration of the external validity of a given phenomenon.

In light of the prominence of the investigation of moderators in many areas of psychology, it is not surprising that discussions of the reasons for failed replications have often acknowledged the potential role of unknown moderators in accounting for differences across original studies and their replications (e.g., see Barsalou, 2016; Cesario, 2014). Although some of the factors identified as potential explanations for differences across studies in these discussions might be regarded as violations of psychometric invariance, other potential moderators discussed would clearly be regarded as more fundamental differences in the relations among psychological constructs.

Along similar lines, many published replications have at least in a general sense acknowledged the possible role of moderators in explaining differences between original research and subsequent replications (e.g., Ebersole et al., 2016; Klein et al. 2014; Open Science, 2015; Wagenmakers et al., 2016). However, as a rule, replication researchers have generally not attempted to systematically specify potential moderators that might explain the differences or empirically evaluate the viability of these moderators as explanations for differences in results. Moreover, specific procedures for identifying and investigating moderators as explanations for nonreplication have not been incorporated into the standard practices of replication initiatives. In a few cases, independent research teams have followed up on failed replication attempts by exploring potential moderators (e.g., Noah, Schul, & Mayo, 2018). Thus, in the current replication literature, external validity is an often acknowledged, but seldom studied, cause of failed replication.

Validity Trade-Offs: Observations and Conclusions

The Prominence of the Validity Types in the Replication Literature

As our review has illustrated, any time a replication study fails to produce the same results as the original study, there are at least four possible sets of explanations for why this failure to replicate might have occurred. These four sets of explanations are not mutually exclusive. In many cases, failures to replicate could be a function of more than one set of causes. However, in our review, we have also noted that the attention accorded to each of these four potential explanations has not been equally distributed within the replication literature. Statistical conclusion validity has been the focal explanation for and proposed solution to ongoing replication concerns. The potential roles of construct validity and external validity have been acknowledged, but systematic explorations of these
explanations have not been considered essential for publishing replications. Likewise, calls for methodological reform and data collection initiatives have not featured concrete protocols for addressing construct validity and external validity as central features in their plans. Finally, for the most part, the potential role of internal validity has been ignored.

One might be tempted to think that our narrative review has exaggerated the asymmetry that exists in attention to the four types of validity within the replication literature. To better assess this possibility, we conducted a more systematic content review of the contemporary replication literature. We identified various books, special issues of journals, and special sections of journals focusing on replication issues. In total, we examined the content of 88 journal articles and book chapters published in eight distinct special issues/sections of journals and one book. For each publication, we examined whether it focused at least in part on issues directly related to each of the four types of validity, and, if so, if there was a predominant focus on one of the types of validity.

A summary of our content analysis is provided in Table 26.1. As can be seen in Column 1, if one simply examines the extent to which issues related to each of the four types of validity are discussed in these publications, there is a clear emphasis on statistical conclusion validity. Approximately 78% of the publications at least in part addressed issues related to statistical conclusion validity. In contrast, 31% touched upon issues related to external validity, and about 16% in some form discussed issues related to construct validity. Only 7% mentioned issues related to internal validity, and 7% made no mention of issues related to any of the four types of validity.

If one focuses on the predominant emphasis of each publication (see Column 2 of Table 26.1), the picture is very similar. Nearly 61% of publications predominantly focused on statistical conclusion validity. About 9% of publications primarily addressed external validity, and approximately 8% were mostly directed at issues related to construct validity. Internal validity was the major emphasis of only about 5% of publications, and the remaining 17% did not have any of the four types of validity as their central theme (e.g., simply advocating for more frequent replication studies without discussing validity issues). Yet, despite the dominance of discussion of statistical validity issues, it is not clear that this type of validity is the leading cause of replication failure. It could be, of course, but no data exist on this yet.

Another way to explore this topic is to examine the submission guidelines of major journals. Reforms to journal submission guidelines have been one of the major ways that consumer psychology and related disciplines have attempted to address concerns regarding replication. To assess the prominence of concerns regarding each type of validity, we examined the guidelines of seven major social-personality journals and six major marketing journals. We coded each guideline with respect

Table 26.1 Frequency of Replication Articles Mentioning, or Having a Primary Focus on, each of the Four Major Validities

<table>
<thead>
<tr>
<th>Validity Type</th>
<th>Validity Type is Mentioned*</th>
<th>Validity Type is Primary Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>78% (n 69)</td>
<td>61% (n 54)</td>
</tr>
<tr>
<td>External</td>
<td>31% (n 27)</td>
<td>9% (n 8)</td>
</tr>
<tr>
<td>Construct</td>
<td>19% (n 14)</td>
<td>8% (n 7)</td>
</tr>
<tr>
<td>Internal</td>
<td>6% (n 6)</td>
<td>5% (n 4)</td>
</tr>
<tr>
<td>None</td>
<td>6% (n 6)</td>
<td>17% (n 15)</td>
</tr>
</tbody>
</table>

Notes: * Because some publications could address more than one type of validity, these percentages would not be expected to sum to 100%.


Additional sources: Psychological Science under Scrutiny (ed. Lilienfeld & Waldman).

497
to whether it was directly related to each of the four types of validity. In the case of social–personality journals, we examined a total of 77 guidelines. As Column 1 of Table 26.2 shows, 45% of guidelines were directly related to statistical conclusion validity (e.g., a guideline regarding sample size and statistical power). For the remaining three validities, only about 1–3% of guidelines addressed internal validity (e.g., a guideline for appropriate causal language when describing cross-sectional mediation analyses), construct validity (e.g., a guideline regarding discussions of the representativeness of experimental stimuli), or external validity (e.g., a guideline regarding the extent to which an effect has been validated across cultures, labs, or samples). Likewise, of the 39 guidelines examined in marketing journals, 26% related to statistical conclusion validity, whereas the other three types of validity were not represented in these guidelines.

In summary, our more formal content analyses support the conclusion of our narrative review. The overwhelming emphasis in the replication literature has been on issues related to statistical conclusion validity. As a result, the fields of consumer psychology and related disciplines have at least implicitly appeared to adopt an assumption that statistical issues are the central cause of failures to replicate and, as such, should be the major solution to enhancing replication. In comparison, construct validity, external validity, and internal validity have at least implicitly been treated as minor or even trivial factors in failures to replicate and advocated actions to enhance replication.

Is the Emphasis on Statistical Conclusion Validity Justified?

Conceptual Justification

Given the overwhelming emphasis on statistical conclusion validity concerns, a natural question is whether this emphasis is justified. On purely conceptual grounds, there does not appear to be a compelling case for assuming that the vast majority of failures to replicate are a result of threats to statistical conclusion validity. As we have seen, all four types of validity have long been regarded as central dimensions for evaluating psychological research, and there is a clear logic for how each of these validities could play a role in failures to replicate. Indeed, the potential roles of construct validity and external validity have been generally acknowledged in the replication literature. However, this recognition of their potential relevance has not translated into these issues being incorporated into reforms for enhancing the replicability of original research or as explanations that must be empirically evaluated in published replications. For instance, although published replications are routinely required to meet certain standards of statistical conclusion validity (e.g., they are routinely required to provide a clear rationale for why their studies should be considered to have adequate power), they are rarely required to provide formal demonstrations that standards of construct validity

<table>
<thead>
<tr>
<th>Validity Type</th>
<th>Social–Personality Journals</th>
<th>Consumer Psychology Journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>45%</td>
<td>26%</td>
</tr>
<tr>
<td>External</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>Construct</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Internal</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>48%</td>
<td>74%</td>
</tr>
</tbody>
</table>


have been met (e.g., they rarely have to present evidence regarding the factor structure of measures or pretests to confirm that manipulations properly capture the constructs of interest). Nor have they been required to compare the participant samples and contexts used in terms of how closely they mirror or differ from the psychological characteristics of the original research.

It is not readily apparent why one form of validity would be considered essential to address in replication whereas the others are not. Advocates of methodological reforms to enhance replicability of studies and researchers conducting replications have not explained why statistical conclusion validity has been emphasized over the other forms of validity. Thus, if there is a clear conceptual basis to argue for the primacy of statistical conclusion validity in replication, as yet it has not been explicitly articulated.

**Empirical Justification**

In the absence of a compelling theoretical argument for assuming that statistical conclusion validity is the primary cause of failures to replicate, perhaps one might argue that there is an empirical basis for this emphasis. For example, one might argue that replication efforts provide some basis for considering statistical conclusion validity to be an important determinant of replicability. In some settings, original studies with power-related characteristics, such as larger effect sizes/smaller standard errors/narrower confidence intervals, have replicated at a higher rate than original studies with fewer such characteristics (e.g., Open Science Collaboration, 2015).

Such findings are certainly consistent with the notion that statistical validity plays an important role in replication. As we have noted, there is a good conceptual basis to expect it is part of the answer. However, the strength of these findings is far from sufficient to provide a basis for focusing exclusively on it and ignoring other possibilities. Moreover, such comparisons have represented far from pure comparisons, as a number of these characteristics have been confounded with the content of the research (which is also associated with the extent to which contextual moderation would be expected—a external validity explanation; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016).

Alternatively, one might point instead to the mathematical analyses linking lack of power to the prevalence of false positive results in the literature (where a large presence of false positives could produce high proportions of study results that do not replicate; e.g., Pashler & Harris, 2012). Yet, these discussions generally make assumptions that limit their applicability to many of the contexts in which research results are often presented (for a critique of these assumptions, see Stroebe, 2016). For example, in these analyses, particularly high false positive rates (e.g., > 50%; Pashler & Harris, 2012) occur only when relatively low levels of power are combined with a very high prior probability of the null hypothesis being true (i.e., 90%) and every study receiving a publication decision in isolation. One major point of most introduction sections in psychological outlets is to provide a compelling rationale for why predicted effects should be expected to occur (i.e., for why null effects should be unlikely rather than likely). In addition, previous arguments have suggested that, at the population level, no null hypothesis is actually true (e.g., Jones & Tukey, 2000). As a result, many methodologists would question the utility of analyses assuming high probabilities of true null hypotheses. Even with rather low levels of power (i.e., 35%), if tested null hypotheses reach even moderate probabilities of being false (e.g., 50%), the false positive rate remains much lower (i.e., 13%) than the “crisis-level” numbers that are generally emphasized in replication-related discussions (cf. Pashler & Harris, 2012).

Even if one were to grant high prior probabilities of true null hypotheses being tested, commentators have not addressed how the obtained proportions of false positives should be applied to a literature that generally consists of multi-study publications. Even if one estimated a high prior probability of a true null hypothesis (because little theory or previous research supports the hypothesis) and low power to reject the null hypothesis, if only half of the obtained rejections represent false positives,
what is one to do, for example, with a set of four studies each reporting significant results? If any of the studies represent “true positive” effects, then it would not seem logical to treat the remaining “demonstrations” of the same relation as false. Moreover, false positives should be just as likely to fall in either direction, so obtaining a number of rejections in one direction with none in the other direction should be a good indication that the population effect is not null (even if one started by believing that the null might likely be true and the power of each individual study is less than some have advocated). As a result, one might want to know how many additional studies have tested the effect and whether any of those studies have produced significant effects in the opposite direction (as should happen if the population effect is zero and the current significant results represent only false positives). However, the rationale for questioning single studies with low power would not seem particularly applicable to sets of studies, even if each individual study has power that is less than remarkable.

In summary, the empirical case for considering statistical conclusion validity as the strongest (and perhaps even the dominant) contributing factor to failures to replicate is not yet convincing. Additionally, although investigations of the role of the other types of validity in failures to replicate are rare, evidence has begun to emerge suggesting that it is unwise to treat these other forms of validity as minor or trivial factors in replication. For example, Ebersole et al. (2016) attempted and failed to replicate the findings of Cacioppo, Petty, and Morris (1983) linking individual differences in need for cognition (NC) to levels of processing of persuasive messages. The original finding was that participants’ motivation to think (measured by the NC scale) interacted with the quality of arguments provided (manipulated by providing participants with arguments that had been pretested to be either strong or weak). Participants’ attitudes were more influenced by the quality of arguments when they were high rather than low in NC.

In response to the failure to replicate, Petty and Cacioppo (2016) argued that there were fundamental problems in the methods used to replicate that could explain the differences in findings. These issues included use of a suboptimal (shortened) scale to measure NC and a weaker manipulation of argument quality (AQ; that was not pretested to match the manipulation used by Cacioppo et al., 1983). Both of these could represent problems with construct validity in the replication study. In addition, whereas Cacioppo et al. (1983) examined a context in which the topic was not particularly high or low in personal relevance to the message recipients, Ebersole et al. (2016) used conditions that had previously been shown to produce high levels of personal relevance, thereby increasing baseline levels of processing across all participants, regardless of their level of NC. Furthermore, they used messages that were shorter than in the original research, also affecting motivation to think. These design features would represent a failure to take known moderators into account, an issue at least partly of external validity. Using such methods, even increasing power of the replication to an extremely high level, would not enhance its likelihood of replicating the original result. Moreover, increasing power of the original research would not have enhanced the likelihood of Ebersole et al. (2016) replicating when Ebersole et al. chose to use methods of unknown (but potentially low) construct validity and chose to test the effect in a context that past research had shown would reduce the likelihood of replication.

In order to empirically evaluate Petty and Cacioppo’s (2016) explanations for nonreplication, Luttrell et al. (2017) conducted a replication of the failed replication, as well as a replication of the original study. That is, they directly compared the materials used by Ebersole et al. (2016; i.e., the same NC measure, AQ manipulation, and high-relevance context) with those that would be optimal for finding the NC by AQ interaction (i.e., a more reliable NC scale, stronger AQ manipulation, and a lower-relevance context with longer messages). Luttrell et al. (2017) successfully replicated Cacioppo et al. (1983) when using the optimal materials but also replicated the failure to replicate when using the nonoptimal materials used by Ebersole et al. (2016). Then, Ebersole et al. (2017) conducted an independent replication of Luttrell et al. (2017), confirming that the materials they used in their attempted replication of Cacioppo et al.’s (1983) finding were not as good for obtaining the effect as were the optimal materials identified by Luttrell et al.
Luttrell et al.'s (2017) work constitutes one concrete example in which it was demonstrated that construct validity and external validity were much more likely explanations for failures to replicate than statistical conclusion validity. More recently, others have explored the role of nonstatistical explanations for failures to replicate other phenomena. For instance, Noah et al. (2018) examined potential explanations for Wagenmakers et al.'s (2016) failure to replicate Strack, Martin, and Stepper's (1988) study examining the facial-feedback effect using the “pen-in-mouth” paradigm (i.e., placing one’s face into a smile can influence how funny cartoons seem). They argued that the presence of a camera in front of participants during the replication experiments constituted an important contextual difference from the original Strack et al. (1988) study (which did not have a camera present). Consistent with this view, in their experiment, the facial-feedback effect was obtained when no camera was present, but failed to emerge when a camera was present. Thus, the Noah et al. (2018) experiment suggests that external validity might be a better explanation for the discrepancy between Strack et al. (1988) and Wagenmakers et al. (2016) than statistical conclusion validity.

**Consequences of Assuming the Primacy of Statistical Conclusion Validity**

A final question that naturally arises is what the implications are of assuming that statistical conclusion validity is the primary reason for failures to replicate? Before turning to those implications, it is important to note that high statistical conclusion validity is a desirable property of research that is important to consider. Statistical conclusion validity concerns might sometimes play a role in failures to replicate and probably do in a good number of cases. We are not arguing that statistical conclusion validity is unimportant in either original studies or replication studies. But the question is whether an emphasis on this form of validity with a relative lack of emphasis on the other forms of validity is problematic.

One answer to this question might be that even if statistical conclusion validity is only one piece in the replication puzzle, surely addressing this piece in an effective way must be regarded as a wholly positive development in the field. Even one step forward in a four step journey should be considered meaningful progress. To quote Lee Corso, a football commentator for ESPN, “not so fast.” Even setting aside the objection that some of the current statistical recommendations in the literature might not actually enhance statistical conclusion validity, such a conclusion assumes that the four types of validity are largely orthogonal. That is, it presumes that emphasis on one validity has no effect on how researchers address the other types of validity. This assumption is probably wrong in at least two ways.

**Trivializing alternative explanations**

One potential risk is that if researchers explicitly or implicitly adopt the assumption that statistical conclusion validity is the most critical determinant of replication success, it will discourage them from carefully considering the role of other validities when designing both original and replication studies. Likewise, when interpreting discrepancies between original studies and their replications, researchers might not seriously entertain alternatives to statistical conclusion validity and thus feel no need to systematically explore these alternative possibilities. In summary, researchers who have “checked the boxes” on the statistical conclusion validity checklist might be overly confident in the inferential foundations of their studies.

At a more extreme level, researchers might be inclined to think that high statistical conclusion validity can in some way “immunize” their studies against threats to other types of validity. For example, they might think that, even if their replication efforts used operationalizations that have not captured the constructs of interest as well as the original study, perhaps this limitation can be offset by simply using a much larger sample size. Unfortunately, as Cook and Campbell (1979) illustrate throughout their discussion of forms of validity, for the most part, being high on one type of validity generally affords at best modest and sometimes no protection against threats to another type...
of validity. For instance, if the key operationalizations represent different constructs than the ones they are intended to capture (low construct validity), no increase in statistical power can compensate for this limitation. Only if the operationalizations represent the same constructs as the original, but weaker, would enhancing sample size be a help. Moreover, even in these cases, if operationalizations are very poor, there might be no practically achievable sample size that is sufficient to offset the liabilities of poor construct validity. Likewise, if the phenomenon is examined in a context where it should not emerge (low external validity) or if a key IV is associated with a threat to internal validity (e.g., differential attrition), simply running a highly powered study will not solve these problems.

**Strategies for Maximizing One Validity Can Undermine Another**

More directly, as noted by Cook and Campbell (1979), there are often trade-offs in maximizing a study on each of the validities. That is, strategies that enhance validity on one dimension can often erode validity on another dimension. Thus, if researchers pursue strategies focused overwhelmingly on statistical conclusion validity, this can sometimes come at the cost of undermining other types of validity. If one or more of these other forms of validity determine replication rates in a given domain, efforts to increase statistical conclusion validity could lead to minimal increases or even decreases in replication rates. To illustrate how seemingly straightforward recommendations designed to enhance one type of validity can sometimes have unanticipated consequences for other types of validity, it is useful to consider an example. Specifically, many journals have implemented guidelines suggesting or 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 a quick and easy criterion for initially evaluating research. The goal of such guidelines would seem to be to increase power of published research and thereby enhance statistical conclusion validity. All else being equal, increased power is of course useful, but all else is rarely equal. There may be unintended consequences of such requirements.

For example, to the extent that researchers become focused on increasing their sample size as a way to increase power, researchers may have recruited participants and collected data in different ways, such as through relatively inexpensive online platforms such as Mechanical Turk. Such a move is not inherently problematic, and clearly these platforms have a useful role to play in psychological research. For example, increased diversity of the obtained sample via these platforms might be beneficial for examining whether results hold across different subpopulations of respondents, though such tests are only useful if sufficient numbers of members of the smaller groups are present in the sample. Unfortunately, simply having diversity of participant characteristics in the sample does not ensure external validity across the subpopulations, and that very same diversity could have potentially negative consequences for construct validity (see also Finkel et al., 2017). That is, if the sample diversity means that a given manipulation or measure differs in its meaning across participants, the construct validity of that manipulation or measure might be undermined.

Thus, although increasing sample size might, in concept, benefit statistical conclusion validity, researchers’ attempts to strengthen this validity might be a detriment to construct validity (and, if construct validity is threatened in this way, the overall effect could be weakened in such a way that power and the resulting statistical conclusion validity would not be helped either). One could certainly view this type of example as a good reason to think more broadly about statistical power, rather than focusing on sample size per se. For example, to the extent that researchers try to maximize power by using more precise measures and manipulations that are consistently interpreted by the participants in one’s study, the increased focus on power could actually benefit construct validity. Likewise, conducting a study in a controlled laboratory setting that minimizes distractions and allows for more precise control of extraneous factors could also increase the efficacy of experimental manipulations and measures, thereby enhancing construct validity as well as statistical conclusion
of validity. For instance, if the key operationalizations represent different constructs than the ones they are intended to capture (low construct validity), no increase in statistical power can compensate for this limitation. Only if the operationalizations represent the same constructs as the original, but weaker, would enhancing sample size be a help. Moreover, even in these cases, if operationalizations are very poor, there might be no practically achievable sample size that is sufficient to offset the liabilities of poor construct validity. Likewise, if the phenomenon is examined in a context where it should not emerge (low external validity) or if a key IV is associated with a threat to internal validity (e.g., differential attrition), simply running a highly powered study will not solve these problems.

**Strategies for Maximizing One Validity Can Undermine Another**

More directly, as noted by Cook and Campbell (1979), there are often trade-offs in maximizing a study on each of the validities. That is, strategies that enhance validity on one dimension can often erode validity on another dimension. Thus, if researchers pursue strategies focused overwhelmingly on statistical conclusion validity, this can sometimes come at the cost of undermining other types of validity. If one or more of these other forms of validity determine replication rates in a given domain, efforts to increase statistical conclusion validity could lead to minimal increases or even decreases in replication rates. To illustrate how seemingly straightforward recommendations designed to enhance one type of validity can sometimes have unanticipated consequences for other types of validity, it is useful to consider an example. Specifically, many journals have implemented guidelines suggesting or 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 a quick and easy criterion for initially evaluating research. The goal of such guidelines would seem to be to increase power of published research and thereby enhance statistical conclusion validity. All else being equal, increased power is of course useful, but all else is rarely equal. There may be unintended consequences of such requirements.

For example, to the extent that researchers become focused on increasing their sample size as a way to increase power, researchers may have recruited participants and collected data in different ways, such as through relatively inexpensive online platforms such as Mechanical Turk. Such a move is not inherently problematic, and clearly these platforms have a useful role to play in psychological research. For example, increased diversity of the obtained sample via these platforms might be beneficial for examining whether results hold across different subpopulations of respondents, though such tests are only useful if sufficient numbers of members of the smaller groups are present in the sample. Unfortunately, simply having diversity of participant characteristics in the sample does not ensure external validity across the subpopulations, and that very same diversity could have potentially negative consequences for construct validity (see also Finkel et al., 2017). That is, if the sample diversity means that a given manipulation or measure differs in its meaning across participants, the construct validity of that manipulation or measure might be undermined.

Thus, although increasing sample size might, in concept, benefit statistical conclusion validity, researchers’ attempts to strengthen this validity might be a detriment to construct validity (and, if construct validity is threatened in this way, the overall effect could be weakened in such a way that power and the resulting statistical conclusion validity would not be helped either). One could certainly view this type of example as a good reason to think more broadly about statistical power, rather than focusing on sample size per se. For example, to the extent that researchers try to maximize power by using more precise measures and manipulations that are consistently interpreted by the participants in one’s study, the increased focus on power could actually benefit construct validity. Likewise, conducting a study in a controlled laboratory setting that minimizes distractions and allows for more precise control of extraneous factors could also increase the efficacy of experimental manipulations and measures, thereby enhancing construct validity as well as statistical conclusion
validity. Furthermore, online environments present limitations for certain kinds of research paradigms, potentially restricting the use of some operationalizations of IVs and DVs. Such a development could further erode construct validity and, perhaps in some cases, internal validity.

Another potential implication of a focus on power is that it could result in fewer but larger studies. Some have advocated this as a way to reduce selectivity in which studies are published (e.g., Schimmack, 2012). However, there are also potential costs of this practice. First, investing more resources into any given study is likely to reduce the number of different operationalizations of the key IVs and DVs within a program of research. But fewer operationalizations can also reduce construct validity. For example, if a single measure or manipulation is used across all of the studies in a paper (which might be more likely to be a single study if power is extremely high), then it becomes more likely that the impact of that measure or manipulation could be due to one of the “irrelevancies” that varies with that particular measure or manipulation. When similar effects are found across different operationalizations (that presumably have different irrelevancies), this triangulation helps to ensure that the results are due to the core construct instead of some unintended feature. Of course, triangulation is not the same thing as including a variety of measures and only reporting a subset of those that produce results. Rather, effective triangulation involves pilot testing and identifying alternative manipulations or measures that effectively capture the intended construct and are then tested (typically separately) across studies.9

None of our comments are intended to suggest that large sample sizes are undesirable. There are clear methodological advantages to having such samples. However, it is not clear that this criterion should be given special status to the exclusion of other criteria. If it is given such status, maximizing it will likely come at the expense of other criteria. The situations and operationalizations most conducive to large sample sizes might be less conducive to other desirable features of studies. Thus, we might expect to find that certain types of research could be viewed quite positively from the perspective of one criterion, but rather poorly from the perspective of a different criterion.

Conclusions

Researchers have long accepted that statistical conclusion validity, internal validity, construct validity, and external validity are all critical considerations in the conduct and evaluation of psychological research. Likewise, researchers have also long recognized that any study involves a balance of these sometimes competing concerns. The optimal weighting of these dimensions will be a function of the goals of the study in question. That being said, pursuing one form of validity at the exclusion of others has generally not been seen as an appropriate strategy for conducting sound psychological research. In our view, the same is also true for the conduct and evaluation of replication studies. Ultimately, an understanding of why studies do not replicate and reforms for enhancing the replicability of research will likely require the same careful consideration and balancing of these four types of validity.

Notes

1 Of course, a variety of theories predict interactions in which one simple effect is supposed to be larger than the other. When a simple effect is larger when it is predicted to be smaller (or smaller when predicted to be larger), that would constitute a “failure” of the theory. But even such interaction hypotheses are typically directional in that one simple effect is predicted to be larger than the other or smaller than the other, not larger by a particular amount or smaller by a particular amount.

2 The distinction between conceptual and exact replications is discussed in more detail later in the chapter.

3 It is important to note that such a context would constitute a “best” case scenario for the acceptance of a false positive. In other cases, some patterns of findings might be conceptually implausible and prove difficult to explain, thus making only certain spurious effects likely to be accepted as valid. In the context of multiple studies, the spurious effects would need to not only be conceptually plausible, but also of a consistent pattern across studies and labs.
4 As noted in our discussion of low statistical power, the case of a single study in which any pattern of an effect could be deemed conceptually plausible constitutes the context most vulnerable to the acceptance of false positives. In cases where some patterns of effects would not be deemed plausible, a researcher would have to engage a set of practices that not only produced a spurious effect, but a spurious effect of a specific pattern or some subset of possible patterns. In the context of multiple studies, the conceptually plausible spurious pattern would also need to be consistent across studies. Moreover, presumably a researcher would also need to report a similar set of practices across studies to produce the spurious effects so as not to arouse suspicion.

5 In some nonexperimental studies, internal validity is irrelevant in that the researcher is not postulating any sort of causal effect, and it would not even be meaningful to designate some variables as IVs and other variables as DVs. However, in other cases, nonexperimental studies might well involve measures where theorizing conceptualizes some variables as IVs and others as DVs. In these contexts, internal validity is a meaningful concept, although achieving high levels of this form of validity is generally difficult, if not impossible. Moreover, in the context of nonexperimental studies, differentiating between threats to internal validity and construct validity (which will be discussed in the next section) can be extremely subtle. Because a full discussion of these issues goes beyond the scope of the present chapter, we will confine ourselves to discussing the role of internal validity in the context of experimental studies.

6 We focus on the threat of attrition because the other common threats to internal validity (e.g., history, maturation) are taken care of when participants are randomly assigned to experimental conditions.

7 It should, of course, be noted that the journal articles and book chapters we examined are by no means a comprehensive compilation of publications in the replication literature. This literature is vast and spans many disciplines, even if one considers only the methodology literature in psychology and related disciplines. That being said, the number of publications we investigated was substantial, and they were drawn from a number of very prominent outlets for this literature. It is not clear that there is any reason to suspect that the sample of publications we examined was in some way biased in favor of statistical conclusion validity concerns. Likewise, the journals we selected were chosen because of their relation to social or consumer psychology. There is no clear reason to expect that these journals would be especially prone to concerns regarding statistical conclusion validity.

8 One might argue that, in a single large study, a researcher could of course include multiple operationalizations of an IV and a DV. This point is certainly true, but multiple operationalizations of an IV will often produce much more complex designs, requiring even larger samples sizes. Likewise, inclusion of multiple DVs in the context of the same study can raise the risk of carryover effects and other threats to validity.

References


