Finkel, Eastwick, and Reis (2016; FER2016) argued the post-2011 methodological reform movement has focused narrowly on replicability, neglecting other essential goals of research. We agree multiple scientific goals are essential, but argue, however, a more fine-grained language, conceptualization, and approach to replication is needed to accomplish these goals. Replication is the general empirical mechanism for testing and falsifying theory. Sufficiently methodologically similar replications, also known as direct replications, test the basic existence of phenomena and ensure cumulative progress is possible a priori. In contrast, increasingly methodologically dissimilar replications, also known as conceptual replications, test the relevance of auxiliary hypotheses (e.g., manipulation and measurement issues, contextual factors) required to productively investigate validity and generalizability. Without prioritizing replicability, a field is not empirically falsifiable. We also disagree with FER2016’s position that “bigger samples are generally better, but . . . that very large samples could have the downside of commandeering resources that would have been better invested in other studies” (abstract). We identify problematic assumptions involved in FER2016’s modifications of our original research-economic model, and present an improved model that quantifies when (and whether) it is reasonable to worry that increasing statistical power will engender potential trade-offs. Sufficiently powering studies (i.e., $\geq 80\%$) maximizes both research efficiency and confidence in the literature (research quality). Given that we are in agreement with FER2016 on all key open science points, we are eager to start seeing the accelerated rate of cumulative knowledge development of social psychological phenomena such a sufficiently transparent, powered, and falsifiable approach will generate.

**Keywords:** direct replication, falsification, power, replicability, transparency

In response to recent suggestions regarding how to improve the quality and trustworthiness of research in the fields of social and personality psychology, some (e.g., Finkel, Eastwick, & Reis, 2015, or FER2015, and Crandall & Sherman, 2016) have argued that proposed reforms could result in a number of undesirable side effects. Specifically, FER2015 conjectured that reducing false-positives increases false-negatives, and that requiring larger sample sizes and direct replications limits the rate of discovery by taxing available resources. We (LeBel, Campbell, & Loving, 2016, or LCL2016) showed that these concerns are unwarranted and reiterated that achieving a literature lending to reliable identification of true discoveries involve some costs, but that gained benefits outweigh such costs. We then presented some simple calculations showing how larger sample sizes can actually increase the efficiency of confirming true discoveries.

We were happy to find several points of agreement in Finkel, Eastwick, and Reis’ (2016, or FER2016) response. Importantly, we now both agree on the crucial need for direct replication in social and personality psychology. They also emphasize that open science and replicability reforms must be tailored to each research domain, something we have previously advocated for relationship science (Campbell, Loving, & LeBel, 2014) and others have done in other domains (e.g., industrial/organizational psychology, Kepes & McDaniel, 2013).

We disagree, however, with FER2016’s concern that recent reform has “focused on a single feature of high-quality science—replicability—with insufficient sensitivity to […] other features, like discovery, internal validity, external validity, construct validity, consequentiality, and cumulativeness” (abstract). We agree that these other goals of science are essential, but argue a more
systematic approach to replication is needed to accomplish these goals. Replication is the *general empirical mechanism for testing and falsifying theory*. Sufficiently methodologically similar replications, or what have sometimes been called *direct replications*, test the basic existence of phenomena and ensure cumulative progress is possible a priori. In contrast, increasingly methodologically dissimilar replications, or what have sometimes been called *conceptual replications*, test the relevance of auxiliary hypotheses (e.g., manipulation and measurement issues, contextual factors) required to productively investigate the validity and generalizability of psychological phenomena. Without prioritizing replicability, a field is not empirically falsifiable (Popper, 1959).

We also disagree with FER2016’s position that “bigger samples are generally better, but . . . that very large samples (‘those larger than required for effect sizes to stabilize’; FER2015, p. 291) could have the downside of commandeering resources that would have been better invested in other studies” (abstract). We identify some problematic assumptions involved in FER2016’s modifications of our original research-empirical model, and present an improved model that quantifies when (and whether) it is reasonable to worry about proposed trade-offs of increasing statistical power.

### The Historical Context of the Replication Movement

Social psychology prior to the “replication crisis” (pre-2011) focused primarily on testing generalizability and internal validity via methodologically dissimilar replications, or what have sometimes been called *conceptual replications*, with much less attention paid to methodologically similar replications, or what have sometimes been called direct replications (LeBel & Peters, 2011; Pashler & Harris, 2012). Though it is difficult to quantify precisely, bibliographic evidence suggests the prevalence rate of so-called *direct replications* in the published psychology literature pre-2011 is most likely less than 1% (Makel et al., 2012; M. Makel, personal communication, November 29, 2012). To correct this imbalance, and to begin more systematically disentangling true from false findings, a growing (and ongoing) movement post-2011 has focused on assessing replicability via methodologically similar *direct replications*. Moving forward, the interplay between replication and other goals of science must be made more explicit. To achieve this, a more fine-grained language, conceptualization, and approach to replication is needed.

### The Replication Continuum

Replications lie on an *ordered* continuum of methodological similarity to an original study, ranging from highly similar to highly dissimilar. A highly methodologically *similar* replication, or what has sometimes been called a *direct replication*, repeats a study using methods as similar as is reasonably possible to the original study such that there is no reason to expect a different result based on current understanding of the phenomenon (Nosek et al., 2012). On the other hand, a highly methodologically *dissimilar* replication, or what has sometimes been called a *conceptual replication*, repeats a study using different general methodology to test whether a finding generalizes to different manipulations, measurements, domains, and/or contexts (Asendorpf et al., 2013; Lykken, 1968; Schmidt, 2009). Crucially, however, and in contrast to common dichotomous views of replication (e.g., FER2015; FER2016; Crandall & Sherman, 2016), both direct and conceptual replications can each reflect different levels of methodological similarity to a previous study (e.g., some direct replications are more similar to a previous study than other direct replications; some conceptual replications are more dissimilar to a previous study than other conceptual replications). Even more problematic, different researchers sometimes use different terminology to refer to replications exhibiting the same level of methodological similarity. Consequently, a more fine-grained language and conceptualization of replication is needed. Rather than using the ambiguous and inconsistently used terms “direct” versus “conceptual” replications, researchers should conceptualize replications with respect to their *relative methodological similarity* to a previous study.

Presented in Figure 1 is a replication taxonomy that aims to achieve just this, informed by earlier taxonomies proposed by Hendrick (1991) and Schmidt (2009). According to this taxonomy, different types1 of increasingly dissimilar replications exist between the highly similar and highly dissimilar poles, each serving different purposes.

In an *exact replication* (1st column), every relevant methodological feature under a researcher’s control is the same except for contextual variables (e.g., history). *Very close replications* (2nd column) employ the same independent variable (IV) and dependent variable (DV) operationalizations and IV and DV stimuli as the original study, but can differ in terms of procedural details (e.g., task instruction wording, font size) and physical setting (e.g., laboratory vs. online), barring any required linguistic and/or cultural adaptations of the IV or DV stimuli (which are part of “contextual variables”; e.g., LeBel & Campbell’s [2013] replications of Vess’ [2012] Study 1 can be considered very close replications given all design facets were the same except minor procedural details such as task instruction wording, font, and font size). *Close replications* (3rd column) employ the same IV and DV operationalizations, but can employ different sets of IV or DV stimuli (or different scale items or versions) and different procedural details (e.g., Domachowska et al.’s, 2016 Study 2 replication of Gable & Harmon-Jones’ [2008] Study 2 can be considered a close replication, given that the same IV and DV operationalizations were used, but with different IV affective stimuli). *Far replications* (4th column) involve different operationalizations for the IV or DV constructs (e.g., Muraven & Slessareva’s [2003] Study 1 can be considered a far replication of Baumeister et al.’s [1998] Study 1: ego depletion was manipulated via thought suppression rather than food temptation). Finally, in *very far replications* (5th column) everything can be different, with the only similarity being a theoretical abstraction of the phenomenon (e.g., as in Bargh, Chen, & Burrows’ [1996] Study 1, 2, and 3). Hence, *exact*, *very close*, and *close replications* are different types of what have sometimes been called *direct replications* whereas *far* and *very far replications* are different types of what have sometimes been called *conceptual replications*, each type involving an increasingly dissimilar methodology relative to an original study.

---

1 These “types” of replications should not be construed too rigidly. The differences in similarity between replications should be regarded as continuous with some unavoidably fuzzy boundaries between types.
Each type of replication serves a different role. **Exact** and **very close** replications establish the basic existence and stability of a phenomenon by falsifying the (null) hypothesis that observations simply reflect random noise (e.g., a Type I error). Increasingly dissimilar replications, such as **close** and **far** replications, establish links between phenomena and relevant contextual factors by falsifying particular auxiliary hypotheses related to the manipulation and/or measurement of the phenomenon under study. As such, increasingly dissimilar replications contribute to validity and allow **empirically justified** generalization of the construct under study. Consequently, the suggestion that prioritizing replicability—via **exact** or **very close** replications—is detrimental to other goals of science (FAZ2016) is, in our view, incorrect. Rather, systematic use of different types of replications is the general method by which replicability and other goals are accomplished.

**Falsification via Different Types of Replications**

**Exact and Very Close Replications Test the Basic Existence of Phenomena**

Replicability must be prioritized by our field because it is the only feature of science that can empirically falsify false hypotheses, ensuring we are dealing with stable phenomena that actually reflect reality. Falsification via **exact** or **very close** replications ensures we are (a) not fooling ourselves (Feynman, 1974) and (b) not being fooled by spurious results reported by another researcher. We, as individuals, are easiest to fool because of several cognitive (e.g., confirmation bias; Nickerson, 1998) and motivational biases (e.g., motivated reasoning; Kunda, 1990), which are substantially amplified by the hyper-competitiveness (Rick & Loewenstein, 2008) and problematic incentives within academia (Nosek, Spies, & Motyl, 2012), and/or possible outside financial interests (Coyne, 2016). These biases are particularly problematic if hypotheses are not pre-registered prior to data collection or if blinded analyses are not used (an approach that minimizes researcher bias by preventing knowledge of results influencing analytic decisions; MacCoun & Perlmutter, 2015).

Falsification via **exact** or **very close** replications also ensures we’re not being fooled by published research. The first scientific organization’s (Royal Society, 1660) defining motto was *Nullius in Verba* (i.e., take no one’s word); all scientific claims must be independently verified via repeatable experimentation rather than appeal to authority. The dictum “trust, but verify” exemplifies the scientific spirit: Though we trust our colleagues are acting in good faith and to the best of their abilities, findings must be independently verified. In this way, prioritizing replicability protects the scientific community against spurious results, whether such results are attributable to unintentional false positives or intentional fraud.

By preventing being fooled by ourselves or published results reported by others, testing replicability via **exact** or **very close** replication ensures we are dealing with real phenomena that are reliably observable and thus actually exist. Consequently, prioritizing replicability guarantees incremental progress. As can be seen in Table 1, if a phenomenon is not replicable (i.e., it cannot be consistently observed under specifiable conditions), it is simply not possible to empirically pursue the other goals of science. For example, making conclusions about internal validity or drawing causal inferences about a finding that is not replicable risks founding theory on vacuous truths.

---

2 The story of “Clever Hans” represents a compelling example of how *close* replications contribute to internal validity (Lykken, 1968, p. 155): “the apparent ability of the remarkable horse to add numbers had been due to an uncontrolled and unsuspected factor (presence of horse’s trainer within horse’s field of view). This factor, not being specified in the methods section, was omitted in the replication which for that reason failed.”

3 The historical case of cold fusion provides a compelling example of how direct replications are required to establish the basic existence of phenomena. Follow-up studies using very different methodology yielded a trickle of positive results whereas methodologically similar studies yielded overwhelmingly negative results (Taubes & Bond, 1993).
Publishing replicable effects that may lack internal validity leaves a field much better off than one in which nonreplicable effects are presented as internally valid because a reader can at least easily speculate about confounds; readers cannot easily gauge the nonreplicability of a finding (Simmons, 2016). Regarding construct validity, if a target phenomenon is not reliably observable and hence may not actually exist, it is not possible to empirically and iteratively test the nomological network of relations among related constructs (Cronbach & Meehl, 1955). One feature that can be said to be more important than replicability is the nonempirical, logical aspects of construct validity (e.g., face validity) whereby a construct must at minimum be logically and theoretically coherent. Like replicability, however, this is too fundamental a requirement to be considered a trade-off. Cumulativeness, consequentiality, construct validity and replicability all must be achieved for a field to be a successful science. Replicability, however, is the only one of these goals that is logically and empirically necessary for achieving the others.

### Falsification via Close and Far Replications

Falsification via close or far replications allows the systematic testing of auxiliary hypotheses and contextual variables required to productively investigate the validity and generalizability of psychological phenomena (Meehl, 1967, 1978). This involves an iterative approach whereby an original finding is followed by increasingly methodologically dissimilar replications in small rather than large steps, and periodically checking replicability in a process that has been termed “systematic replication” (Hendrick, 1991) or “replication batteries” (Rosenthal, 1991). For example, an original finding from another lab should be followed by a close replication (third replication type in Figure 1) employing the same general methodology for the IV and DV as the original study but different stimuli (or more efficiently, by including both very close or close replication “anchor-cells” and separate far replication “generalizability cells”; Hendrick, 1991; Rosenthal, 1991). This is in contrast to following up with a far or very far (“conceptual”) replication, a common approach in social psychology pre-2011 (Makel et al., 2012). If a close replication is unsuccessful, a researcher should subsequently follow-up with a more methodologically similar study (i.e., a very close replication). If a close replication is successful, however, then a subsequent follow-up study could use more dissimilar methodology via a far replication. In a further step, the target phenomenon could be examined in a different domain via a very far replication. Whenever difficulty is encountered in replicating a phenomenon, a researcher should either repeat the experiment at an equivalent or earlier point on the replication continuum rather than run a highly dissimilar study instead.

### Problems When Replicability Is Not Prioritized

Problems can arise when highly dissimilar far and very far replications are consistently prioritized over more similar very close or close replications, attempting to establish the outermost limits of generalizability before confirming the basic existence of phenomena or testing the relevance of immediate contextual factors (e.g., as espoused by FER2016 and Crandall & Sherman, 2016). Published findings then acquire inertia—or an incumbent advantage—that leads to a conservatism bias (Edwards, 1982) whereby researchers resist updating their beliefs according to new evidence (a form of anchoring bias). Further, an independent researcher who executes a far or very far replication study in a different region, culture/ethnicity, country, or political climate is unlikely to question the existence of the original finding if such replication is unsuccessful. In an attempt to first show the generalizability of a finding, the essential replicability step is skipped. Instead of testing and establishing replicability, this approach implicitly assumes replicability (e.g., Zanna, 2004), leaving most false-positive findings unidentified and the theories used to explain those findings virtually never falsified (Ferguson & Heene, 2012). Indeed, Feynman (1974) argued that not testing replicability along the way was exactly the flaw that prevented psychology from being a cumulative science: “it seems to have been the general policy then to not try to repeat psychological experiments, but only to change the conditions and see what happens” (p. 13). In this way, falsification via testing replicability is the key, nonoptional principle undergirding science’s success.

Unfortunately, these problems have emerged in several highly influential research areas in social psychology. Researchers have been unable to replicate, via high-powered designs that closely matched methodology of original studies (i.e., exact, very close, or close replications), key phenomena across several different operationalizations (e.g., ego-depletion, superiority-of-unconscious decision-making effect, Macbeth effect, power posing, mood on helping effect, money priming, cleanliness priming, elderly priming, achievement priming, professor priming, God/religion priming, font disfluency on math performance, color priming, mate priming, U.S. flag priming, heat priming, honesty priming, distance priming, embodiment of secrets, embodiment of warmth; see Appendix A at https://osf.io/8sro6/ for citations to original and replication studies and http://CurateScience.org for more examples). This is in the context of an even longer list of highly cited social psychological findings that have been systematically replicated in small steps, with opportunistic replications of highly influential research (Makel et al., 2012). The problem is not that these findings are unreplicable, nor that researchers are resistant to testing replicability. The problem is that replicability is systematically prioritized over generalizability in each instance, leading to the problems listed above.

---

4 Bem’s (2011) precognition offers a compelling case in point. It is not sensible to launch into a series of long-winded construct validity investigations (Cronbach & Meehl, 1955) of precognition if the phenomenon’s basic existence is unconfirmed (as eventually was shown to be the case; Galak, LeBoeuf, Nelson, & Simmons, 2012; see CurateScience.org for other unsuccessful replications).
encountered replication difficulties across a single operationalization of an original finding (e.g., grammar-intentionality effect, relationship commitment priming, facial feedback effect, playboy effect, Romeo and Juliet effect, ovulation on face preferences, ovulation on voting, color on physical attraction, status legitimacy effect, stereotype threat). That said, a few successful replications of social psychology effects have been reported (e.g., approach-motivated positive affect constrains attention, Domachowska et al., 2016; sex differences in distress to infidelity, IJzerman et al., 2014; see Appendix A at https://osf.io/8srcld/ for more examples). Unfortunately, however, such successful replications are much fewer relative to the hundreds of unsuccessful replications that have now been published.

Can’t We Just Strive for Moderate Power?

Sadly, no. Running a moderately powered study (e.g., 50%) renders nonstatistically significant results uninterpretable (i.e., is my nonstatistically significant result due to falsity of the tested hypothesis or due to insufficient power?). Consequently, a researcher can get caught up futilely adjusting an experimental design in a misguided attempt to recapture a statistically significant phenomenon that never existed in the first place. Sufficient power (e.g., >80%) to detect plausible effect sizes increases the interpretability of nonstatistically significant results. If the phenomenon exists it would have been detected, assuming sound auxiliaries. But if it was not detected, it either does not exist or is a subtle effect (i.e., a faint star), in which case a much more powerful telescope is needed (Simonsohn, 2015).

But how much power is sufficient? And is there a point beyond which increasing power incurs more costs than benefits? Both LCL2016 and FER2016 attempted to approach this question through some simple models of the research process. By publicly sharing the R code of our LCL2016 model (https://osf.io/hpwqdl/), we have facilitated open discussion, which has led to productive extensions of our original model.

Clarifying and Extending LCL2016’s Model

LCL2016’s original model had the specific goal of estimating the collective replication resources (i.e., exact or very close replications) that would be required to distinguish true from false findings when original studies are planned for low (25%; status quo) versus high (80%) power (see Appendix B at https://
FER2016 sought to extend only the efficiency portion of our original model to gauge overall research efficiency of the field. Their model combined participant costs from original and replication studies and divided this by the number of true positives produced by original researchers to sufficiently replicate low-powered (25%) versus high-powered (80%) original studies. Under these assumptions, results suggested low-powered original research requires spending more replication resources per discovery to later confirm findings compared to high-powered original research. Consequently, though a larger number of underpowered original studies can be run given a finite amount of original study resources, the trade-off is greater resources per discovery to later confirm findings compared to high-powered original research. Consequently, though a larger number of underpowered original studies can be run given a finite amount of original study resources, the trade-off is greater resources per discovery to later confirm findings compared to high-powered original research.

We implemented such a model with fixed, more realistic amounts of replication resources (5%, 10%, or 20% of total resources) and using 80%-powered rather than 95%-powered replications (R code implementation of our model available at https://osf.io/wp6an/). Note that the specific values chosen for replication resources are less important than the trend that emerges from varying them (though current values are generous: see Appendix B at https://osf.io/3vr8a/ for more details).

Our updated model is depicted in Figure 2, representing a realistic scenario where replication resources reflect 5% of total resources and the base rate of true hypotheses is 20% (Miller & Ulrich, 2016, equation A.10). The moderate-power approach (top panel A) involves a larger proportion of false negatives (10%) and smaller proportion of positive results (14%) compared with high-power approach (4% and 20%, respectively). Crucially, a smaller proportion of positive results can be followed-up via replication for the moderate-power approach (18.7%) compared with high-power approach (26.5%). Hence, the moderate-power approach leads to fewer confirmed true discoveries via replications (57%) compared with the high-power approach (64%), and consequently larger overall resources per confirmed true discovery ("NTCD").

5 FER2016 implied our model did not consider these types of trade-offs, but as just described, it examined precisely this issue.
The main results of our updated model are presented in Figure 3. The moderately powered approach is suboptimal in terms of both efficiency and overall quality. Focusing on the 5% replication resource scenario (most realistic current value) for the left panel (α = .05), high-powered research (circle) incurs fewer overall resources than moderately powered research (triangle) and yields higher overall confidence in a literature’s results (83% vs. 74%). Similar results emerge when α = .10 (right panel). Increasing power uniformly increases both research efficiency and quality, with diminishing returns occurring only when either (a) alpha level is very stringent (α < .01) or (b) an unrealistically large proportion of the field’s total resources (e.g., 20%) is spent on direct replications. Given current resources spent on direct replications are most likely considerably less than our model values (Makel et al., 2012; M. Makel, personal communication, November 29, 2012), it is very unlikely that sufficient statistical power will result in undesirable trade-offs in research efficiency or quality. That said, optimal power levels depend on the utility and disutility of different study outcomes (Miller & Ulrich, 2016), which may vary across research areas. Our general recommendation of sufficient power (>80%) should be seen as a default power level unless unambiguous and objective outcome (dis)utility suggests otherwise.

6 If replications are powered at 90% or 95%, highly similar results emerge that lead to the same conclusion.

References


New Policy for the Journal of Personality and Social Psychology

The Journal of Personality and Social Psychology is inviting replication studies submissions. Although not a central part of its mission, the Journal of Personality and Social Psychology values replications and encourages submissions that attempt to replicate important findings previously published in social and personality psychology. Major criteria for publication of replication papers include the theoretical importance of the finding being replicated, the statistical power of the replication study or studies, the extent to which the methodology, procedure, and materials match those of the original study, and the number and power of previous replications of the same finding. Novelty of theoretical or empirical contribution is not a major criterion, although evidence of moderators of a finding would be a positive factor.

Preference will be given to submissions by researchers other than the authors of the original finding, that present direct rather than conceptual replications, and that include attempts to replicate more than one study of a multi-study original publication. However, papers that do not meet these criteria will be considered as well.

Submit through the Manuscript Submission Portal at (http://www.apa.org/pubs/journals/psp/) and please note that the submission is a replication article. Replication manuscripts will be peer-reviewed and if accepted will be published online only and will be listed in the Table of Contents in the print journal. As in the past, papers that make a substantial novel conceptual contribution and also incorporate replications of previous findings continue to be welcome as regular submissions.