Running Head: FEATURES OF A HIGH-QUALITY SCIENCE

Replicability and Other Features of a High-Quality Science:

Toward a Balanced and Empirical Approach

Eli J. Finkel

Northwestern University

Paul W. Eastwick

University of Texas at Austin

Harry T. Reis

University of Rochester

In Press, *Journal of Personality and Social Psychology*

July 2, 2016

Author note. The authors thank Galen Bodenhausen, Jim Coan, Alison Ledgerwood, and Brian Nosek for their insightful feedback on this article.

Abstract

Finkel, Eastwick, and Reis (2015; "FER2015") argued that psychological science is better served by responding to apprehensions about replicability rates with contextualized solutions than with one-size-fits-all solutions. Here, we extend FER2015's analysis to suggest that much of the discussion of best research practices since 2011 has focused on a single feature of high-quality science—replicability—with insufficient sensitivity to the implications of recommended practices for other features, like discovery, internal validity, external validity, construct validity, consequentiality, and cumulativeness. Thus, although recommendations for bolstering replicability have been innovative, compelling, and abundant, it is difficult to evaluate their impact on our science *as a whole*, especially because many research practices that are beneficial for some features of scientific quality are harmful for others. For example, FER2015 argued that bigger samples are generally better, but also noted that very large samples ("those larger than required for effect sizes to stabilize"; p. 291) could have the downside of commandeering resources that would have been better invested in other studies. In their critique of FER2015, LeBel, Campbell, and Loving (2016; "LCL2016") concluded, based on simulated data, that ever-larger samples are better for the efficiency of scientific discovery (i.e., that there are no tradeoffs). As demonstrated here, however, this conclusion holds only when the replicator's resources are considered in isolation. If we widen the assumptions to include the original researcher's resources as well, which is necessary if the goal is to consider resource investment for the field as a whole, the conclusion changes radically—and strongly supports a tradeoff-based analysis. In general, as psychologists seek to strengthen our science, we must complement our much-needed work on increasing replicability with careful attention to the other features of a high-quality science.

Replicability and Other Features of a High-Quality Science: Toward a Balanced and Empirical Approach

> The research process can be viewed as a *series of interlocking choices*, in which we try *simultaneously to maximize several conflicting desiderata*.
>
> — McGrath, 1981, p. 179 (italics in original)

When pursuing the goal of conducting high-quality science, researchers must learn to live with intractable dilemmas, making decisions that optimize a study's overall scientific contribution despite the fact that no method can produce maximal value on every feature of good science. McGrath (1981) illustrated this point by demonstrating that external validity, experimental control, and experimental realism—three features that are, in isolation, unmitigated scientific goods—are inherently incompatible, because maximizing one of them makes it impossible to maximize the others. In the terminology of our "Best Research Practices in Psychology" article (Finkel, Eastwick, & Reis, 2015; "FER2015"), all research strategies involve *tradeoffs* among the desirable features of a high-quality science—among McGrath's "desiderata." Consequently, in FER2015, we expressed enthusiasm for the increased attention that scholars have brought to the issue of replicability since 2011 (e.g., Klein et al., 2014; Open Science Collaboration, 2015), but we also noted that many current proposals for improving replicability could unintentionally weaken other features of a high-quality science. Without such tradeoff-based thinking, we argued, the field cannot even ask questions regarding whether a given research practice is ultimately beneficial or harmful for our science *as a whole* (i.e., across the full range of desirable features), even if it is clearly beneficial for a particular feature (e.g., replicability).

Building on FER2015's appeal for tradeoff-based thinking, LeBel, Campbell, and Loving (2016; "LCL2016") considered the costs and benefits of certain research practices. Especially innovative was their simultaneous consideration of the scientific features of *discovery* (i.e., finding

evidence in support of novel hypotheses) and *replicability* (i.e., finding that results emerge in other random samples that capture the most important facets of the research design; Asendorpf et al., 2013). LCL2016 tested the potential tradeoff between these two features in a series of simulations that examined how the efficiency of *true discoveries* (statistically significant effects that withstand rigorous replication attempts) changes depending on whether researchers allocate their available research participants (*N*) to few high-powered studies or many low-powered studies. On the basis of these simulations, LCL2016 concluded without caveats that the former approach is better than the latter for the efficiency of *our science as a whole*, regardless of how large the sample in question already is. Indeed, their simulations and accompanying online app suggest that adding power to a single study (by increasing the sample size) *always* increases the efficiency of scientific discovery—that doing so has no tradeoffs in the pursuit of true discoveries. However, as we demonstrate herein, this conclusion results from the aforementioned omission of the original researcher's resources in LCL2016's efficiency calculation. When the resources of both the original researcher and replicator are included, the simulations further bolster the FER2015 conclusion that tradeoff-based thinking is crucial as we seek to establish which research practices are best for our science. Before describing these simulations in more detail, however, we first situate this discussion in a broader epistemological context, one that considers the core features of a high-quality science and focuses on the importance of adopting a tradeoff-based approach to evaluating research practices.[1]

The Core Features of a High-Quality Science

Replicability is a necessary feature of a healthy scientific discipline, and doubts about the replicability of published effects (Simmons, Nelson, & Simonsohn, 2011) catalyzed psychology's

[1] In this report, our goal is neither to address every point of disagreement with LCL2016 nor to set the record straight regarding all of the cases where (in our view) LCL mischaracterized what we said in FER2015. Rather, we focus on the issues that afford the best opportunity for constructively moving the discussion forward.

evidentiary value movement. We argued in FER2015 that this increased emphasis on replicability is excellent for the field because it can help reduce false-positive error rates—and that it is also important to consider whether specific proposals designed to bolster replicability might exacerbate false-negative error rates. Building on Fiedler, Kutzner, and Krueger's (2012) analysis, we adopted an expansive definition of "false negatives" that includes cases in which true-positive findings are omitted from the scholarly literature due to an increasingly stringent editorial formulary.

The present analysis extends FER2015's "error balance" logic to emphasize tradeoffs *among features of a high-quality science* (among scientific desiderata). When seeking to optimize the quality of our science, scholars must consider not only how a given research practice influences replicability, but also how it influences other desirable features. Beyond discovery (i.e., results that document support for novel hypotheses) and replicability (i.e., results that reflect those obtained with other random samples), what other features are essential for building a high-quality science? We make no attempt to provide a comprehensive list of such features here, nor do we attempt to discern the circumstances under which certain features are more important than others. Rather, we discuss a set of features in the hope that our efforts will contribute to a robust field-wide discussion about what our core features of scientific quality are and the extent to which we should prioritize each of them in a given research context.

Figure 1 provides a preliminary list, beginning with discovery and replicability (see boxes under "proximal means"). Cook and Campbell (1979, pp. 38-39) discuss two more: *internal validity* ("the validity with which statements can be made about whether there is a causal relationship from one variable to another") and *external validity*, also called representativeness or generalizability ("the approximate validity with which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times").

Presaging McGrath's (1981) and FER2015's tradeoff-based analysis, Campbell (1957, p. 297) observed that both internal validity and external validity "are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness." Another key feature is *construct validity*, which refers to correctly linking the theoretical constructs to the operationalizations that were conducted in the study itself (Brewer & Crano, 2014; Cook & Campbell, 1979).

Even if a scientific discipline aligns strongly with these first five features (i.e., it has discovered a substantial number of replicable findings that are high in internal, external, and construct validity), there is no guarantee that it is flourishing. It could be that most or all of these findings are low in *consequentiality*—that they are unimportant or uninfluential. For example, a discipline might lack consequentiality if its findings are not theoretically interesting, if other sciences do not draw from its insights, or if it fails to yield findings that can be effectively applied (e.g., to improve humanity's average quality of life). Or it could be that the findings lack *cumulativeness*—they fail to cohere in a manner that affords conceptual integration across studies (Mischel, 2006). For example, a discipline might lack cumulativeness if its articles test disconnected hypotheses, or if scholars fail to draw connections to conceptually related findings from other laboratories and subfields.

On Tradeoffs: No Study Can Accomplish Everything, and Resources Are Finite

When considering large collections of studies, it is important to pursue all of the features of a high-quality science. Depending on the context, some features might be prized more than others, but the collection of studies must achieve a reasonably high level of all features to be considered a mature research space. However, as we narrow the focus from a discipline to a topic area to a research program to an individual study, tradeoffs among the features loom ever larger. These tradeoffs emerge for two reasons. First, no single study can accomplish everything. In the wake of

a given study, for example, there will always be alternative explanations for the effectiveness of a manipulation (i.e., doubts about internal validity), real-world contexts to which the finding may not generalize (i.e., doubts about external validity), and the possibility that the results capitalized on chance (i.e., doubts about replicability). Second, resources are finite. Each resource (time, money, research participants, etc.) that a scholar invests in a study oriented toward bolstering replicability is a resource that she does not invest in a study oriented toward, say, bolstering internal validity. Such tradeoff-based analysis, which is widespread among methodologists and philosophers of science (e.g., Brewer & Crano, 2014; Cartwright, 2007; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002; Smith & Mackie, 2000), dovetails with the one presented by McGrath (1981, italics in original): "It is *always desirable (ceteris paribus) to maximize*" various desiderata, but "alas, *ceteris is never paribus*, in the world of research"; "[t]here is no way—in principle—to maximize all…desiderata" at once (p. 184-186).

Given this dilemma, the researcher must consider the multiple features of a high-quality science as she decides which study she will run next. She has many options at her disposal. She might conduct a study on a new research topic (e.g., to bolster discovery), or she might pursue one of several types of replication. *Direct replications*, sometimes called "exact" or "close" replications, are studies that precisely repeat the original procedure toward the goal of ascertaining the replicability of the original result. Such studies, which have fortunately become much easier to publish due to the evidentiary value movement, can help the researcher gauge the replicability of the original result. But if she wishes to build confidence in the internal and construct validity of a finding, she might instead prioritize *conceptual replications*, which are studies that vary the operationalizations of the original theoretical concepts (Fabrigar & Wegener, in press; Ledgerwood, Soderberg, & Sparks, in press; Lykken, 1968). Through conceptual replications, she can eliminate alternative explanations for the effect of a particular manipulation on a measure, and

she can triangulate on the theoretically relevant constructs of interest by using a variety of manipulations and measures. Alternatively, she can build toward external validity by conducting real-world extensions and/or *systematic replications*, which are studies that vary elements of the original procedure that should be unrelated to the effect of interest (Ledgerwood et al., in press). Through systematic replications, she can begin to test the generalizability of a finding across variations in procedure and setting.

Once the researcher has considered the various options, and weighed them in light of her available resources, she can settle on a study that reflects her research priorities. In this way, it is epistemologically sensible to expect that topic areas ultimately work to bolster all features of a high-quality science, but single studies will always prioritize some features over others.

Should Scientists Consistently Prioritize Replicability Above Other Core Features?

In the evidentiary value movement, replicability is sometimes treated as *equivalent to* scientific quality. To be sure, it may be that nobody actually believes that these two things are equivalent, but such equivalence is sometimes implied. Consider, for example, what a newcomer would likely conclude from reading the seminal texts in psychology's evidentiary value movement (e.g., John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). We have great respect for these texts, and we have no reason to doubt that their authors see value in desiderata other than replicability. But the newcomer could be forgiven for drawing the incorrect inference that replicability and scientific quality are one and the same, as these texts neglect these other features and do not consider potential costs that particular recommendations for bolstering replicability might have for them. The newcomer's inference would perhaps be reinforced by reading the most influential discussions on social media, including Michael Inzlicht's (2015) influential "Check Yourself before you Wreck Yourself" blogpost, which has been shared or retweeted nearly 9,000

times as of this writing and in which he explicitly (albeit perhaps inadvertently) treated replicability as equivalent to scientific quality.

One could argue that replicability is unique in that it is the first thing a researcher should want to know in the wake of a new discovery. The *prima facie* case for this suggestion is strong: In the absence of evidence for replicability, the original researcher and other scholars who learn of the study would be wasting their time trying to follow up with conceptual or applied extensions. But even this suggestion may vary in its applicability to different research domains, because studies vary in the *ease* with which the original conditions can be replicated.[2] Consider the famous first test of Einstein's theoretical perspective on gravitational light deflection, which capitalized on a solar eclipse (Dyson, Eddington, & Davidson, 1920), or studies of stress reactions conducted in the days after September 11th (Schuster et al., 2001). These studies have superlative internal and external validity, respectively, and the fact that direct replication is difficult or impossible need not discount their contributions to science. As with the other features of a high-quality science, (direct) replicability may not be demonstrable for some very high-quality studies, even as a healthy proportion of the studies in a high-quality body of research must be amenable to direct replication and yield supportive results in those replications.

In our view, the field's discussion of best research practices should revolve around how we prioritize the various features of a high-quality science and how those priorities may shift across our discipline's many subfields and research contexts. The various new initiatives seeking to ascertain replicability are impressive (e.g., Nosek et al., 2015; Simons, Holcombe, & Spellman, 2014), and as we work to craft strategies to improve replicability, we will want to ensure that we are crafting smart strategies—strategies that improve replicability without accidentally redirecting resources away from studies that bolster other desirable scientific features. After all, a discipline

[2] Furthermore, one could make a case that other features of a high-quality science, like internal validity, should be demonstrated first and foremost—what good is a highly reproducible artifact of a particular study design?

that is weak in, for example, internal validity or consequentiality would be scarcely better than a literature that is weak in replicability.

Different Research Practices Bolster Different Core Features

Figure 1 also presents specific research practices that influence one or more of the features (see boxes under "Distal Means"). Here, again, we seek to be illustrative rather than exhaustive, but even this brief list provides many examples in which a specific research practice may increase alignment with at least one core feature while decreasing alignment with one or more of the others. Consider calls for all data to be made publicly available, and the possibility that other researchers could publish novel findings from those data before the original researcher has had the opportunity to pursue her multiple-article publication plan. As FER2015 noted, such a free-for-all may bolster discovery by letting everybody delve into the data simultaneously, but it might also undermine cumulativeness by fostering piecemeal publication. Worse yet, it might disincentivize the massive resource investment required to design and conduct the sorts of methodologically ambitious studies—such as longitudinal field experiments—that tend to have strong external validity and consequentiality.

Similarly, requiring very large sample sizes increases replicability by reducing false-positive rates and increases cumulativeness by reducing false-negative rates, but it also reduces the number of studies that can be run with the available resources, so conceptual replications and real-world extensions may remain unconducted. Also, large sample size norms and requirements may limit the feasibility of certain sorts of research, thereby reducing discovery. That is, such norms and requirements are likely to increase the prevalence of the sorts of research that employs inexpensive, easy-to-access data (e.g., the sort currently exemplified by studies using Amazon's Mechanical Turk) while decreasing the prevalence of the sorts of research that employs expensive,

time-consuming, or difficult-to-access data (e.g., the sort FER2015 discussed in detail). Such changes may also compromise external validity, construct validity, and consequentiality.

Or consider the implications of increasingly comprehensive disclosure norms for the ability of researchers to present their articles in a compelling manner. As we work toward the laudable goal of greater transparency—which promotes replicability both by enabling editors and reviewers to evaluate the work more accurately and by strengthening scholars' ability to conduct near-direct replications following publication—we become increasingly vulnerable to producing indigestible articles. We must develop new writing norms (perhaps with liberal use of supplemental online materials) that accommodate the need for greater transparency while still affording authors the opportunity to write in a clear, cogent manner that aligns with how readers process information (Pinker, 2014). Lucid writing is likely to increase both cumulativeness (e.g., by making research reports easier for other scholars to digest and connect to their own research interests) and consequentiality (e.g., by making reports more accessible to people outside the field, including reporters, policymakers, and scholars in related disciplines).

Recently, the Association for Psychological Science began rewarding some research practices with a system of badges. Specifically, *Psychological Science* currently rewards researchers for three of the distal means presented in Figure 1—preregistration, open materials, and open data—all of which should have positive effects on replicability. It is clear why such badges have been prioritized in light of the field's intensifying focus on replicability. But, in principle, a large range of practices could be rewarded with badges. Why do we not reward, for example, the development of an artifact-free manipulation with a strong manipulation check to bolster internal validity, or the use of non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic; Henrich, Heine, & Norenzayan, 2010) samples to bolster external validity, or the creation of a new intervention with clear-cut potential to actually help people in the real world to bolster consequentiality, or the

bridging of two previously unconnected literatures to bolster cumulativeness? To be clear, we are not calling for more badges; we simply wish to raise awareness about research practices linked to other core features of a high-quality science that are not receiving much attention and could—if not nurtured—wane in favor of other, rewarded practices.

An Aspiration: Toward the Quantification of the Core Features of a High-Quality Science

Among the many contributions of the evidentiary value movement is an intensive emphasis on quantifying replicability, with scholars offering thoughtful discussions of how to quantify replication success and failure (e.g., Braver, Thoemmes, & Rosenthal, 2014; Etz & Vandekerckhove, 2016). Does this term refer to a statistically significant effect in the same direction as the original, or to an effect within the 95% confidence interval of the original? Does it refer to the comparability of effect sizes between the key result of the original and the replication study? Does it refer to the results of a meta-analysis across multiple replication attempts? Such debates are orthogonal to our goals here, but they illustrate the point that replicability is, in principle, quantifiable.

All of the other features of a high-quality science are quantifiable, too, even if the quantification process for them is every bit as complex as it is for replicability, or perhaps even more so. Consider external validity: There are compelling empirical demonstrations that laboratory and field studies align better in some domains of psychology than in others (Mitchell, 2012). If the first half-decade of psychology's evidentiary value movement has been devoted predominantly to understanding how certain research practices increase or decrease replicability, we hope that the next half-decade will be devoted also to aligning our research practices with *all* of the core features of a high-quality science. Such an effort could leverage the preliminary list in Figure 1 to pursue a robust discussion of what these core features are, and then generate quantifiable metrics for each of them. Then it could evaluate the field—or a topic, a research

program, a research practice, etc.—according to these metrics. Presumably, these metrics would be weighted according their importance in a given context (e.g., external validity might be more important for some research areas than for others), how precisely a given metric taps the underlying feature, and so forth.

Developing strong empirical metrics for the various features of a high-quality science will facilitate sharper tradeoff-based decision-making and help to ensure that any research practice is evaluated with respect to the full collection of desirable scientific features before being anointed a "best practice." Our view is that psychological science—and, presumably, the other empirical sciences—will be better served if the scholarly analysis of optimal research practices, a burgeoning discipline invigorated by the evidentiary value movement, ultimately produces recommendations that are framed in terms of broad, tradeoff-based principles or guidelines rather than in terms of strict policies that focus on one or a subset of features while neglecting the others.

As we move in this direction, it will be important for different laboratories to assess the costs and benefits of certain research practices for their particular subfield and research program, as FER2015 emphasized. There is no need for researchers to wait for definitive top-down recommendations before improving their research practices in light of new knowledge—adopting sharper theory, better statistics, or tighter methods relevant to their research program (Ledgerwood, 2014). Imagine a given researcher assessing, for example, how conducting fewer conceptual and more direct replications has bolstered her work's replicability (i.e., she chases fewer false positives) but harmed its cumulativeness (i.e., the narrower conceptual scope of her findings reduces their ability to contribute to a shared understanding of a topic area). If the field institutes mechanisms for the researcher to publically disseminate this cost-benefit evaluation, then we can all learn from these efforts. Across-the-board, top-down changes in research practices (procrustean editorial policies, caveatless exhortations for ever-larger samples, etc.) are likely to

be less beneficial than such locally generated empirical lessons, especially as we seek to make better-informed decisions about how to allocate our limited resources.

Indeed, LCL2016 engaged with this goal of using resources efficiently. They introduced an intriguing construct called "*N*-per-true-discovery," a metric for considering the possible tradeoffs involved when trying to maximize both replicability and discovery. We now offer a detailed response to LCL2016's discussion of these possible tradeoffs, demonstrating how their central conclusion—that larger *N*-per-study is uniformly better for efficiently discovering true findings (i.e., that there are no tradeoffs)—holds only if scholars care *exclusively* about being efficient with resources dedicated to replications. If scholars also care about being efficient with resources dedicated to the production original findings—if they care about resources for the field as a whole—the conclusion changes radically.

Reconsidering LCL2016's *N*-Per-True-Discovery Analysis

We argued in FER2015 that, when it comes to the sample sizes that researchers allocate to a particular study, "bigger is better" (p. 291). Yet we noted that this recommendation should be weighed against the opportunity costs that emerge when researchers draw from a fixed pool of resources:

> An important caveat is that the use of very large sample sizes—those larger than required for effect sizes to stabilize—will obviate the possibility of running other studies that might have been conducted with the excess participants and, consequently, increase theoretical false negatives. For example, in many cases, running 10,000 participants in one study focusing on one research question provides worse value—in terms of total scientific yield—than would allocating those 10,000 participants across a set of studies focusing on distinct research questions (or on replications of an initial effect). To our knowledge, scholars have not delved deeply into issues related to the opportunity costs associated with the allocation of research participants across studies. (p. 291)

One basic question implied in this excerpt is this: Does science benefit when researchers run few studies with larger *N*-per-study or many studies with smaller *N*-per-study? We were pleased to see LCL2016 tackle this question head-on—by simulating how a researcher's decision to allocate

her fixed pool of sample size resources (e.g., *N*=5,000) to many versus few studies alters the efficiency of scientific discovery. That is, given the risk of false positive and false negative errors in conducting a particular research study, and given that positive findings should ideally be replicated to afford confidence that it is a true discovery, how many studies should the researcher conduct with her *N*=5,000 to achieve the largest number of true findings?

After close inspection and correspondence with the third author of LCL2016 (Loving, personal e-mail correspondence, February 20, 2016), it became evident that LCL2016 focused solely on maximizing efficiency for researchers attempting to replicate the original researcher's finding (whether the replication is conducted by the researcher herself or by other scientists). That is, the assumptions baked into their simulations imply that there is nothing to be gained by striving for efficiency with the original researcher's resources—that the field is equally well-served by her discovering 1 or 10 or 100 true findings with her *N*=5,000. These assumptions would have been reasonable if LCL2016 had asked questions and drawn conclusions targeted exclusively toward the replication process. But LCL2016 asked questions and drew conclusions for the field as a whole—the collective resources available to the field for original research and replications—which produced a major disconnect between their simulations and their conclusions. In this section, we fix this disconnect by considering the efficiency not only of the process of replicating original results, but also of the process of generating those results in the first place. In doing so, we begin developing a data-driven picture of how tradeoffs can play out across a range of research scenarios.[3]

In their simulations, LCL2016 illustrated how a novel metric called the *true discovery rate*—the proportion of significant findings that reflect true rather than false positives—can aid researchers in making decisions about the most efficient use of participant pool resources.

[3] All simulations in this section adopt LCL2016's defaults unless otherwise stated, and all conclusions assume that the logic and math underlying LCL2016's app (http://shinyapps.org/apps/N-per-discovery/) are valid.

LCL2016 demonstrated that, to the extent that an original study was highly powered, replicators (whether the original researcher or other researchers) must invest fewer *N*-per-true-discovery. That is, when original researchers conduct few high-powered studies instead of many low-powered studies, replicators can use their resources more efficiently to determine whether the original result was a true rather than a false positive. For example, LCL2016's Table 3 shows that the *N*-per-true-discovery decreases from *N*=1,742 when the original research is statistically powered at 25% to *N*=917 when the original research is statistically powered at 95%.

Figure 2 presents *N*-per-true-discovery as a function of *N*-per-study used in the original research across the range from 10% power (*N*=12/study) to 95% power (*N*=311/study).[4] Consistent with LCL2016's conclusion, the dotted line slopes downward from left to right, indicating that *replicators* will spend their resources more efficiently (smaller *N*-per-true-discovery) when *original researchers* prioritize *higher-powered studies*. From this analysis, LCL2016 concluded that field-wide calls for increased statistical power will not reduce the pace of scientific progress but rather will foster the most efficient use of limited participant resources.

As noted previously, however, LCL2016 neglected to mention that this conclusion applies *only* to the replicator's resources (because the original researcher's *N*=5,000 were omitted from LCL's efficiency calculations). Once we account for the original researcher's resources—which is required if we wish to draw field-wide conclusions—the conclusion changes radically. For example, consider the dashed line in Figure 2, which illustrates *N*-per-true-discovery from the perspective of the original researcher—the *N*=5,000 resources without which there would be no studies to replicate. In direct opposition to LCL2016's conclusion, this line slopes upward from left to right, indicating that *original researchers* will be more efficient (smaller *N*-per-true-discovery) when they prioritize *lower-powered studies*. That is, when assuming that an original

[4] It is not entirely clear from LCL2016 or from their app what research design they are using in these simulations, but it appears to be a two-cell between-subjects design.

researcher wishes to spend her resources efficiently to unearth many true effects, plans never to replicate her own work, and is insensitive to the resources required to replicate her studies, she should run many weakly powered studies.

Given the conflicting efficiency goals between original researchers and replicators, whose goals shall we prioritize? Both roles are essential, of course, and researchers often play both roles—they generate original findings, and they replicate their own and others' findings. Whereas LCL2016 focused exclusively on the replicator's efficiency goals, our view is that, if the goal is to draw conclusions for the field as a whole (as LCL2016 sought to do), we must prioritize *the field's* efficiency goals rather than either the replicator's or the original researcher's in isolation. The solid line in Figure 2 illustrates *N*-per-true-discovery from the perspective of the field—when the original researcher's 5,000 participants are added to the pool of participants used by the replicator. This line forms a U-shaped pattern, suggesting that *the field* will be more efficient (smaller *N*-per-true-discovery) when original researchers prioritize *moderately powered studies*. In short, the replicator's efficiency is indeed maximized when the original researcher conducts higher-powered studies, but the original researcher's efficiency is maximized when she conducts lower-powered studies, and, most importantly, the field's efficiency is maximized when she conducts moderately powered studies.

Might these conclusions be one-off outliers resulting from LCL2016's default base-rate estimate of true hypotheses (10%) and effect size (d=.41)? Figure 3 addresses this question by illustrating the *N*-per-true-discovery as a function of *N*-per-original-study across the range from 10% to 95% power for the original studies—but this time for two different effect sizes (d=.41 and .80) and four different base-rate estimates of true hypotheses (10%, 25%, 50%, and 75%). Results from all of these simulations yield conclusions that align with those from Figure 2. For replicators,

all eight lines slope downward (Panel A). For original researchers, all eight lines slope upward (Panel B). Most importantly, for the field, all eight lines are U-shaped (Panel C).

Figure 3 reveals auxiliary findings of interest. For example, the curvilinear pattern in Panel C was especially pronounced for LCL2016's default of moderate effect size (d=.41) and low *a priori* likelihood of the hypothesis being true (10%). Larger effect sizes and higher *a priori* likelihoods revealed flatter curves, suggesting that weakly and highly powered studies are comparably efficient for the field.[5] Additionally, and perhaps startlingly, for hypotheses that are likely to be true, the most efficient use of field-wide N emerged when original researchers powered their studies poorly. For example, if a hypothesis is 75% likely to be true, which might be the case if the finding had a strong theoretical foundation, the most efficient use of field-wide N appears to favor power of ~25% for d=.41 and ~40% for d=.80.

Of course, LCL2016's app does not incorporate all possible considerations when it comes to evaluating efficiency; for this reason, we are reluctant to make any real-world recommendations based on these simulations. For example, running weakly powered studies implicitly assigns no cost to Type II errors and is clearly unwise if researchers wish to draw conclusions from null findings (i.e., only with large samples can one conclude from a null finding means that the effect is small or nonexistent). Also, if replications cannot be published and publicized, then false positives might live as zombie findings in the published literature despite (file-drawered) replication failures. Furthermore, instead of replicating only positive findings, perhaps there is also value in vigorously replicating negative findings to avoid the possibility that negative findings needlessly discourage future attempts to unearth important phenomena. Finally, these simulations also assume that there is just one true effect size for each hypothesis and no heterogeneity, an assumption that is often unfounded (Klein et al., 2014; McShane & Böckenholt, 2014).

[5] In the absence of p-hacking and file-drawering, a meta-analytic synthesis is likely to yield comparable conclusions regardless of whether it includes many small-N studies or few large-N studies (Stroebe, in press).

Simulations are only useful insofar as their underlying assumptions map onto real data and real research practices; neither LCL's simulations (nor ours) sufficiently deal with complexities like the costs of Type II error or effect size heterogeneity. Ultimately, informed recommendations will emerge with the aid of broad and flexible tools for calculating efficiency of resource expenditure that incorporate such complexities. One intriguing effort along these lines was recently offered by Miller and Ulrich (in press), who introduced the idea, and developed a formal model, of *total research payoff*—the greatest scientific yield for the investment of a given set of resources. Simulations derived from this model yielded the conclusion that "optimal choices for researchers depend on the characteristics of their research area, and this means that it is impossible to identify a universally optimal set of choices that would apply across all areas" (p. xx). This conclusion aligns precisely with the conceptual analysis offered in FER2015, but it misaligns with LCL2016's conclusion that ever-larger sample sizes are good for the field as a whole, as exemplified in these excerpts (emphasis added):

- "increasing sample sizes, while potentially costly for individual researchers, is crucial *for the field* if we wish to make important and replicable discoveries" (p. xx)
- "in actuality larger sample sizes and the execution of replication studies is required *for overall scientific progress of the collective field*" (p. xx)

LCL are not merely arguing that we need to increase sample sizes from, say, small to medium; on the contrary, their simulations and conclusions do not specify any sort of upper bound on their "larger sample sizes" conclusions—they do not account for tradeoffs. For example, in the calculations underlying their app, increasing power *always* increases efficiency. It is this lack of upper bound that we question. We, and FER2015, certainly agree with the need for the field to use larger samples as a normative practice (larger than typical circa 2011, for example), but "bigger is

always better" is unlikely to be an adequate heuristic as researchers decide how to be efficient with the field's collective resources, not to mention their own.

Despite our disagreements with LCL2016, however, we are pleased that all parties in this debate are engaging seriously with the ideas (a) that determining which research practices are optimal in a given context requires a consideration of tradeoffs (at least in theory) and (b) that such determinations can be based on data (including simulated data). Although LCL2016's simulations calculated efficiency only for the replication process, which does not really permit one to draw conclusions for the discipline as a whole, their innovative app has helped to underscore an important set of tradeoffs for researchers to consider when making resource-allocation decisions.

Near-Consensus on Open Practices

LCL2016 claimed that their article challenges statements by FER2015 and others not only regarding tradeoff considerations in terms of sample sizes, but also about potential costs of open practices. But with regard to open practices, we see little disagreement between their views and the ones we expressed in FER2015 (aside from minor quibbles about the circumstances under which preregistration has more versus less value, for example). Here is how LCL2016 characterized their philosophy on open science (emphasis in original): "Our personal open science position advocates a *sufficiently open science*, which is science that is sufficiently open to allow for (1) accurate peer-review evaluation, (2) independent verification of analytic reproducibility of results, and (3) the execution of diagnostic direct replications" (p. xx). This position entirely aligns with FER2015—and, we suspect, with the opinions of the vast majority of researchers in the field. Apparent disagreements between FER2015 and LCL2016 tended to result from their mischaracterization of our views. For example, they suggested (pp. 27-28) that FER2015's Discussion-section comments on intellectual property applied to authors shielding their data from scholars wishing to evaluate a submitted or published finding. But FER2015 were clear on this

point—our comments applied only to the possibility that a researcher could be scooped with her own data if policies were to require that data from unpublished variables be made publicly available *for others to publish novel findings on their own*.

Our tone regarding this issue—suggesting that addressing it successfully "will require collaborations among, at minimum, psychologists, ethicists, and legal scholars" (FER2015, p. 293)—is illustrative of our broader approach to the complex issues under discussion in the evidentiary value movement. Rather than anointing specific distal means as "best" practices, our view is that although we, as a field, seem to be moving toward "better" practices, we have not yet done the sort of nuanced, tradeoff-based thinking required to warrant top-down, one-size-fits-all rules or norms.

Conclusion

Since 2011, psychological science has witnessed major changes in its standard operating procedures—changes that hold great promising for bolstering the replicability of our science. We have come a long way, we hope, from the era in which editors routinely encouraged authors to jettison studies or variables with ambiguous results, the file drawer received only passing consideration, and $p<.05$ was the statistical holy of holies. We remain, as in FER2015, enthusiastic about such changes.

Our goal is to work alongside other metascientists to generate an empirically grounded, tradeoff-based framework for improving the overall quality of our science. Scholars must be willing to alter their research practices and assess how the quality of their scientific output changes as a result. We, as a field, need many more data—beyond John et al. (2012) and Fiedler and Schwarz (2016)—regarding what researchers' actual practices (rather than their *assumed* practices) really look like, and what the implications of those practices are. We need simulations that leverage these data to understand how we can strengthen our science. These are significant

challenges, and our ability to meet them depends upon us leveraging our core strengths as scientists.

To sharpen our understanding of best research practices, we need a much greater emphasis on tradeoffs among the features that a flourishing discipline should possess. We have certainly benefited from the intensive recent emphasis on identifying which practices increase versus decrease replicability. But we must evaluate the extent to which a given research practice strengthens our discipline across the full range of scientific desiderata. In particular, we must focus greater attention on establishing which features are most important in a given research context, the extent to which a given research practice influences the alignment of a collective knowledge base with each of the relevant features, and, all things considered, which research practices are optimal in light of the various tradeoffs involved. Such an approach will certainly prioritize replicability, but it will also prioritize other features of a high-quality science, including discovery, internal validity, external validity, construct validity, consequentiality, and cumulativeness.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333-342.

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

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, *54*(4), 297-312.

Cartwright, N. (2007). *Hunting causes and using them: approaches in philosophy and economics*. Cambridge University Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *220*(571-581), 291-333.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS one*, *11*(2), e0149794.

Fabrigar, L. R., & Wegener, D. T. (in press). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7,* 661–669. http://dx.doi.org/10.1177/1745691612462587

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

Finkel, E. J. (2016, January). *Taking stock of the evidentiary value movement: Where from here?* Paper presented at the annual SPSP Training Preconference preceding the meeting of the Society for Personality and Social Psychology (SPSP), San Diego, CA.

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

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, *21*, 1180-1187.

Inzlicht, M. (2015). Check yourself before you wreck yourself. Viewed 3/21/2016 at http://sometimesimwrong.typepad.com/wrong/2015/04/guest-post-check-yourself-before-you-wreck-yourself.html.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142-155.

LeBel, E. P., Campbell, L., Loving, T. J. (2016). Benefits of Open and High-Powered Research Outweigh Costs. *Journal of Personality and Social Psychology*.

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, *9*(3), 275-277.

Ledgerwood, A., Soderberg, C., & Sparks, J. (in press). Designing a study to maximize informational value. In J. Plucker & M. Makel (Eds.), *Doing good social science: Trust, accuracy, & transparency.* Washington, DC: American Psychological Association.

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

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *The American Behavioral Scientist*, *25*(2), 179-210.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612-625.

Mischel, W. (2006). Bridges toward a cumulative psychological science. In P. A. M. Van Lange (Ed.), *Bridging social psychology* (pp. 437-446). Mahwah, NJ: Erlbaum.

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, *7*, 109–117.

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

Pinker, S. (2014). *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. New York, NY: Penguin.

Schuster, M. A., Stein, B. D., Jaycox, L. H., Collins, R. L., Marshall, G. N., Elliott, M. N., ... & Berry, S. H. (2001). A national survey of stress reactions after the September 11, 2001, terrorist attacks. *New England Journal of Medicine*, *345*(20), 1507-1512.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science*, 9, 552-555. DOI: 10.1177/1745691614543974

Simonsohn, U. (2012). It does not follow evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in Press). *Perspectives on Psychological Science*, *7*(6), 597-599.

Smith, E. R., & Mackie, D. M. (2000). *Social psychology* (2nd ed.). Philadelphia, PA: Psychology Press.

Stroebe, W. (in press). Are most published social psychological findings false? *Journal of Experimental Social Psychology*.

Figure 1. How to achieve a high-quality science: Engage in research practices (distal means) that increase net alignment with the core desiderata of science (proximal means).

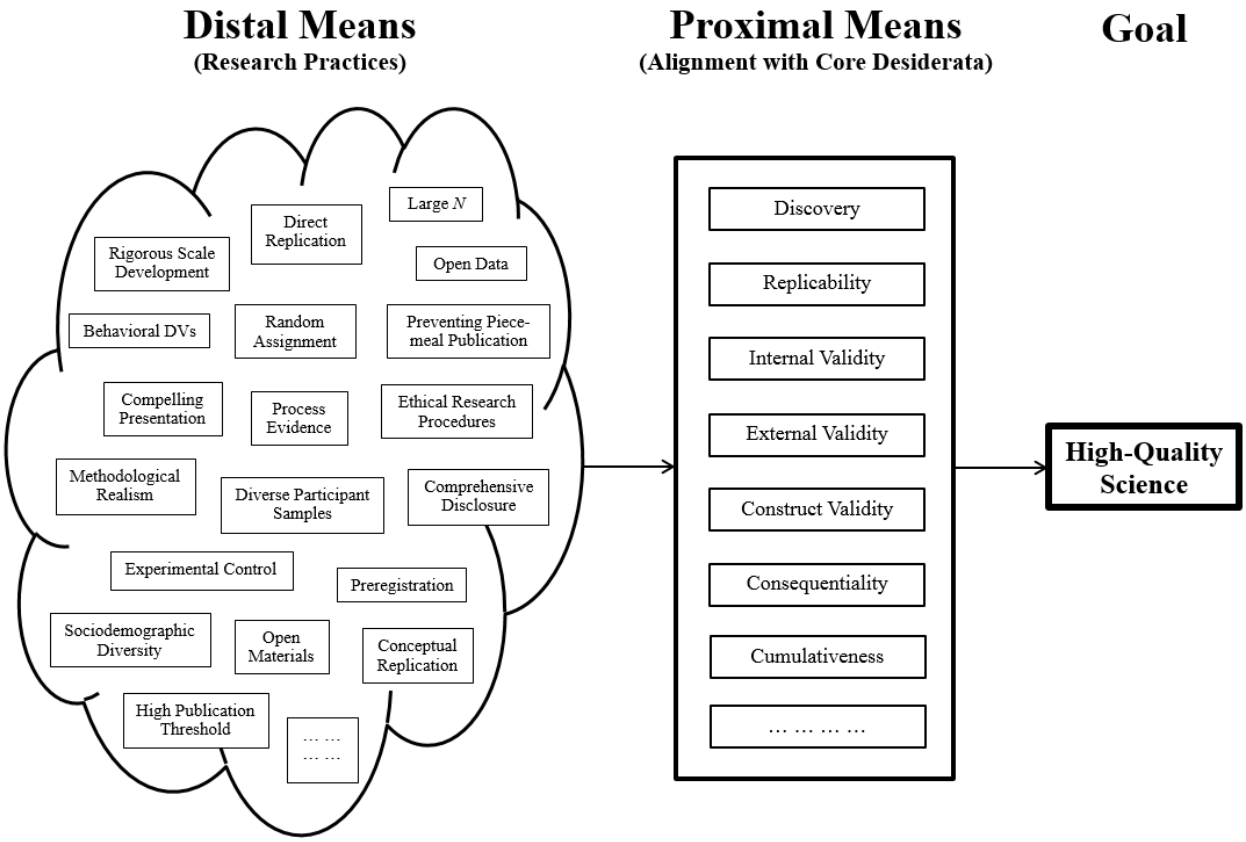
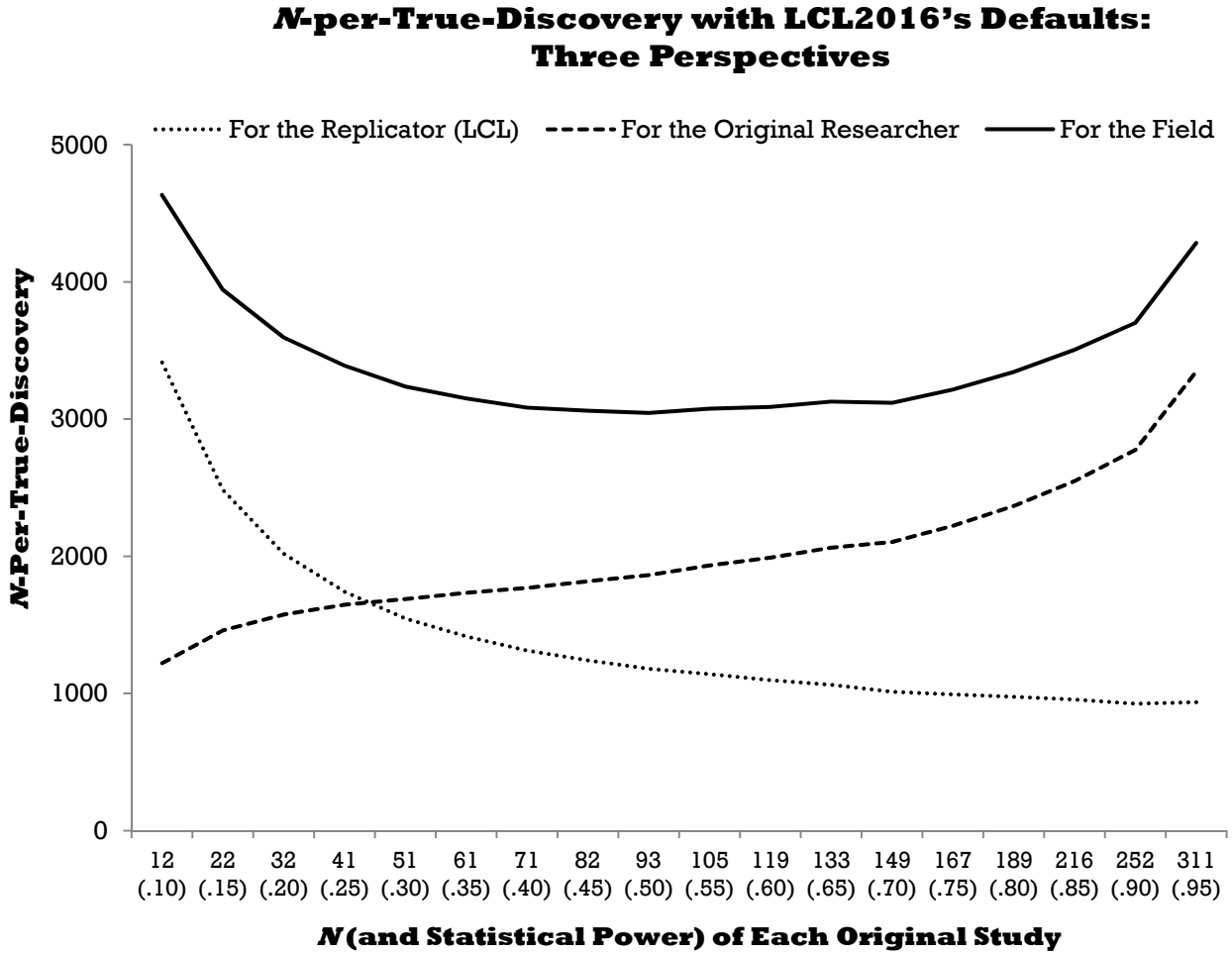


Figure 2. *N*-per-true-discovery for the replicator, for the original researcher, and for the field when adopting LCL2016's defaults.

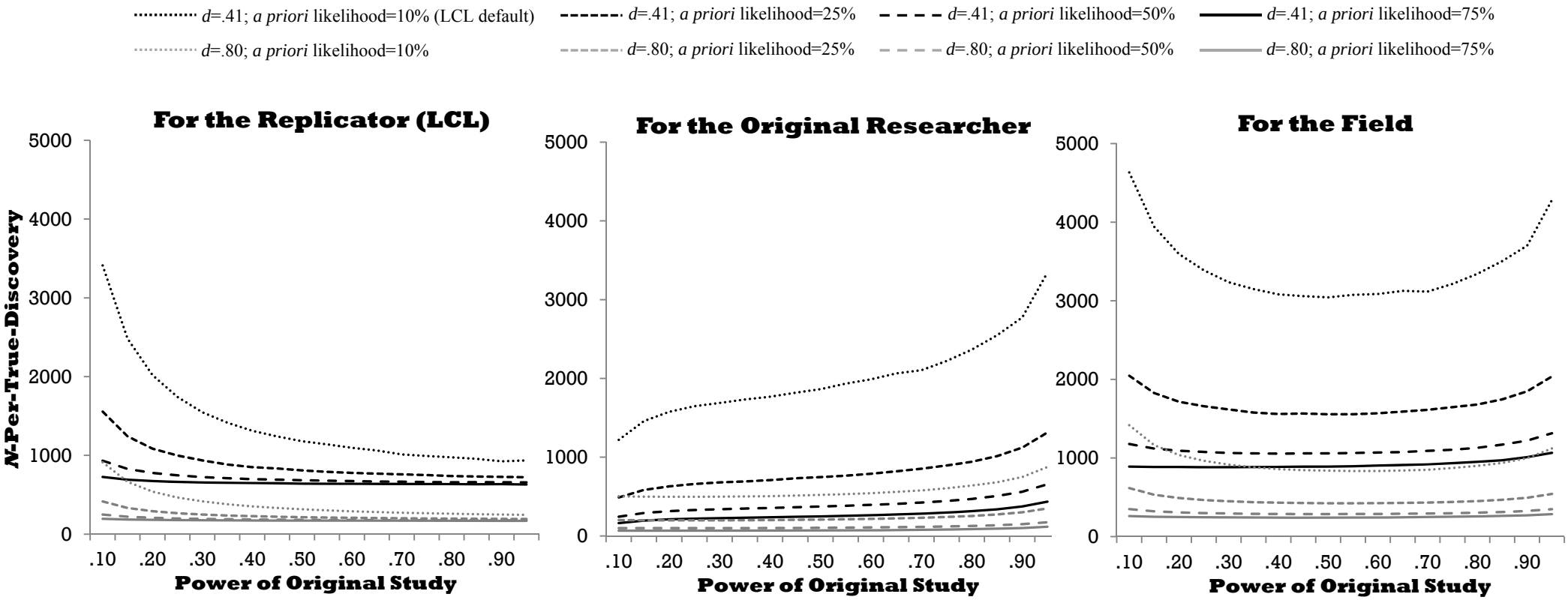


Note. All plotted simulations adopt LCL2016's six defaults: α=.05, *d*=.41, total *N* available to the original researcher=5,000, base-rate of true hypotheses=.10, number of replications per statistically significant original study=2, and power of replications=.95. These estimates are calculated using the data for (a) "Total N of replication studies required to distinguish true from false discoveries," (b) "No. studies yielding positive results," and (c) "True discovery rate (TDR)" from the LCL2016 app (http://shinyapps.org/apps/N-per-discovery/; data collected February 20-21, 2016). We used a separate spreadsheet (available from the first author on request) to plot the lines according to these formulae (with the *a*, *b*, and *c* referring to the parameters from the previous sentence):

- Dotted line—the replicator's perspective=a/(b×c)
- Dashed line—the original researcher's perspective=5,000/(b×c)
- Solid line—the field's perspective (including both the original researcher's and the replicator's perspectives)=(a+5,000)/(b×c).

The results for the dotted line align with those from LCL2016's Table 3, although that table exhibits rounding error. For example, the *N*=311 x-axis data point for the replicator (dotted line) in the figure above reads 917 in LCL's Table 3 (see the bottom-right cell in that table), but the true y-axis value (depicted in the figure here) is 937.2. LCL (LeBel, Campbell, and Loving) focused exclusively on the replicator's perspective (dotted line), entirely neglecting the original researcher's perspective (dashed line) and, crucially, the field's perspective (solid line).

Figure 3. *N*-per-true-discovery (a) for the replicator, (b) for the original researcher, and (c) for the field as a function of both the true effect size, *d*, and the *a priori* likelihood that the hypothesis is true.



Note. Aside from the variation in assumptions regarding the actual effect size and the *a priori* likelihood that the hypothesized effect is true, all calculation procedures mirror those for Figure 2.