Benefits of Open and High-Powered Research Outweigh Costs

Etienne P. LeBel and Lorne Campbell
University of Western Ontario

Timothy J. Loving
University of Texas at Austin

Several researchers recently outlined unacknowledged costs of open science practices, arguing these costs may outweigh benefits and stifle discovery of novel findings. We scrutinize these researchers' (a) statistical concern that heightened stringency with respect to false-positives will increase false-negatives and (b) metascientific concern that larger samples and executing direct replications engender opportunity costs that will decrease the rate of making novel discoveries. We argue their statistical concern is unwarranted given open science proponents recommend such practices to reduce the inflated Type I error rate from .35 down to .05 and simultaneously call for high-powered research to reduce the inflated Type II error rate. Regarding their metaconcern, we demonstrate that incurring some costs is required to increase the rate (and frequency) of making true discoveries because distinguishing true from false hypotheses requires a low Type I error rate, high statistical power, and independent direct replications. We also examine pragmatic concerns raised regarding adopting open science practices for relationship science (preregistration, open materials, open data, direct replications, sample size); while acknowledging these concerns, we argue they are overstated given available solutions. We conclude benefits of open science practices outweigh costs for both individual researchers and the collective field in the long run, but that short term costs may exist for researchers because of the currently dysfunctional academic incentive structure. Our analysis implies our field's incentive structure needs to change whereby better alignment exists between researcher's career interests and the field's cumulative progress. We delineate recent proposals aimed at such incentive structure realignment.

Keywords: open science practices, independent replication, cumulative knowledge, analytic and design flexibility

A growing open science movement has emerged in psychology and related social and biomedical sciences. Though diversity in opinion exists within the movement, researchers have collectively called for major modifications to research practices and journal policies. These proposed changes reflect a concerted bid to increase the cumulative nature and validity of published findings so as to accelerate (theoretical) understanding of human behavior. For example, some open science proponents have called for more transparent analysis and reporting of studies (Campbell, Loving, & LeBel, 2014; LeBel et al., 2013; Simmons, Nelson, & Simonsohn, 2011) and more open sharing of experimental materials and raw data (Nosek & Bar-Anan, 2012). Others have advocated for the preregistration of studies and analytic plans (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) and execution of adequately powered studies (Asendorpf et al., 2013), while others have called for a stronger focus on the execution (and publication) of independent direct replication attempts (Pashler & Harris, 2012; LeBel & Peters, 2011). Overall, researchers have begun to heed such recommendations, as evidenced by the increasing number of psychologists who publicly post their data (Wicherts, 2015), publish preregistered studies (Chambers, 2013; e.g., Matzke et al., 2015), and execute, report, and publish independent direct replications in prominent journals (Srivastava, 2012; e.g., LeBel & Campbell, 2013).

Though the open science movement continues to grow and journals are implementing new norms and policies, some researchers have raised concerns regarding possible unacknowledged (or underappreciated) costs of adopting these new practices. It is argued that such costs may in fact outweigh any benefits, particularly by limiting the discovery of novel findings going forward. The primary objective of this article is to scrutinize these potential costs with an eye toward evaluating the cost–benefit trade-off of adopting open science practices. Specifically, we examine epistemological concerns that the heightened stringency of open science practices involves undesirable tradeoffs in terms of (a) statistical issues (i.e., reducing false-positives will by definition increase false-negatives), and (b) metascientific issues (i.e., requiring larger

Etienne P. LeBel and Lorne Campbell, Department of Psychology, University of Western Ontario; Timothy J. Loving, Department of Human Development and Family Sciences, University of Texas at Austin.

We thank Brent Donnellan, Richard Lucas, Hal Pashler, Rogier Kievit, and E.-J. Wagenmakers for valuable feedback on an earlier version of this article. We also thank Felix Schönbrodt for hosting the Shiny App we created for Table 3.

Correspondence concerning this article should be addressed to Etienne P. LeBel, Department of Psychology, University of Western Ontario, Social Science Centre, London, Ontario, N6A 5C2 Canada. E-mail: etienne.lebel@gmail.com

http://dx.doi.org/10.1037/pspi0000049

samples and replications engender opportunity costs that will decrease the rate of new discoveries). Our analysis of the relative costs versus benefits of open science practices for both individual researchers, as well as the field overall, leads us to conclude that the benefits outweigh the costs for both individual researchers and the collective progress of our field in the long run, but that the current incentive structure needs to change for this to be true for individual researchers in the short term. As a secondary objective, we scrutinize the pragmatic concerns raised regarding the application of open science practices to a specific subfield of psychology, relationship science, with respect to preregistration, open materials, open data, close replications, and sample size recommendations. Investigating the merits of such epistemological and pragmatic concerns is critical given that overestimating the costs of adopting more open science practices may interfere with the growing momentum of arguably the most important methodological revolution to have emerged in the history of modern science (Borsboom, 2014; Chambers, 2014).

Epistemological Concerns of Adopting More Open Science Practices

The primary epistemological concern that has been raised centers on the assumption that decreasing Type I errors will necessarily increase Type II errors,[1] a trade-off that can have important unintended negative consequences for scientific progress. For example, Fiedler, Kutzner, and Krueger (2012) state that ". . . setting the criterion for the funding, investigation, and publication of findings at an *extremely conservative* level is indeed a strategy that reduces false positives, but this reduction comes at the price of an unknown increase in false negatives" (p. 663; emphasis added). Relatedly, Murayama, Pekrun, and Fiedler (2014) argued that recent contributions of researchers within the open science movement have placed a disproportionate emphasis on reducing *inflated* Type I error rates at the expense of paying sufficient attention to extant methodological practices that reduce false-positive rates (e.g., multiple replication-study papers, a priori theoretical hypotheses). Such unbalanced efforts are believed to undermine scientific progress by constraining research endeavors unnecessarily, causing researchers to underestimate the validity of their "discoveries," influencing journal editors to make "unreasonable decisions," and potentially causing the public to consider psychology as less scientific than it really is (Murayama et al., 2014).

In a similar spirit, Wigboldus and Dotsch (2015) argued that ". . . in science, there is no such thing as a questionable research practice when it concerns data analyses" (p. 4), as long as researchers share the details of all analyses conducted.[2] They argue that the label "questionable research practices" will likely discourage researchers from data exploration that may help reduce false-positive rates but will also very likely increase false-negative rates by virtue of limiting the opportunity for new discoveries. Furthermore, they argue that data exploration is not inherently questionable; rather, it is not reporting the details of data exploration that is questionable, and thus, they propose the new label "questionable reporting practices" to incentivize both data exploration and heightened reporting of these analytic endeavors. Wilson (2014) has also argued that researchers within the open science movement have placed inordinate emphasis on false-positives relative to false-negatives and this bias is particularly problematic in the instance when one researcher is not able to replicate the results of another researcher. These failures to replicate potentially represent false-negatives and damage both the reputation of the original researcher and the progression of science in terms of discovering novel and groundbreaking findings. Wilson argued that "there is just as much evidence that we have a crisis of false-negatives as we do a crisis of false-positives" (p. 1). Specifically, false-negatives can arise from inexperienced replicators lacking expertise or via the confirmation bias of skeptical replicators who may act in ways that reduce the likelihood of replicating an original finding—a practice Wilson dubs "*p* squashing."

Finally, consistent with many of the positions reviewed above, Finkel, Eastwick, and Reis (2015) argued that careful consideration of epistemological and pragmatic issues must be addressed to maximize the value of the open science movement's recommendations. Concerning epistemology, Finkel et al. argued that open science proponents have overly focused on reducing false-positive rates and "seem not to have accounted sufficiently for false-negative rates" (p. 280). This type of thinking, according to Finkel et al., is shortsighted because "heightened stringency regarding false-positive rates . . . will almost certainly increase false-negative rates" (p. 278; see also Lieberman & Cunningham, 2009). According to Finkel et al.'s (2015) error-balance perspective, one type of error should not be valued more than another.

Before examining these concerns, we would like to acknowledge the importance of these researchers' contributions. First, they provide unique views on how the open science movement can help psychologists optimize research practices to increase scientific discovery and validity of the psychological literature. Second, these arguments derive from *within* a field (i.e., social psychology) that is currently at the center of the overall open science debate. It is an exciting and positive development that such discussions are now occurring among researchers *within* the field rather than by researchers *outside* the field (which was previously typically the case). Finally, these contributions are important because they will help sharpen and advance the discussion of the merits of open science practices which ultimately will improve research practices in psychology.

The common assumption running through the opinions reviewed above is that open science practices—though valuable in several respects—are likely to increase Type II errors, manifested in *literal* or *theoretical* false-negatives.[3] Below, we scrutinize the accuracy of this zero-sum assumption. We focus on statistical and then metascientific concerns that scientific practices designed to reduce Type I errors will necessarily increase Type II errors.

[1] Following standard terminology, a Type I error involves incorrectly concluding an effect exists when in reality it does not, whereas a Type II error involves incorrectly concluding an effect does *not* exist when in reality it does.

[2] Importantly, this proposal relies on researchers being completely transparent with respect to their data analyses across studies.

[3] A *literal* false-negative is erroneously concluding—based on an empirical study—that no effect exists when one in fact exists, whereas a *theoretical* false-negative involves the same error *without* having executed an empirical study (failure to pursue a valid alternative hypothesis, see Fiedler et al., 2012, p. 663).

Table 1
Error Rates, Power Level, and Sample Sizes of Status Quo and Open Science Approaches

Research approach	α	Type I error rate	Type II error rate	Statistical power	*N*
Status quo	.05	.35	.75	.25	40
Open Science 1	.05	.05	.75	.25	40
Open Science 2	.05	.05	.20	.80	190
Error equivalence	.05	.05	.05	.95	311

Note. α = nominal alpha criterion level. Statistical power values refer to the probability of detecting an effect size of $d = .41$, the mean effect size of social psychology studies over the past 100 years (Richard et al., 2003), with the sample sizes (*N*) used and/or advocated by proponents of the different research approaches.

The Type I Versus Type II Error Trade-Off: Statistical Considerations

Decreasing the nominal α level—for example, from .05 to .01—would certainly increase the Type II error rate. But open science proponents are not proposing such a recommendation (cf. Colquhoun, 2014). Rather, a strong case can be made that, in practice, the Type I error rate is in fact much higher than the nominal α level of .05 because of study design and data analytic flexibility (Simmons et al., 2011). Such flexibility—also known as *researcher degrees of freedom*—refers to the typically large number of ways a researcher can analyze data to test a particular hypothesis, which can easily lead to an effective false-positive rate that is dramatically higher than the nominal $\alpha = .05$, even for the best-intentioned researcher. For example, typically more than one dependent variable is included in a study, observations are excluded using post hoc criteria, additional data are collected after results are known, or conditions are combined or contrasted in ways that may not have been specified a priori. Based on simulations, Simmons et al. vividly demonstrated that the combined effect of (unintentionally) exploiting such researcher degrees-of-freedom easily increases the Type I error rate beyond .50 (see their Table 1, p. 1361). Importantly, such exploitation of design and analytic flexibility does not necessarily stem from malicious intent, but rather may be driven by the fact that such decisions are most often complex and ambiguous and hence researchers can easily fall prey to confirmation bias (Nickerson, 1998) and/or motivated reasoning (Kunda, 1990).[4]

Is there evidence that researchers engage in such practices? According to a large-scale survey of over 2,000 psychologists by John, Loewenstein, and Prelec (2012), the answer is *yes* (but see Fiedler & Schwarz, 2016). The majority of respondents admitted that they have (a) not always reported all dependent measures, (b) collected more data after results were known, and (c) reported only studies that "worked" (see their Table 1, p. 525). Furthermore, about 40% of respondents admitted to excluding observations after looking at the impact of doing so on the results and almost 30% admitted to not reporting all of a study's conditions. Of course, such admissions do not necessarily indicate that researchers engage(d) in these practices in each and every one of their studies. However, a recent analysis of published psychology studies from a competitive grant program where all study materials and data were made available, does suggest that these questionable research practices occur fairly frequently. Specifically, 40% of studies failed to report all experimental conditions, 70% of studies failed to report all outcome variables included in questionnaires, and the reported effect sizes were almost twice as large and 3 times more likely to be statistically significant compared to unreported effect sizes (Franco, Malhtra, & Simonovits, 2015). Additionally, O'Boyle, Banks, and Gonazalez-Mulé (2014) compared results of the same research projects as presented in dissertations and subsequent published papers and found a much higher ratio of supported to unsupported hypotheses in the peer-reviewed publications; this metamorphosis of "ugly" findings into "beautiful" results (what they termed the *chrysalis effect*), was achieved via the application of questionable research practices (e.g., dropping statistically nonsignificant results, altering hypotheses to be consistent with unexpected results). Additional evidence from PsychDisclosure.org also suggests such practices are indeed common and even sometimes demanded by editors and reviewers during the peer-review process (LeBel et al., 2013, see Figure 1, p. 427).

Given this state of affairs, proponents of the open science movement have recommended using more open science practices, such as publicly posting data and materials, conforming to higher reporting standards, and preregistering study hypotheses when feasible, not to reduce the α level below .05, but rather to bring the Type I error rate (likely > .35) back down closer to the nominal α level of .05. Doing so will ensure that studies have higher evidentiary value and hence are more likely to contribute to a cumulative knowledge base. Additionally, the position that the open science movement's suggestions disproportionately address false positives at the expense of false negatives is questionable given that more than 10 distinct voices within the open science movement have argued that we need to increase statistical power in order to *decrease* our already high Type II error rate (Asendorpf et al., 2013; Bakker, Hartgerink, & Wicherts, 2012; Button et al., 2013; Fraley & Vazire, 2014; Ioannidis, 2005, 2012; Lakens & Evers, 2014; Lucas & Donnellan, 2013; Nosek et al., 2012; Pashler & Harris, 2012; Perugini, Gallucci, & Costantini, 2014; Schimmack, 2012; Schimmack & Dinolfo, 2013; Simons, 2014).

To illustrate these points, Table 1 lists the alpha level, Type I and II error rates, sample size (*N*), and power levels of four different research approaches, assuming an effect size of $d = .41$ ($d = .41$ is the average effect size [*not* corrected for publication bias] of social psychology studies in the past 100 years as estimated in a meta-meta-analysis by Richard, Bond, & Stokes-Zoota, 2003). The first row, representing the status quo approach in social

[4] Indeed, Silberzahn et al.'s (2016) recent "one data set, many analysts" metascientific investigation vividly illustrates how flexibility in data analytic choices can substantially influence results. Asked to test the same hypothesis (is soccer players' skin color associated with probability of being given a red card) using the same data set, 29 different research teams ended up reaching a wide variety of conclusions.

psychology, exhibits a Type I error rate of about .35 due to design, analytic, and reporting flexibility (John et al., 2012; Simmons et al., 2011). At the same time, the status quo approach exhibits an even higher Type II error rate of about .75 (Bakker et al., 2012) given the typically small sample sizes used (i.e., $N = 40$ given the median cell size of $n = 20$ in papers in *Psychological Science* and *Journal of Personality and Social Psychology*, Simonsohn, 2014b; see also Marszalek, Barber, Kohlhart, & Holmes, 2011).

Given this state of affairs, open science proponents ("Open Science 1," second row) have called for more transparent research reporting and preregistration to reduce the inflated Type I error rate closer to the nominal $\alpha = .05$ level. Reducing only the Type I error rate in this way helps minimize false positives, but does not reduce the Type II error rate or increase statistical power.

That is why open science proponents have also called for the use of larger sample sizes to reduce the inflated Type II error rate ("Open Science 2," third row) so researchers have sufficient power to detect effects of realistic magnitude (i.e., raising statistical power from .25 to .80 to detect a $d = .41$ requires an $N = 190$ rather than $N = 40$). Hence, the open science movement's recommendations are sensitive to the problems of each type of error and are thus geared toward decreasing both Type I *and* Type II errors. Lastly, for illustrative purposes, an error equivalence approach (fourth row) that balances Type I and II error rates at .05 (one possible way to formalize Finkel et al.'s, 2015, "error-balance" approach), requires even larger sample sizes (i.e., $N = 310$ to have .95 power to detect $d = .41$).

The argument that open science practices will reduce both Type I and Type II errors has been advocated previously. Lakens and Evers (2014)—in the context of discussing how to increase the informational value of studies—explicitly state, "Increasing the statistical power of a study increases the likelihood of finding true positives while decreasing the likelihood of finding false-negatives and, therefore, increases the informational value of studies" (p. 284; see also their Table 2 on how to power studies adequately). In another instance, Asendorpf et al. (2013; a 16-author paper)—when discussing the fundamental logic of Type I and II errors—state that ". . . this may give the *misleading* impression that one has to choose between the two types of errors when planning a study. Instead, it is possible to minimize both types of errors simultaneously by increasing statistical power" (p. 110, emphasis added; see also Maxwell, Kelley, & Rausch, 2008). In yet another article, Button et al. (2013) explicitly recommend that researchers always perform a priori power calculations and work collaboratively with other labs if insufficient resources exist to achieve adequate statistical power (see their Box 2, p. 10).

Underscoring the importance of executing sufficiently powered-studies, Perugini, Gallucci, and Costantini (2014) have even proposed a "safeguard power analysis" approach that overcomes the serious problem that effect size estimates from original studies are noisy and virtually always overestimated due to publication bias (Simonsohn, 2013). The logic of this approach is to calculate power based on a lower bound of the original effect size estimate, which "safeguards" a researcher in the event the true effect size is indeed lower than is reported in the original published study. Schimmack and Dinolfo (2013) have argued that researchers should use their limited resources to conduct fewer studies with high statistical power and that "editors should focus on total power and reward manuscripts that report studies with high statistical power because statistical power is essential for avoiding Type I and II errors" (p. 133; see also Maxwell, 2004). Lucas and Donnellan (2013) have even gone so far as to suggest that editors and reviewers should strongly consider desk-rejecting manuscripts that include underpowered studies. They specifically recommend that authors need to demonstrate they had sufficient statistical power to detect plausible effect sizes typical for their field unless compelling evidence is presented suggesting the specific effect under investigation is of particularly large magnitude.

Taken together, it is clear that several voices within the open science movement have explicitly stated and recommended executing high-powered studies to reduce Type II, as well as Type I, error rates to generate more cumulative knowledge. The concern that heightened stringency regarding false-positives will increase false-negatives, therefore, is questionable given that open science proponents do not propose to reduce the nominal alpha level but rather recommend open science practices to bring the inflated Type I error rate—due to design and analytic flexibility—back down to the nominal alpha level of .05. Furthermore, several voices within the open science movement have explicitly called for higher-powered research; thus, open science practices effectively reduce both Type I *and* Type II error rates (at the statistical level).

Table 2
True Discovery Rates for Different Type I Error Rates, Power Levels, and Base Rates of True Hypotheses for Different Research Approaches

Research approach	Type I error rate	Power	Proportion of studies yielding true positives	Proportion of studies yielding false positives	True discovery rate
Base rate of true hypotheses = .10					
Status quo	.35	.25	.025	.315	.074
Open Science 1	.05	.25	.025	.045	.357
Open Science 2	.05	.80	.080	.045	.640
Error equivalence	.05	.95	.095	.045	.679
Base rate of true hypotheses = .25					
Status quo	.35	.25	.063	.263	.192
Open Science 1	.05	.25	.063	.038	.625
Open Science 2	.05	.80	.200	.038	.842
Error equivalence	.05	.95	.238	.038	.864

Note. True discovery rates can be calculated for other Type I error rates, power levels, and base rates of true hypotheses using Zehetleitner and Schonbrodt's (2016) web application available at http://87.106.45.173:3838/felix/PPV/.

The Type I Versus Type II Error Trade-Off: Metascientific Considerations

A secondary—though arguably equally important—epistemological concern that has been raised involves a metascientific trade-off whereby it is believed that heightened stringency in terms of requiring larger sample sizes and execution of replications may engender opportunity costs that will decrease the rate of making new scientific discoveries by increasing theoretical false negatives and consequently stifling overall scientific progress. Given individual researchers have limited resources, the logic underlying this concern seems to be that a lab cannot increase the sample size (and/or execute replications) without electing to conduct fewer original studies. Consequently, running fewer original studies may translate to fewer scientific discoveries across the board. For example, Finkel et al. (2015) state,

> In some cases, stricter publication policies emerging in the wake of the evidentiary value movement will replace a true positive with a (literal or theoretical) false negative, clearly a bad trade. The issue is that nobody knows what the actual effect is in the broader population—otherwise hypothesis tests would be superfluous. Our point here is not that heightened stringency regarding false-positive rates is bad, but rather that it will almost certainly increase false-negative rates, which renders it less than an unmitigated scientific good. (p. 278)

In a similar spirit, Fiedler et al. (2012) argued that isolated discussions about—and interventions designed to reduce—false-positives in the published literature without considering the importance of false-negatives is more likely to hinder rather than promote the growth of knowledge in psychology. Fiedler et al. contend that rewarding strong inference (Platt, 1964) is a more productive approach to achieving progress than is tightening standards for the publication of original findings. In their own words,

> Notwithstanding the worthwhile aims of the call for more control of false positives, science would hardly prosper if unrealistically high thresholds inhibited the publication and dissemination of innovative ideas, discouraged (young) scientists from conducting bold research . . . and forced researchers to concentrate on the reliability of local research questions rather than engaging in open-minded validity tests of global research questions. (p. 667)

In both cases, the common theme appears to be that the heightened stringency of open science practices entails opportunity costs (due to limited resources) that will increase theoretical false-negatives and hamper overall scientific progress.

We respectfully disagree with such a position and present evidence from several metascientific investigations to support our point. To individual researchers, embracing more open science practices does indeed involve new non negligible costs (e.g., making data openly available takes additional time, larger samples take longer to collect and typically cost more money, executing independent direct replications takes time away from original studies). Given that time is a zero-sum game, adopting more open science practices by an individual researcher will result in publishing fewer papers, likely containing fewer studies, a cost that may seem to outweigh the benefits of adopting more open practices (e.g., citation advantage of open data, see Piwowar & Vision, 2013; improved organization and archiving of study materials, and lower likelihood of losing track of data over time, see Nosek & Bar-Anon, 2012). Hence, individual researchers may be less able to benefit from an open science approach given the current incentive structure that rewards publishing more versus fewer papers, despite the fact that generating cumulative knowledge within a field is dramatically hindered by a largely underpowered research literature stemming from such a dysfunctional incentive structure (see Bakker, van Dijk, & Wicherts, 2012).

We argue that it is critical to draw a distinction between what is good for scientific progress versus what is good for an individual researcher's curriculum vitae, though we realize this may be a hard pill to swallow. Specifically, the current incentive structure is not conducive to generating cumulative knowledge, which, after all, is (should be) *the* goal of our collective scientific efforts (Ioannidis, 2005, 2012; Nosek, Spies, & Motyl, 2012). Indeed, there is growing evidence that the social psychology literature has serious replicability problems, thwarting cumulative knowledge and theoretical progress. For example,

- In the recently published Reproducibility Project (Open Science Collaboration, 2015), 76% (22/29) of results of studies in social psychology were not replicated.
- In a special issue on replication in the journal *Social Psychology* (Nosek & Lakens, 2014), 70% (16/23) of originally published results were not corroborated.
- In the Many Labs 3 project (Ebersole et al., in press), 88% (7/8) of results of studies in social psychology were not replicated, and this project included multisite replication attempts across 20 different labs.

Furthermore, there are now over 100 documented unsuccessful high-powered independent direct replications of many highly influential, and thus highly cited, social psychological findings (see LeBel, 2015a, for a list of 111 such replications). Additionally, there is compelling evidence that a large proportion of published articles contain statistical and/or reporting errors (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, J. M. 2014), and retractions due to errors (representing the majority [60%] of retractions) have increased tenfold within the past decade (Van Noorden, 2011). Taken together, the current incentive structure has produced a large number of published findings that are not corroborated in direct replication studies; yet, replication is a basic requirement for cumulative knowledge and theoretical progress.

Though it may seem on an intuitive level that the costs of doing open and high-powered research will lead to fewer true discoveries across the board, in actuality incurring these costs is required to increase the rate of making true discoveries (i.e., to be able to distinguish true from false hypotheses).[5] Specifically, McElreath and Smaldino (2015) showed in a comprehensive mathematical model of scientific discovery that high frequency and proportion of true discoveries require (a) a low Type I error rate, (b) high statistical power, and (c) the execution and high communication rate of direct replications. The direct implication of their results is that increasing the frequency and proportion of true discoveries in

[5] It is important to note that a "new discovery" should only be considered a true new discovery once follow-up independent direct replications have corroborated the reliability of a new finding. This is true for all discoveries, but, arguably, especially those that involve underpowered studies and/or nonconfirmatory analyses.

social psychology is only possible by (a) reducing our (inflated) Type I error rate, (b) increasing statistical power, and (c) executing *and* publishing independent direct replications (see also Popper, 1959; Feynman, 1974). This reasoning will be unpacked in Table 2, Figure 1, and Table 3.

The crucial idea to consider in Table 2 is that to gauge the true rate of scientific discovery, the *base rate* of true hypotheses needs to be considered in addition to Type I and II error rates. That is, from the pool of all hypotheses tested by psychologists, what proportion of these hypotheses is actually true a priori? This is of course difficult to estimate; however, a review by Ioannidis (2005, 2012) of several scientific domains subjected to intense and systematic independent replications reveal that a conservative value for the base rate of true hypotheses is .10 (i.e., 10% of all hypotheses tested turn out to be true). This low base-rate value seems reasonable for psychology given the highly complex and dynamic nature of psychological phenomena (McElreath & Smaldino, 2015) and embryonic state of theory (Meehl, 1978, 1990). Such a low base-rate value also makes sense if we are talking about the base rate of true hypotheses construed to be "field-changing discoveries," that is, findings that open up innovative new doors to our understanding of social behavior. For comparison purposes, we also consider a higher base rate of true hypotheses (.25) for hypotheses that could be construed as more obviously correct a priori, such as with incremental findings. Hence, Table 2 presents the true discovery rates for research with different Type I error rates and power levels for base rates of true hypotheses of .10 ("field-changing") and .25 ("incremental").

When the base rate of true hypotheses is .10, the status quo research approach—involving high Type I and II error rates—yields a very low true discovery rate of only .074 (i.e., true positives divided by [True positives + False negatives], or .025/.34 = .074). However, reducing the Type I error rate from .35 to .05, as advocated by open science proponents ("Open Science 1" row in Table 2), increases the true discovery rate to .36. And concurrently increasing power from .25 to .80, also advocated by open science proponents ("Open Science 2" row in Table 2), further increases the true discovery rate to .64. An error equivalence approach that balances both types of errors at .05 yields a true discovery rate of .68.

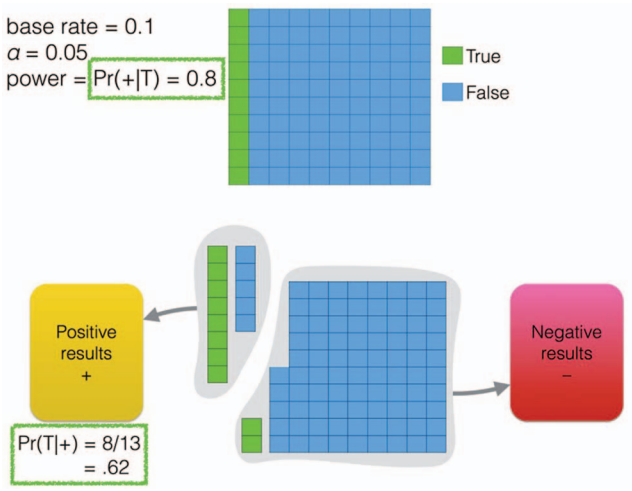

Figure 1. The rate of making true discoveries depends critically on the base rate of true hypotheses that are tested (relative frequency of green [or darker shade] to blue [or lighter shade] squares = .10), in addition to Type I (α = .05) and II (β = .20) error rates. In depicted example, true discovery rate is .62 with values for Type I and II error rates advocated by open science proponents (diagram reproduced with permission from McElreath & Smaldino, 2015). See the online article for the color version of this figure.

The logic underlying these calculations can be further unpacked by considering Figure 1, which visually depicts the "Open Science 2" approach of Table 2 (base rate = .10). The 10 green (or darker shade) blocks at the top half of Figure 1 represent the .10 base rate of true hypotheses whereas the 90 blue (or lighter shade) blocks represent the false hypotheses. A power level of .80 means eight of the 10 true hypotheses will be detected. A Type I error rate of .05 means five out of the 90 false hypotheses (technically 4.5 given .05 × 90 = .045) will incorrectly appear as positive results (left-hand side, bottom half of figure).

Consequently, such a research approach yields a true discovery rate of .62 (technically .64; small deviation due to rounding the 4.5 false hypotheses to 5 blue [or ligther shade] blocks) given that only eight out of 13 positive results actually reflect true hypotheses. Returning to Table 2, the same general conclusions hold when considering the higher base rate of true hypotheses of .25. Specifically, the status quo approach yields an unacceptably low true discovery rate of only .19, whereas a low Type I error rate and high-powered approach advocated by open science proponents yields a true discovery rate of .84.

It is important to consider, however, that the only way to determine which of the supported hypotheses are true (vs. false) is for researchers to conduct high-powered replication studies for each of the studies yielding positive results. Consequently, to fully evaluate Finkel et al.'s (2015 and others') metascientific concern that heightened stringency in terms of requiring larger sample sizes and executing replication studies may engender opportunity costs that will decrease the rate of making new scientific discoveries, we need to specifically compare the rate and frequency of making true discoveries relative to the collective resources imposed on the field in terms of follow-up replication studies required to distinguish true from false hypotheses.

Consequently, Table 3 describes the total resources required relative to the number of true discoveries unearthed for different research approaches (last column). To derive the estimates presented in the table, we made the following assumptions: The base rate of true hypotheses = .10, Type I error rate = .05, number of follow-up replication studies required to verify the robustness of each positive result (i.e., to distinguish a true from a false discovery) = 2, power of detecting an effect size of d = .41 in the replication studies = .95, and the subject pool resources available to researchers (N) = 5,000. The status quo (low power) approach allows a researcher to execute a much higher number of studies compared to the high power approach (i.e., 125 vs. 26) for a fixed amount of resources (e.g., subject pool of N = 5,000, chosen for ease of exposition). Among those studies, the low power approach yields a much higher number of studies with positive results (i.e., 9) compared to the high power approach, which only yields 3 (i.e., [True positives + False positives] × Number of studies, or [.08 + .05] × 26 = 3). Among those positive results, the low power approach yields 3 true discoveries (given true discovery rate of .36,

Table 3
Total Resources Required Relative to Number of True Discoveries Unearthed for Different Research Approaches

Research approach	Power	*N*	Proportion of studies yielding true positives	Proportion of studies yielding false positives	True discovery rate	No. of studies	No. of studies yielding positive results	No. of true discoveries	No. of replications required	Total *N* of replications	Total *N* per true discovery
Low power	.25	40	.03	.05	.36	125	9	3	18	5,443	1,742
High power	.80	190	.08	.05	.64	26	3	2	7	2,046	972
Error equivalence	.95	311	.10	.05	.68	16	2	2	5	1,400	917

Note. Calculations for all research approaches assume the following: base rate of true hypotheses = .10, Type I error rate = .05, number of follow-up replication studies per positive studies = 2, power of replication studies = .95, subject pool resources of individual researcher (*N*) = 5,000. Power reflects probability of detecting social psychology's average effect size of $d = .41$ (Richard et al., 2003) using typical sample sizes for between-subjects design (Simonsohn, 2014b). Using different values for these assumptions leads to highly similar conclusions. Indeed, we created a Shiny App using R code (available at http://shinyapps.org/apps/N-per-discovery/) to allow readers to easily change such values and explore the cost-effectiveness of different research approaches.

i.e., $.36 \times 9 = 3$) compared to the high power approach, which yields only 2 (i.e., $.64 \times 3 = 2$). However, this advantage for the low power approach reflects false progress because to distinguish the three true discoveries from the nine positive results, many more follow-up replication studies need to be executed by other researchers compared to the high power approach. Specifically, the low power approach requires 18 replications to distinguish the three true discoveries from the nine positive results (i.e., 2 replications per positive result [minimum], or $2 \times 9 = 18$), whereas the high-power approach requires only 7 replications to distinguish the 2 true discoveries from the 3 positive results (i.e., $2 \times 3.3 = 7$). Assuming .95 power for the replication studies, a much larger number of total subjects is required overall for the low-power approach (i.e., $N = 5{,}443$) compared to the high-power approach, which requires only a total of $N = 2{,}046$ subjects (i.e., $6.58 \times 311 = 2{,}046$). Considered this way, the high-power approach actually reflects a much more effective use of resources given that much fewer total resources are required per true discovery (i.e., $N = 972$) compared to the status quo approach (i.e., $N = 1{,}742$). Hence, even though at first blush it may appear to an individual researcher with limited resources that calls for increasing statistical power can stifle scientific progress, in actuality larger sample sizes and the execution of replication studies is required for overall scientific progress of the collective field. These observations reinforce our position (and the open science movement's general position) that we critically need to report our research more transparently (to help bring down the Type I error rate), run high-powered studies (to help bring down the Type II error rate), *and* execute and publish high-powered replication studies.[6]

Taken together, our analysis points to the dire need to change our field's incentive structure whereby better alignment exists between what is good for an individual researcher's career advancement and what is good for a field's cumulative progress (Ioannidis, 2012; Nosek et al., 2012). But how can we change the incentive structure to improve such alignment? A full discussion of this issue is beyond the scope of this article, but we offer here preliminary suggestions in the hopes of further promoting these important discussions. One approach is for us to all collectively be guided by the goal of doing good science rather than the personal goal of advancing our own careers by running many underpowered studies (admittedly, the current incentive structure rewards a focus on personal career goals). In other words, if we all start publishing fewer papers with fewer—but higher quality—studies (Nelson, Simmons, & Simonsohn, 2012; Schimmack & Dinolfo, 2013), what is considered "productive" will change and both individual researchers and the collective field will benefit over time in terms of generating reliable and valid explanations of human behavior.

There are strong headwinds limiting the likelihood of this proposal proving effective, however, given it involves a classic tragedy of the commons (Everett & Earp, 2015). Specifically, unless the entire field buys in to putting scientific progress ahead of individual gain, it remains advantageous for individual researchers to maintain the status quo to the detriment of the collective field (see also Gervais, Jewell, Najle, & Ng, 2015). But by individually adopting open science practices and promoting such practices publicly, (trailblazing) researchers can slowly nudge the incentive structure such that it better aligns with what benefits scientific progress (Ferguson, 2015). Indeed, Buttliere (2014) has argued that the best way to realign individual researcher and collective interests is to create a centralized web platform that facilitates and rewards postpublication peer review and open data (see http://CurateScience.org for one such centralized platform). Also, LeBel (2015b) has proposed a new replication norm for psychology whereby as a service to the field (paralleling the extant peer-review norm), researchers aim to (directly) replicate an important finding in their own area of research for every four original studies they publish. Finally, Hartshorne and Schachner (2012) have proposed a new system to evaluate the quality of researchers by estimating the replicability of their findings as a way to incentivize the publication of reliable results, which would go a long way to improving the current academic incentive structure.

Pragmatic Concerns

Although concerns regarding the costs of open science practices have focused primarily on the false-negatives versus false-positives trade-off, Finkel et al. (2015) also laid out a series of

[6] Our personal open science position advocates a *sufficiently open science*, which is science that is sufficiently open to allow for (a) accurate peer-review evaluation, (b) independent verification of analytic reproducibility of results, and (c) the execution of diagnostic direct replications. We are also of the position that rewarding open science practices is more effective in improving scientific practices than punishing nonopen science practices (Buttliere, 2014).

additional pragmatic concerns against overzealous "procrustean" applications of open science recommendations. To facilitate our discussion, we begin by directly quoting an excerpt from Finkel et al. that succinctly conveys their general pragmatic concern:

> To address why we believe it is so important for the evidentiary value movement to account for variation across subfields, let us consider a scenario in which (a) psychology develops strict new norms and rules but (b) variation in research questions and optimal methodology across subfields means that Subfield 1 and Subfield 2 are differentially able to adhere to those norms and rules. Relative to the research questions and methods of Subfield 1, the research questions and methods Subfield 2 are inherently less amenable to the conduct of close replications, to strict preregistration, to the efficient sharing of research materials, to data sharing, and so forth [for example, increasing sample sizes]. As we look forward 10 or 20 years, it seems likely that Subfield 1 will gain status over time while Subfield 2 will lose it, with straightforward consequences for representation in top journals, allocation of grant resources, and implications for hiring and promotion decisions. (p. 292)

The core pragmatic concern reflected in this scenario rests on the assumption that the entire field of psychology might develop and implement "strict new norms and rules" that are insensitive to the research realities of the various subfields within the discipline (e.g., behavioral neuroscience, clinical, cognitive, comparative, developmental, educational, evolutionary, industrial/organizational, personality, social psychology, etc.). However, there is no evidence at this time that this hypothetical scenario is occurring in reality. In fact, there is no single organization in existence that governs or "rules" *all* of these subfields.[7] More generally, the scientific method inescapably involves calibrating one's beliefs according to the quality of evidence, which needs to be independently verified and replicated before scientists, practitioners, policymakers, or the general public place much confidence in a set of findings (see Ferguson, 2015). Though it is true that achieving such confidence may be substantially more difficult to achieve in certain domains of inquiry relative to others (e.g., relationship science, neuroscience, and developmental psychology relative to experimental psychology), in the scientific arena this is simply unavoidable. Furthermore, it is unclear whether open science practices are in fact less amenable to relationship scientists given that the majority of relationship science studies involve data from individuals (vs. couples) and is similar to experimental studies (rather than dyadic or longitudinal studies; see Kashy, Campbell, & Harris, 2006).

As an example of their hypothetical scenario, Finkel et al. (2015) call attention to "badges" that are now being attached to articles in some journals (e.g., *Psychological Science*) if they adhered to specific open science practices, and they suggest some subfields' (e.g., relationship science) research methodologies may be less amenable to receiving those badges than others (e.g., experimental psychology). It is important to point out that open science badges are meant as small incentives (in the short term) to help nudge researchers toward scientific behaviors that benefit any research area. Such badges are not (and were never) meant to allow researchers to deem findings as "scientifically valid only if they have been honored with at least one (or maybe all) research integrity badges" (Finkel et al., 2015, p. 292). Indeed, as just mentioned, researchers should only have confidence in findings that have been independently verified and corroborated via high-powered direct replications, which is the only way to eventually separate true from false findings (Mcelreath & Smaldino, 2015; Feynman, 1974). Furthermore, relationship researchers should avoid comparing the gradual embracement of more open science practices across areas of research, but rather they should strive to compare progress in open science practices across time within relationship science (i.e., we are *now* more transparent and use larger sample sizes than previously). In the long run, the ultimate path to gaining status as a scientific field is to develop accurate theories that withstand grave refutation in the face of true empirical discoveries. Again, this path is only possible by reducing the Type I error rate, increasing statistical power (and therefore reducing the Type II error rate), and executing and publishing independent direct replications (Mcelreath & Smaldino, 2015, see Figure 1). We now examine issues related to the open science movement's recommendations with respect to preregistration, open materials, open data, close replications, and sample size.

Preregistration

A pragmatic concern regarding preregistration is that for some studies (e.g., longitudinal, large-scale surveys), once a study has begun or data collection completed, it seems impossible on the face of it to preregister any new hypotheses. In fact, however, Finkel et al. (2015) correctly note that one indeed *can* preregister new hypotheses after data collection has started for longitudinal studies or for any preexisting data sets, as long as the hypotheses and data analytic plan are registered prior to conducting the newly proposed analyses with the existing data. Indeed, researchers who preregister new hypotheses or analyses for preexisting data sets qualify to earn *Psychological Science's* preregistration badge (with a special "DE" indicating that data exist) as long as any deviations from the preregistered plan are disclosed.

More generally, however, Finkel et al. (2015) mention that psychological science will benefit from careful consideration and discussion of the optimal use of preregistered versus non preregistered studies, suggesting that ". . . [preregistration] may be nonsensical when data collection involves intensive and/or longitudinal methods" (p. 292). But we contend that a compelling case can be made that preregistration should be seen as *particularly valuable* for expensive and/or time-consuming studies. This is the case because, for small-scale cross-sectional studies (whether online or in the lab), it is relatively simple to run additional confirmatory studies after conducting entirely exploratory studies or after encountering unexpected results. On the other hand, conducting a confirmatory study to replicate the effects of a large-scale longitudinal study of married couples, for example, is generally not feasible. Therefore, knowing more details of how the study was conducted is of utmost importance to accurately evaluate the veracity of the presented findings. Indeed, this is precisely the position of Frank (2013, a developmental psychologist). For costly longitudinal studies, the cost of preregistration is minimal relative to the costs of conducting the longitudinal study, and the benefits

[7] That being said, overarching scientific organizations (e.g., the Association for Psychological Science) may ultimately come to value the extent to which individual subfields (e.g., relationship science) do the best they can with respect to open science goals, rather than holding individual subfields to the standards of areas where such goals may be easier to attain.

of preregistration greatly outweigh the costs of executing a confirmatory direct replication of a finding from a longitudinal study.

Open Data

Making data openly available has been suggested as one way to facilitate the verification of, and confidence in, research findings. This suggestion seems particularly relevant given recent research showing that only 38% of authors that had published articles in four American Psychological Association journals in 2012 shared their data when requested by other competent professionals (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015), even though the American Psychological Association code of conduct states sharing data with competent professionals is required postpublication. Finkel et al. (2015) correctly point out, however, that ensuring confidentiality of partner dyadic data represents a unique challenge for relationship scientists who wish to publicly post their de-identified and anonymized data to a public data repository. Couple-level data sets often include information that can be used to identify partner pairs within the data even when those data lack a couple ID number. For example, inclusion of wedding dates or locations, specific open-ended responses, partner name and/or initials, and other items routinely collected by researchers could be used to identify pairs of couples within a data set, particularly when an individual has intimate knowledge about one or more couple members represented in a couple-level data set. As a result, it is theoretically possible that a vengeful (ex-)partner, or simply a curious partner, might access a publicly posted data set and probe into his or her partner's responses. As a result, Finkel et al. state that "it might be practically impossible to share such data publicly" (p. 286).

Though we acknowledge that such challenges are real, we believe there are several approaches available to researchers to overcome these challenges. Before reviewing such approaches, it is important to clarify that publicly sharing data (or making data "publicly available"), does not necessarily mean making *all raw data* publicly available. Rather, the spirit of the 'open data' aspect of open science is that researchers publicly post only the portions of the raw data set that are required to reproduce the main results reported in a published article (indeed PLOS' new policy regarding mandatory public availability of data upon publication of an article refers to such data sets as "minimal datasets," Silva, 2014). Hence, any personally identifiable information or open-ended responses can (and should) be removed before posting the data to a public repository, for all data sets, including those without a dyadic partner structure (see Mackinnon, 2014, who summarizes key identifiers to consider removing when de-identifying data; see also Chadwick, 2015). Returning to the vengeful ex-partner scenario, a simple approach to practically eliminate the possibility that a subject identifies their partner's responses by remembering their own item-level responses is to only publicly share composite scores rather than item-level scores. Alternatively, or in addition, another approach is to upload a data set to the Open Science Framework, but only make those data available to other researchers (something that is not too difficult to confirm) with the agreement that those data will not be publicly posted. Another option would be to encrypt a data set and make the encryption key available after confirming the identity of an individual requesting access to the data set; alternatively, professional societies or publishers could distribute such encryption keys as a succession plan to overcome situations whereby individual researchers do not respond or pass away.

The common theme here is that there are a number of ways that researchers can safely and ethically achieve open data without giving carte blanche access to the entire data set (see Burnham, 2014, for details). Moreover, many of the recommended solutions to the confidentiality concern also minimize fears of what would happen if a particular online depository is breached, such as what happened with the now infamous 2008 Facebook data set (Zimmer, 2010; see Finkel et al., 2015, for brief overview). Encrypting data sets and/or posting data sets, but keeping them private until the identity and intentions of interested parties are confirmed, is a reasonable work-around when researchers are not confident they have appropriately scrubbed their data sets of possible variables that could be used to identify participants. We refer readers to Fraser and Willison (2009) as well as Friedlin and McDonald (2008) who have developed tools to efficiently de-identify potentially identifying information from sensitive data sets (see also Cavoukian & Emam, 2011). Overall, our central argument on this issue is that there are plenty of examples in the literature, from fields that deal with arguably more sensitive data than do relationship scientists, regarding how to share relevant data with the research community in a manner that protects subject confidentiality.

Another pragmatic issue raised with regard to open data is the concern that researchers' ideas may be "scooped" by other researchers who may unfairly benefit from publishing articles based on open data. Our response to such a concern is twofold. First, publicly posting one's data and materials (whether for a published article or prior to publication) can actually safeguard oneself from being scooped because there is a public (and time/date stamped) record that you are testing (or have tested) particular hypotheses with corresponding evidence. Second, we are unaware of any documented cases where a researcher has published a paper based on open data without inviting the data set creator as a coauthor and/or simply properly crediting the data source (the current social norm). Of course it is *possible* that such inappropriate data scooping could occur; however, it is important to mention that new metrics are currently being developed to give researchers who make their data/materials publicly open the appropriate credit they deserve (e.g., ImpactStory.org, Altmetric.org).[8] Also, Figshare.com and the OpenScienceFramework.org both track the number of downloads and views of one's open data/materials and even provide a permanent DOI to cite data directly so that researchers can get appropriate credit for open data, even if the data are unpublished.

An additional issue that has been raised in public discussions concerning open sharing of data is that journals may somehow "own" the rights to a researcher's data if those data are submitted to the journal as part of open science practices. Such concerns likely arise from cursory readings of many journal publishing agreements. For example, the *Psychological Science* publishing agreement indicates the following:

[8] Furthermore, we hope social psychologists can avoid the open data prisoner's dilemma and rationally cooperate for the collective benefit of all.

> In the event Contributors provide Supplemental Materials, as defined in Section 2 of the Terms of the Agreement, Contributors hereby grant to Association the nonexclusive right and license to reproduce, publish, republish, create derivative works, distribute, sell, license, transfer, transmit, and publicly display copies of, and otherwise use, the Supplemental Materials, in whole or in part, alone or in compilations, in all formats and media and by any method, device, or process, and through any channels, now known or later conceived or developed, for the full legal term of copyright and any renewals thereof, throughout the world in all languages and in all formats, and through any medium of communication now known or later conceived or developed, and the nonexclusive right to license or otherwise authorize others to do all of the foregoing, and the right to assign and transfer the rights granted hereunder.

Such legal language is certainly intimidating, and coupled with the fact that "Supplemental Materials may include, but is not limited to, data sets . . .," we appreciate the concern. We should point out, however, that contributor agreements do not transfer rights to the journals for data sets. In fact, as we were informed by a representative from Sage Publications (and consistent with information discussed on the Center for Open Science listserv; F. Schönbrodt, e-mail communication, April 30, 2015), "no one can own data in a copyright sense." That said, we understand that the mere thought of a journal having legal right to redistribute a data set shared as part of a publication is less than ideal. Thus, we suggest a very simple fix: Upload data to the Open Science Framework and provide a link to the project when provision of data for publication is desired and/or required. In such a case, data are not submitted as supplementary material, so the concern is completely mitigated.

Open Materials

Finkel et al. (2015) are generally open to the public sharing of experimental materials and procedures, but state that ". . . intellectual property issues emerging in the wake of the evidentiary value movement are extremely complicated, and it is likely that addressing them successfully will require collaborations among, at minimum, psychologists, ethicists, and legal scholars" (p. 293). We respectfully disagree with this position given that a fundamental aspect of the scientific method requires that independent researchers have access to all relevant details of the research process to allow for accurate evaluations and interpretation of empirical findings. If an original researcher is unwilling or unable to share such details, then it is impossible for an independent researcher to gauge the veracity and strength of the reported evidence. For example, consider the situation where a private for-profit company (e.g., E-harmony) publishes a set of peer-reviewed findings reporting that they have the most successful matching system for single individuals seeking partners. Such a set of findings should of course not be outright ignored; however, researchers (and other private interests) should not place much confidence in those findings until sufficient methodological and procedural details have been shared (as suggested by Finkel, Eastwick, Karney, Reis, & Sprecher, 2012, in the case of companies that use matching algorithms to pair clients on dates but do not disclose the details of those algorithms). That being said, we agree that the field needs to think more deeply about these intellectual property issues, but that simply continuing to not share such methodological details in any form is an undesirable option at this point in time.

Additionally, all consumers of science should be given the right to draw their own conclusions about what a specific set of published findings really mean. The only way to fully evaluate scientific conclusions is to fully understand the methods that underlie those conclusions. Of particular concern are those situations in which a researcher chooses to not disclose or otherwise provide detailed information about exactly what study participants experienced because she or he does not think some information is relevant to a specific set of analysis. For example, imagine a scenario in which a researcher tests the association between variables A (e.g., attachment style) and C (e.g., relationship satisfaction), but chooses to not provide details about variable B (e.g., conflict behavior), which was collected in between the two target variables. In such a case, knowing about B may very well affect interpretation of the association between A and C. Furthermore, anything that happened before the assessment of the target variables should also be reported.

Such disclosures, however, shouldn't be limited to preceding or 'sandwiched' variables. In some cases, information collected after the target variables may also be relevant for appropriate interpretation of study results. For example, using the scenario depicted above, If D was collected after A and C, and D is an interchangeable outcome variable with C, then such information should also be disclosed as analysis of D provides important information on the robustness of the A–C association (e.g., gauging robustness of association between attachment style [A] and relationship satisfaction [C] would require disclosure of relationship quality [D]). Perhaps the only situation in which a variable or study detail does not need to be disclosed, for the sake of interpretation, is when there are additional clearly unrelated outcome variables.[9] In such cases, knowledge of these variables does not necessarily affect interpretation of reported findings. However, in the interest of future reanalyses and reinterpretations from different theoretical perspectives, we contend scientific progress can be substantially accelerated when such information is also provided.

Close Replications

We strongly agree with Finkel et al. (2015) recommendation that for longitudinal studies relationship scientists can "devote some components of their intensive and/or longitudinal studies to close replications of one or more published findings" (pp. 287–288), and then augment this with components that test new hypotheses. However, we want to clarify important issues regarding Finkel et al.'s position with respect to "what close replications and conceptual replications can and cannot achieve" (p. 288). Finkel et al. state that conceptual replication will typically have further-reaching implications for testing the theoretical propositions under study, whereas direct replications render conclusions susceptible to idiosyncracies of the original stimuli and/or methods. This position is undeniably true, given that a finding that is only replicable using specific stimuli and/or measures at a particular point in time is unlikely to be important theoretically or in terms of real-world

[9] Note that after results are known, confirmation bias could influence a researcher's perception regarding the relatedness of an outcome variable, especially when much time has passed between formulating hypotheses and analyzing data. Preregistration effectively eliminates such unintentional confirmation bias problems.

applications. That being said, we believe that their stance *overemphasizes* the execution of conceptual replications relative to direct replications, which can be harmful for a field. If every positive finding is immediately followed up with a conceptual replication using a different manipulation and/or outcome measure, then such positive findings can never be disconfirmed because "failed" conceptual replications can always be attributed to the intentional changes in methodology rather than the falseness of the original hypothesis (LeBel & Peters, 2011; Pashler & Harris, 2012).

Increasing Sample Sizes

Finkel et al. (2015) state that

> . . . even in the nearly bullet-proof case that our science requires larger sample sizes, it is necessary to add the caveat that procrustean applications of stricter sample size policies may sometimes be ill-advised, such as in cases where participant recruitment is particularly difficult or expensive. (p. 292)

Though it is understandable that situations where participant recruitment is difficult and/or expensive can be personally frustrating to researchers, it is simply an unavoidable fact of inferential statistics that one needs sufficient statistical power to detect effect sizes that can reasonably be expected to truly exist in certain areas of research. Without sufficient statistical power, it is a waste of time and precious resources—and arguably unethical to participants (Rosenthal, 1994)—to execute such underpowered studies, no matter how difficult and/or expensive participant recruitment is. If a researcher runs into such situations, one potential solution is to combine resources with other labs interested in testing the same hypothesis (e.g., Grahe et al., 2012), as has been done in genome-wide association studies (Manolio, 2009). Indeed, the examples we discussed previously make it clear that increasing sample sizes, while potentially costly for individual researchers, is crucial for the field if we wish to make important and replicable discoveries. Indeed, as Tables 2 and 3 illustrate, increasing sample size is a more cost-effective solution for true scientific progress than maintaining the status quo.

Finally, in relation to sample size considerations, we want to address Finkel et al.'s (2015) power calculations that led them to conclude the field of relationship science is on "relatively solid ground" (p. 291). In their Table 2 (p. 290), Finkel et al. report effect sizes that can be detected with .80 power for different effect size types with different forms of nonindependence. Although .80 power is typically considered ideal in psychological research, it corresponds to a Type II error rate of .20, or 4 times the nominal alpha level of .05 (the Type I error rate). Given that Finkel et al.'s main epistemological concern is that heightened stringency regarding false-positive rates will increase false-negative rates, it seems plausible that they would have considered arguing for increasing power beyond .80 as one certain way to decrease the Type II error rate. Additionally, Finkel et al.'s conclusion regarding power in relationship science studies is based on the assumption that researchers primarily assess simple correlations. However, many studies actually test more complex hypotheses involving interactions, which require much larger sample sizes to achieve the same level of statistical power (Simonsohn, 2014a). Furthermore, when relationship researchers collect data from both partners they often assess partner effects (i.e., the association between Partner A's independent variable and Partner B's dependent variable), effects that tend to be smaller than actor effects (Kenny, Kashy, & Cook, 2006) and therefore require larger sample sizes to reliably detect (see also Bakker, Hartgerink, & Wicherts, 2015 who find the vast majority of researchers overestimate power afforded by specific research designs). Indeed, an analysis by Schimmack (2015) revealed that articles in two journals that publish research focusing only on close relationships (i.e., *Personal Relationships* and *Journal of Social & Personal Relationships*) in the years 2010 to 2014 had average post hoc power of only .52 and .56, respectively (with a negative trend detected for the *Journal of Social & Personal Relationships* such that 2015 articles yielded power of only .39). These power levels, however, clearly overestimate actual power given they are based on post hoc power levels of only the studies reported, which does not account for the file drawer problem (John et al., 2012; Rosenthal, 1979). Furthermore, in another metascientific analysis, Fraley and Vazire (2014) found that studies in the *Journal of Personality and Social Psychology*'s *Interpersonal Relations and Group Processes* (between 2006 and 2010) had power of only .49 to detect an $r = .20$, the average effect size from Richard et al.'s (2003) metameta-analysis of social psychological effects (which importantly *also* does *not* account for publication bias). Taken together, based on more nuanced considerations of the different designs used in relationship science and metascientific analyses estimating power for relationship science journals, we conclude that power levels in relationship science are insufficiently high and need to be much improved.

Summary

In this article, we scrutinized concerns recently raised by several researchers regarding the possible unacknowledged (or underappreciated) costs of adopting open science practices, which such researchers argued may outweigh the benefits and hence stifle the discovery of novel findings. Specifically, we questioned these researchers' epistemological concerns that the heightened stringency of open science practices involves undesirable trade-offs in terms of (a) statistical concerns (i.e., reducing false-positives will by definition increase false-negatives) and (b) metascientific concerns (i.e., larger samples and replications engender opportunity costs that will decrease the rate and frequency of true discoveries). Regarding their statistical concern, we argued that the concern that heightened stringency regarding false-positives will by definition increase false-negatives is unwarranted given that open science proponents (a) do not propose reducing the nominal α level (from .05 to .01 for e.g.) but rather recommend such practices to bring down the inflated Type I error rate from $\sim$.35 (due to design and analytic flexibility; Simmons et al., 2011) back down to .05 and (b) simultaneously have called for high-powered research to reduce Type II error rates (which is currently $\sim$.75, Bakker et al., 2012). Thus, open science practices reduce both Type I *and* Type II error rates. Regarding their metascientific concern, we demonstrate that though it may seem on an intuitive level that the costs of doing open and high-powered research will lead to fewer true discoveries across the board, in actuality incurring these costs is required to increase the rate of making true discoveries. This is the case because increasing the frequency and proportion of true discoveries (i.e., to be able to distinguish true from false hypotheses) is only possible by (a) reducing Type I error rate, (b) increasing

statistical power, and (c) executing *and* publishing independent direct replications (Feynman, 1974; McElreath & Smaldino, 2015; Popper, 1959). Overall then, our analysis of the relative costs versus benefits of open science practices for both individual researchers and the field overall, leads us to conclude that the benefits outweigh the costs for both individual researchers and the collective progress of our field in the long run, but that in the short term this may not be the case for individual researchers due to the (still) broken academic incentive structure. An important implication of our analysis points to the dire need of changing our field's incentive structure whereby better alignment exists between what is good for an individual researcher's career advancement and what is good for a field's cumulative progress (Ioannidis, 2012; Nosek et al., 2012). In this spirit, we mention several recently proposed ideas on how to begin such incentive structure realignment.

As a secondary objective, we scrutinized pragmatic concerns raised regarding the adoption of open science practices with respect to preregistration, open materials, open data, close replications, and sample size recommendations. In each of these cases, we show that the concerns are overstated given available solutions that already exist to mitigate such concerns.

We end by commending these researchers for their contributions that have helped sharpen and advance the discussion of the merits of open science practices, as evidenced by the existence of our article. In this sense, we agree with Finkel et al. (2015) that "research practices in our field will be better—in terms of scientific discovery and validity—in 2020 than they were in 2010" (p. 294). Ultimately, these discussions will improve research practices in psychology and consequently accelerate theoretical progress in our quest to understand human behavior.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. http://dx.doi.org/10.1002/per.1919

Bakker, M., Hartgerink, C. H. J., & Wicherts, J. M. (2015, October 12). Power intuitions. Retrieved from https://osf.io/5t0b7/

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. http://dx.doi.org/10.1177/1745691612459060

Borsboom, D. (2014, November 12). Facilitating radical change in publication standards. Retrieved from http://osc.centerforopenscience.org/2014/11/12/facilitating-radical-change/

Burnham, B. (2014, February 4). Open data and IRBs. Retrieved from http://centerforopenscience.github.io/osc/2014/02/05/open-data-and-IRBs/

Buttliere, B. T. (2014). Using science and psychology to improve the dissemination and evaluation of scientific work. *Frontiers in Computational Neuroscience, 8,* 82. http://dx.doi.org/10.3389/fncom.2014.00082

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Campbell, L., Loving, T. J., & LeBel, E. P. (2014). Enhancing transparency of the research process to increase accuracy of findings: A guide for relationship researchers. *Personal Relationships, 21,* 531–545. http://dx.doi.org/10.1111/pere.12053

Cavoukian, A., & El Emam, K. (2011). *Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy*. Information and Privacy Commissioner of Ontario, Toronto: Ontario, Canada. Retrieved from https://www.ipc.on.ca/images/Resources/anonymization.pdf

Chadwick, I. (2015, May 21). Can I really share that? Working with sensitive and confidential data. Retrieved from http://www.open.ac.uk/blogs/the_orb/?p=458

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49,* 609–610. http://dx.doi.org/10.1016/j.cortex.2012.12.016

Chambers, C. (2014, January 24). The changing face of psychology. Retrieved from http://www.theguardian.com/science/head-quarters/2014/jan/24/the-changing-face-of-psychology

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1,* 140216. http://dx.doi.org/10.1098/rsos.140216

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

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6,* 1152. http://dx.doi.org/10.3389/fpsyg.2015.01152

Ferguson, C. J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist, 70,* 527–542. http://dx.doi.org/10.1037/a0039405

Feynman, R. P. (1974). Cargo cult science. *Engineering & Science, 37,* 10–13.

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 & Personality Science, 7,* 45–52. http://dx.doi.org/10.1177/1948550615612150

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest, 13,* 3–66. http://dx.doi.org/10.1177/1529100612436522

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. http://dx.doi.org/10.1037/pspi0000007

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9,* e109019. http://dx.doi.org/10.1371/journal.pone.0109019

Franco, A., Malhtra, N., & Simonovits, G. (2015). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological & Personality Science*. Advance online publication.

Frank, M. (2013, July 25). Thoughts on preregistration. Retrieved from http://babieslearninglanguage.blogspot.ca/2013/07/thoughts-on-preregistration.html

Fraser, R., & Willison, D. (2009). Tools for de-identification of personal health information. *Pan Canadian Health Information Privacy (HIP) Group*.

Friedlin, F. J., & McDonald, C. J. (2008). A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association, 15,* 601–610.

Gervais, W. M., Jewell, J. A., Najle, M. B., & Ng, B. K. L. (2015). A powerful nudge? Presenting calculable consequences of underpowered research shifts incentives toward adequately powered designs. *Social*

Psychological & Personality Science, 6, 847–854. http://dx.doi.org/10.1177/1948550615584199

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science, 7,* 605–607.

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6,* 8. http://dx.doi.org/10.3389/fncom.2012.00008

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. http://dx.doi.org/10.1177/1745691612464056

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Kashy, D. A., Campbell, L., & Harris, D. W. (2006). Advances in data analytic approaches for relationships research: The broad utility of hierarchical linear modeling. In A. Vangelisti & D. Perlman (Eds.), *The Cambridge handbook of personal relationships* (pp. 73–90). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511606632.006

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9,* 278–292. http://dx.doi.org/10.1177/1745691614528520

LeBel, E. P. (2015a, October 13). A list of successful and unsuccessful high-powered direct replications of social psychology findings. Retrieved from https://proveyourselfwrong.wordpress.com/2015/10/13/a-list-of-successful-and-unsuccessful-high-powered-direct-replications-of-social-psychology-findings/

LeBel, E. P. (2015b). A new replication norm for psychology. *Collabra Open Access Journal, 1,* Art. 4. http://dx.doi.org/10.1525/collabra.23

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.org: Grassroot support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8,* 424–432. http://dx.doi.org/10.1177/1745691613491437

LeBel, E. P., & Campbell, L. (2013). Heightened sensitivity to temperature cues in individuals with high anxious attachment: Real or elusive phenomenon? *Psychological Science, 24,* 2128–2130. http://dx.doi.org/10.1177/0956797613486983

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15,* 371–379. http://dx.doi.org/10.1037/a0025172

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience, 4,* 423–428. http://dx.doi.org/10.1093/scan/nsp052

Lucas, R. E., & Donnellan, B. (2013). Stop me if you think you have heard this before: The challenges of implementing methodological reforms. *European Journal of Personality, 27,* 130–131.

Mackinnon, S. (2014, January 29). Privacy in the age of open data. Retrieved from http://osc.centerforopenscience.org/2014/01/29/privacy-and-open-data/

Manolio, T. A. (2009). Collaborative genome-wide association studies of diverse diseases: Programs of the NHGRI's office of population genomics. *Pharmacogenomics, 10,* 235–241. http://dx.doi.org/10.2217/14622416.10.2.235

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112,* 331–348. http://dx.doi.org/10.2466/03.11.PMS.112.2.331-348

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144,* e1–e15. http://dx.doi.org/10.1037/xge0000038

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods, 9,* 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563. http://dx.doi.org/10.1146/annurev.psych.59.103006.093735

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS ONE, 10,* e0136088. http://dx.doi.org/10.1371/journal.pone.0136088

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834. http://dx.doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244. http://dx.doi.org/10.2466/pr0.1990.66.1.195

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18,* 107–118. http://dx.doi.org/10.1177/1088868313496330

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry, 23,* 291–293. http://dx.doi.org/10.1080/1047840X.2012.705245

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2,* 175–200. http://dx.doi.org/10.1037/1089-2680.2.2.175

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23,* 217–243. http://dx.doi.org/10.1080/1047840X.2012.692215

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology, 45,* 137–141. http://dx.doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631. http://dx.doi.org/10.1177/1745691612459058

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. Advance online publication. http://dx.doi.org/10.3758/s13428-015-0664-2

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2014). The chrysalis effect how ugly initial results metamorphosize into beautiful articles. *Journal of Management*. Advance online publication. http://dx.doi.org/10.1177/0149206314527133

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. http://dx.doi.org/10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. http://dx.doi.org/10.1177/1745691612463401

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9,* 319–332. http://dx.doi.org/10.1177/1745691614528519

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1,* e175. http://dx.doi.org/10.7717/peerj.175

Platt, J. R. (1964). Strong inference. *Science, 146,* 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. Oxford, UK: Basic Books.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–648. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5,* 127–134. http://dx.doi.org/10.1111/j.1467-9280.1994.tb00646.x

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17,* 551–566. http://dx.doi.org/10.1037/a0029487

Schimmack, U. (2015, September 28). Replicability ranking of 27 psychology journals. Retrieved from https://replicationindex.wordpress.com/2015/09/28/replicability-ranking-of-27-psychology-journals-2015/

Schimmack, U., & Dinolfo, G. (2013). Increasing replicability requires reallocating research resources. *European Journal of Personality, 27,* 132–133.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Nosek, B. A. (2016). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. Retrieved from https://osf.io/j5v8f/

Silva, L. (2014, February 24). PLOS' new data policy: Public access to data. Retrieved from http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

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. http://dx.doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN 2160588*. http://dx.doi.org/10.2139/ssrn.2160588

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76–80. http://dx.doi.org/10.1177/1745691613514755

Simonsohn, U. (2013). Small telescopes: Detectability and the evaluation of replication results. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2259879

Simonsohn, U. (2014a, March 12). No-way interactions. Retrieved from http://datacolada.org/2014/03/12/17-no-way-interactions-2/

Simonsohn, U. (2014b, May 1). We cannot afford to study effect size in the lab. Retrieved from http://datacolada.org/2014/05/01/20-we-cannot-afford-to-study-effect-size-in-the-lab

Srivastava, S. (2012, September 27). A Pottery Barn rule for scientific journals. Retrieved from https://hardsci.wordpress.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/

Van Noorden, R. (2011). Science publishing: The trouble with retractions. *Nature News, 478,* 26–28. Retrieved from http://www.nature.com/news/2011/111005/full/478026a.html

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra Open Access Journal, 1,* 3.

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS one, 9*(12), e114876. http://dx.doi.org/10.1371/journal.pone.0114876

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632–638. http://dx.doi.org/10.1177/1745691612463078

Wicherts, J. (2015, January 2). Will 2015 be the year in which most Psychological Science articles have open data? Tweet retrieved from https://twitter.com/jeltewicherts/status/551039830650785792

Wigboldus, D. H., & Dotsch, R. (2015). Encourage playing with data and discourage questionable reporting practices. *Psychometrika*. Advance online publication. http://dx.doi.org/10.1007/s11336-015-9445-1

Wilson, T. (2014, June 15). Is there a crisis of false negatives in psychology? Retrieved from https://timwilsonredirect.wordpress.com/2014/06/15/is-there-a-crisis-of-false-negatives-in-psychology/

Zehetleitner, M., & Schönbrodt, F. (2016, January 14). When does a significant p-value indicate a true effect? Understanding the positive predictive value (PPV) of a p-value. Retrieved from http://shinyapps.org/showapp.php?app=http://87.106.45.173:3838/felix/PPV

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology, 12,* 313–325. http://dx.doi.org/10.1007/s10676-010-9227-5

Received June 9, 2015
Revision received January 27, 2016
Accepted January 29, 2016 ■