Running Head: BEST PRACTICES IN PSYCHOLOGY

Best Research Practices in Psychology: Illustrating Epistemological and Pragmatic Considerations with the Case of Relationship Science

Eli J. Finkel

Northwestern University

Paul W. Eastwick

University of Texas at Austin

Harry T. Reis

University of Rochester

November 6, 2014

(This Draft Replaces the November 4th draft that circulated previously)

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

Abstract

In recent years, a robust movement has emerged within psychology to increase the evidentiary value of our science. This movement, which has analogs throughout the empirical sciences, is broad and diverse, but its primary emphasis has been on the reduction of statistical false positives. The present article addresses epistemological and pragmatic issues that we, as a field, must consider as we seek to maximize the scientific value of this movement. Regarding epistemology, this article contrasts the *false-positives-reduction* (FPR) approach with an alternative, the *error balance* (EB) approach, which argues that any serious consideration of optimal scientific practice must contend simultaneously with both false-positive and false-negative errors. Regarding pragmatics, the movement has devoted a great deal of attention to issues that frequently arise in laboratory experiments and one-shot survey studies, but it has devoted less attention to issues that frequently arise in intensive and/or longitudinal studies. We illustrate these epistemological and pragmatic considerations with the case of relationship science, one of the many research domains that frequently employ intensive and/or longitudinal methods. Specifically, we examine six research prescriptions that can help to reduce false-positive rates—preregistration, prepublication sharing of materials, postpublication sharing of data, close replication, avoiding piecemeal publication, and increasing sample size. For each, we offer concrete guidance not only regarding how researchers can improve their research practices and balance the risk of false-positive and false-negative errors, but also how the movement can capitalize upon insights from research practices within relationship science to make the movement stronger and more inclusive.

Abstract word count: 250

Best Research Practices in Psychology: Illustrating Epistemological and Pragmatic Considerations with the Case of Relationship Science

According to Greek mythology, Poseidon's son Procrustes regularly offered weary travelers hospitality for the night. Once inside, he made them repose on an iron bed, which was in fact a torture device. If the traveler was too short or too tall for the bed, Procrustes stretched or hacked him until his body was precisely the same length as the bed. In doing so, Procrustes killed the traveler and claimed his money and possessions.

In contemporary usage, the adjective *procrustean* refers to an entity requiring that everything fit a preconceived standard. In the domain of scientific conduct, procrustean research practices could refer to a process through which scholars first decide what effect they wish to see (the procrustean standard) and then manipulate the data to produce that effect. Indeed, many psychological scientists manipulate their data ways that artificially increase the likelihood that they will find evidence to support an effect that the scientists want them to support (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). Increasing recognition of this fact has helped to launch a robust social movement, the *evidentiary value movement*, to increase the extent to which the evidence that scholars accumulate and disseminate provides information that helps the field converge on truth over time. To date, the dominant perspective within this movement has been the *false-positives-reduction* (FPR) approach; indeed, the level of dominance has made it easy to conflate the FPR approach with the broader movement rather than recognizing that the reduction of false positives is simply one means—albeit a crucial one—for increasing the evidentiary value of our science. Specifically, the FPR approach seeks to bolster the quality of our science by reducing the prevalence of false positives. In doing so, it seeks to increase confidence that statistically significant effects in the scientific literature provide valid evidence for true effects in the broader population.

The evidentiary value movement has commanded enormous attention in many corners—in scholarly journals, at conferences, in task forces, in granting agencies, in the mainstream and social media, and so forth. However, this attention has not been equally distributed across subfields within psychology. After all, the movement has frequently offered recommendations—pertaining to topics such as the number of participants required per cell of a research design or secretly jettisoning data from certain conditions—that are much more relevant to some research domains than to others. We suggest that this unequal distribution of attention is producing two adverse consequences. First, scholars in those subfields that are not the primary focus of the movement are paying less attention to developments in that movement and, consequently, are at risk for missing out on an important opportunity to improve their research practices. Second, given that much of the emphasis in the movement has revolved around prototypical research methods—laboratory experiments and one-shot surveys—the leaders of the movement, in conjunction with the scientific policymakers they influence (journal editors, leaders of scientific societies, granting agencies, etc.), are at risk for neglecting sensible variation in research practices across subfields and, consequently, creating norms and policies that inadvertently marginalize those subfields that employ research methods that deviate from the prototype.

These adverse consequences have procrustean parallels of their own. First, those scholars who are neglecting developments in the movement may be in for a sudden stretching or hacking when they next seek to publish their findings. Second, just as it can harm science for a scholar to concoct a standard (e.g., a mean difference between conditions) and then stretch or hack the data to fit that standard, it can harm science to develop norms or standards that are optimal for a select group of research domains and then stretch or hack other research domains to fit that standard.

We illustrate these considerations, which pervade the empirical sciences, with a discussion of relationship science. As do scientists in many other disciplines and in many other subfields within

psychology, relationship scientists employ a broad range of research methods. They sometimes employ laboratory procedures or one-shot surveys, but they frequently employ resource-intensive and/or longitudinal procedures on nonindependent units, especially married or dating couples. The movement speaks directly to the former types of methods, but, thus far, it has spoken much less directly to the latter types of methods.

Article Overview

We pursue two primary goals in this article. First, we provide a broad epistemological framework for considering the important issues raised by the evidentiary value movement. In particular, we couch the discussion of these issues in terms of the overarching principles of discovery and validity, which are the central benchmarks against which we should assess efforts to improve scientific practice. In doing so, we contrast the FPR approach with the *error balance* (EB) approach, which emphasizes (a) that efforts to reduce false-positive rates will frequently exacerbate false-negative rates and (b) that the circumstances under which one of these types of errors is more damaging than the other vary in complex ways across research contexts. Second, we seek to move beyond procrustean, one-size-fits-all solutions to the false-positives problem in favor of a more nuanced discussion of how scholars can implement the insights emerging from the evidentiary value movement while simultaneously attending to the diverse pragmatic issues they confront as they seek to contribute valid findings to the scholarly literature.

It is difficult, even ill-advised, to consider these broad epistemological and pragmatic issues exclusively in the abstract. As such, we complement our discussion of these issues with a detailed illustration of how they play out in everyday practice within one particular research domain: relationship science. Although we would be delighted if relationship scientists (and other scholars who employ intensive and/or longitudinal methods) find this detailed case study especially useful, our intended audience is psychological scientists more generally. We intend for our general

approach in this case study—a systematic consideration of the epistemological and pragmatic issues relevant to optimal scientific conduct—to generalize across research domains.

In pursuit of our primary goals, we first situate the evidentiary value movement within a broader historical context, contrast the FPR and EB approaches to increasing the evidentiary value of our science, and introduce the sorts of pragmatic considerations that can emerge as scholars consider the implications of the evidentiary value movement for their own research practices. Next, we provide concrete discussions of these epistemological and pragmatic issues vis-à-vis six research recommendations emerging from the evidentiary value movement, especially the FPR approach: preregistration, prepublication sharing of materials, postpublication sharing of data, close replication, avoiding piecemeal publication, and increasing sample size. Finally, we provide a broad discussion of how psychology can address the issues raised by the evidentiary value movement in a manner that sets our discipline on the strongest course toward scientific excellence.

The Evidentiary Value Movement in Historical Context

The bedrock of the evidentiary value movement consists of two observations (e.g., Bakker, van Dijk, Wicherts, 2012; Fanelli, 2012; Humphreys, de la Sierra, & Van der Windt, 2013; Ioannidis, 2005, 2008; John et al., 2012; Kerr, 1998; Nosek, Spies, & Motyl, 2012; Pashler & Harris, 2012; Simmons et al., 2011; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). The first observation is that researchers have strong incentives to find statistically significant results. After all, virtually all meaningful professional rewards—getting hired or promoted, procuring grant funding, garnering the respect of one's peers, and so forth—have long depended upon publishing empirical articles, and journals have long favored articles reporting statistically significant findings over articles reporting statistically nonsignificant findings, regardless of the quality of the research methods or the importance of the research question. The second observation is that researchers have data-

analytic latitude, or "researcher degrees of freedom," for pushing *p*-values below .05, a process known as *p-hacking* (Simmons et al., 2011, p. 1359). Aside from simply not reporting studies that failed to yield statistically significant results, scholars can jettison conditions or outliers, report only those dependent variables that reach statistical significance, snoop on their findings and terminate data collection as soon as the desired effect becomes significant, tinker with the inclusion versus exclusion of covariates, and so forth. When the incentives to find statistically significant results are combined with *p*-hacking, the true rate of false positives substantially exceeds the nominal α-level (say, .05), which logically implies that the published literature contains many more false-positive findings than are indicated by that α-level (more than 5%).

The roots of the movement date back to the middle of the 20th century (e.g., de Groot, 1969/2014; Sterling, 1959), and, indeed, at least one scholarly journal—*Representative Research in Social Psychology* (*RRSP*)—was long dedicated to the publication of replications and null effects. In 1970, a handful of graduate students at the University of North Carolina at Chapel Hill (UNC), including Robert B. Cialdini, procured funding from the American Psychological Association and several bake sales to launch *RRSP*, which was (with occasional gaps) published annually well into the twenty-first century (Chamberlin, 2000).[1] Perhaps because the field was not yet ready to prioritize the journal's emphasis on replications and null effects or because of the editorial turnover associated with a journal run entirely by graduate students at a single university, *RRSP* never became a high-impact journal, and it went defunct around a decade ago.

Although *RRSP* may have remained on the field's periphery, concerns about false-positive error inflation have roiled the field in the past. In particular, the *Journal of Personality and Social Psychology* (*JPSP*) prioritized these issues during Anthony G. Greenwald's editorial tenure in the

[1] During his graduate student days at UNC, the first author of the present article served as an associate editor of volume 25 of *RRSP* (2001).

1970s. In the editorial introducing his policies, Greenwald (1976) observed that "There may be a crisis in personality and social psychology, associated with the difficulty often experienced by researchers in attempting to replicate published work" (p. 2) and cautioned against "selective presentation of results that are favorable to the author's preferred hypothesis" (p. 6). Greenwald's editorial policies were far stricter than those of previous (and subsequent) *JPSP* editors, and the number of articles published in the journal plummeted on his watch. These changes drew strong objections and ultimately resulted in the premature termination of his editorial tenure.

It turns out that scholars like Cialdini and Greenwald were ahead of their time. Their ideas presaged those in the evidentiary value movement, which did not begin to coalesce until the new millennium and did not become a major force until quite recently. Many crucial catalysts transpired outside of psychology, including John P. A. Ioannidis' (2005) article in *PLOS Medicine* entitled "Why Most Published Research Findings Are False" and Jonah Lehrer's (2010) article in *The New Yorker* entitled "The Truth Wears Off: Is There Something Wrong with the Scientific Method?" The movement first became a major force in psychology in 2011, and it rapidly commanded the attention of many of the most influential individuals and organizations. Various developments converged during that year. For example, Daryl J. Bem (2011) published an article in *JPSP* presenting evidence for precognition, triggering an immediate backlash among scholars decrying research practices that can yield false positives (e.g., Galak, LeBoeuf, Nelson, & Simmons, 2012). Jelte M. Wicherts, Marjan Bakker, and Dylan Molenaar (2011) discovered not only that many researchers frequently are unwilling to share the raw data underlying their published reports, but that this unwillingness is linked both to weaker evidence for the statistically significant results and to an increased prevalence of apparent errors in the reporting of those

results. Most importantly, Joseph P. Simmons and colleagues (2011) published their hugely influential "False-Positive Psychology" article on *p*-hacking.[2]

The movement gained steam after 2011. *Perspectives on Psychological Science* (*PPS*) has published a series of special sections on the topic (Volume 7, Issue 6; Volume 8, Issue 4; Volume 9, Issue 3; Volume 9, Issue 6), and it introduced an entirely new type of journal article called registered replication reports, in which many independent laboratories follow "an identical, vetted protocol designed to reproduce the original method and finding as closely as possible" (Simons, Holcombe, & Spellman, 2014, p. 552). The Open Science Collaboration (2012), spearheaded by Brian A. Nosek, initiated a massive undertaking: the replication of a large swath of studies published in high-profile psychology journals in 2008. Nobel laureate Daniel Kahneman (2012) emailed a widely circulated open letter to influential social priming researchers observing that "questions have been raised about the robustness of priming results," stating that "I see a train wreck looming," and recommending that the scholars "should collectively do something about this mess." Scholars developed procedures that allow observers to use the published literature to detect evidence for elevated false-positive error rates. For example, Ulrich Schimmack (2012) introduced the "incredibility index" to quantify the probability that a multiple-study article has reported all relevant studies and data analyses rather than surreptitiously burying at least one additional effect that did not support the researchers' hypothesis (also see Francis, 2012; Sterling et al. 1995), and Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons (2014) introduced the concept of "*p*-curves" to allow scholars to compare the distribution of statistically significant *p*-values for a set

[2] Another major development was the discovery of systematic fraud in the publications of several prominent psychologists, notably Diederik Stapel in social psychology and (apparently) Marc Hauser in cognitive psychology. Although the evidentiary value movement is not primarily oriented toward addressing outright fraud, the sort of close scrutiny afforded by the changes advocated by the movement can help investigators ferret it out (Simonsohn, 2013).

of studies to the theoretical distribution of *p*-values that should emerge if the relevant effect actually exists in the population.

Changes in research norms and policies followed. The Society for Personality and Social Psychology (SPSP) Task Force on Publication and Research Practices offered a number of recommendations, including that scholars (a) report information on their sample size decisions, (b) avoid research practices that increase the likelihood of false positives, (c) make research materials available for independent replication, and (d) pursue high-quality direct replication studies (Funder et al., 2014). High profile journals—including *JPSP*, *PPS*, *Psychological Science*, *Personality and Social Psychology Bulletin* (*PSPB*)*, Journal of Experimental Social Psychology* (*JESP*), and *PLOS One*—have adopted new publication policies. For example, *Psychological Science* now offers electronic badges for (a) preregistering studies and analysis plans, (b) making research materials publicly available, and (c) publicly sharing data files. Major funding agencies, including the U.S. National Institutes of Health (NIH), have altered their policies and funding priorities toward the goal of reducing false-positive rates (e.g., Collins & Tabak, 2014). In short, the breadth and depth of the evidentiary value movement's influence since 2011 have been staggering.

Epistemological considerations:

The False-Positives-Reduction Approach and the Error Balance Approach

In considering how to leverage the insights of the evidentiary value movement in the most constructive manner possible, we must revisit the basic logic underlying hypothesis testing. This imperative is especially strong given the dominance of the FPR approach. A false positive, also called an "α-error" or a "Type I error," is one of four possible outcomes from a hypothesis test. Figure 1 incorporates principles from signal detection theory (Tanner & Stets, 1954) to illustrate the logic underlying null hypothesis statistical testing (also see Fiedler, Kutzner, & Krueger, 2012). A *false positive*, which is depicted in the upper-right quadrant of Figure 1, emerges when

the researcher incorrectly concludes from a study that an effect exists in the population. A *false negative*, which is depicted in the lower-left quadrant (and which is also called a "β-error" or a "Type II error"), emerges when the researcher incorrectly concludes from a study that an effect does not exist in the population. A *true positive* or a *true negative* emerges when the researcher draws correct (i.e., accurate, valid) conclusions about the presence or absence, respectively, of an effect in the population.[3]

Psychological scientists typically set α (the theoretical possibility of a false positive) at .05, and, following Cohen (1988), they frequently set β (the theoretical possibility of a false negative) at .20. In other words, the field has, in principle, been willing to accept false positives 5% of the time and false negatives 20% of the time (although *de facto* false-positive and false-negative rates almost certainly have been higher than these nominal α and β levels). These rates derive from convention rather from some sort of platonic ideal, and, indeed, there are many circumstances under which scholars might prefer α or β levels that are stricter or looser than is conventional.

As widely discussed in our research methods and statistics textbooks, the setting of α and β levels involves tradeoffs. For example, according to Keppel and Wickens (2004),

> Increasing the chance of a Type I error decreases the chance of a Type II error. Every researcher must strike a balance between the two types of error, and decreasing the chance of a Type I error increases the chance of a Type II error. When it is important to discover new facts, we may be willing to accept more Type I errors and thus enlarge the rejection region by increasing α. But when we want to avoid Type I errors—for example, not to get started on false leads or to clog up the literature with false findings—we may be willing to accept more Type II errors and decrease the rejection region. (p. 48)

Tradeoffs between false positives and false negatives extend well beyond the setting of α and β levels. Indeed, pending how reviewer and editor norms change, such tradeoffs are relevant to

[3] Some scholars have argued that hypothesis testing of this sort is fundamentally flawed and should be eliminated altogether. That important topic is beyond the scope of the present article, which is in large part a response to the central emphasis on hypothesis testing among scholars adopting an FPR approach to maximizing the quality of our science.

many, perhaps most, recommendations emerging from the evidentiary value movement. For example, one major emphasis is a call for researchers to be more open about their research practices, even if (or perhaps especially when) the practices a researcher is tempted to hide are inconsistent with the hypothesis she seeks to support. If reviewers and editors become more tolerant of imperfect or messy data (e.g., marginally significant main effects, simple effects that fall short of statistical significance in a subset of the studies), as recommended by some scholars (Maner, 2014; Simmons et al., 2011), then greater openness will yield more information for readers to digest, but it will not necessarily alter the ratio of false positive to false negative errors. In contrast, if reviewers and editors retain their longstanding standards for how robustly, or how cleanly, the data support the researcher's hypothesis, we are likely to see a marked decline in journal acceptance rates. Such a decline would almost certainly reduce false positives rates by, for example, making it harder for her to publish findings that required the deletion of outliers to obtain a statistically significant finding. But it would simultaneously increase false negatives rates by eliminating from the scientific literature cases in which outlier-deletion reveals a scientific truth.

Klaus Fiedler and colleagues (2012) use the term *theoretical false negatives* to refer to instances in which potentially true effects are overlooked or omitted from the scholarly literature. 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. In other cases, stricter publication policies will replace a false positive with a true negative, clearly a good 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.

Given that the central goal of the FPR approach is the reduction of false-positive error rates, and given the dominance of the FPR approach within the evidentiary value movement, it is not surprising that many of the recommendations emerging from this movement are systematically oriented toward the reduction of such error rates. In the present article, we contrast the FPR approach with an alternative that we call the *error balance* (EB) approach. As shown in Table 1, the EB approach has three central tenets. The first tenet is that both false positives and false negatives undermine the superordinate goals of science, which are discovery and validity, with *validity* defined as "the best available approximation to the truth or falsity of propositions" (Cook & Campbell, 1979, p. 37). Of particular relevance to the present discussion, Thomas D. Cook and Donald T. Campbell (1979) define "statistical conclusion validity" as the truth or falsity of inferences regarding the correspondence between (a) the conclusion drawn from a sample-based hypothesis test and (b) the truth regarding that conclusion in the population from which the sample was selected.

The second tenet of the EB approach is that neither type of error—neither false positives nor false negatives—is uniformly a greater threat to validity than the other type. To be sure, certain scholars have staked out strong claims that one type of error is, by and large, worse than the other. For example, Simmons and colleagues (2011, p. 1359) have argued that false positives are the more costly of the two types of error because, among other problems, such errors (a) tend to be difficult to weed out once they have been published, (b) place the field at risk for losing credibility, and (c) waste resources by inspiring "investment in fruitless research programs and can lead to ineffective policy changes." In contrast, Fiedler and colleagues (2012, pp. 666-667) have argued that false negatives are the more costly type of error because, among other problems, (a) overcoming false negatives is much more likely to yield theoretical innovations than is abandoning false positives; (b) research strategies that seek to overcome false negatives "can yield

existence proofs and new discoveries," whereas research strategies that seek to reduce false positives "only yield ambiguous nonproofs"; and (c) both statistical and theoretical false positives are more prevalent than equivalent false negatives (due to small sample sizes in the field, normative lack of correction for sampling and measurement error, and researchers' inattentional blindness to alternative hypotheses). According to the EB approach, both of these sets of arguments are compelling, and there is no reason why either type of error should uniformly be given more weight than the other. Rather, the circumstances under which either type of error is worse than the other will involve complex considerations regarding whether it is more important, in a given research context, to avoid drawing an incorrect conclusion about a given effect (in which case false-positive errors are worse) or to discover new truths (in which case false-negative errors are worse). For example, whether it is worse for scholars to develop and publicize an intervention to reduce domestic violence that in fact lacks efficacy (a false positive) or for scholars to fail to identify a beneficial intervention that is efficacious in reducing domestic violence (a false negative) involves the consideration of injury rates, resource use, and many other factors.

The third tenet of the EB approach is that any serious consideration of optimal scientific practice must contend with both types of error simultaneously. In particular, all conceptual analyses that compare the relative scientific value of certain research practices must explicitly address the implications of those practices for both types of error, not only in terms of the putative accuracy of one's findings, but also in terms of the practical implications of each type of error. Now that policymakers have gleaned major new insights from the evidentiary value movement—especially from scholars adopting an FPR approach—they face a dilemma: Should they move as quickly as possible to develop new rules and norms oriented toward the reduction of false positives, or should they move more deliberately to allow for a robust discussion regarding (a) how such rules and norms are likely to alter false-negative rates and (b) the costs and benefits

resulting from such tradeoffs between false-positive and false-negative rates? Although we are persuaded that false-positive rates are almost certainly higher than most members of our discipline had appreciated circa 2010, we believe that the possibility that the field will overreact to these new insights may be just as threatening to the quality of our discipline—as evaluated in terms of discovery and validity—than the possibility that the field will underreact (also see Sbarra, 2014).

Pragmatic considerations: The Problem with Procrustes

As noted previously, one major priority in the evidentiary value movement is to formalize and implement new norms and policies regarding scientific conduct. For example, the "False-Positive Psychology" article included a list of six requirements for authors, including "authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification" and "authors must decide the rule for terminating data collection before data collection begins and report this rule in the article" (Simmons et al., 2011, p. 1362). In general, we admire efforts along these lines, but it is important to recognize that both the relevance and the consequences of these rules vary considerably across research domains. For example, the "20 observations per cell" requirement is largely irrelevant to research domains in which experimental participants are rare (e.g., people with a certain type of brain damage), and the "data termination" requirement is tricky to implement in research domains in which participant recruitment is especially unpredictable. This issue of variation across research domains has been complicated further as scholars have recommended additional rules beyond Simmons et al.'s (2011), as some of these rules are much more extreme. For example, Wagenmakers et al. (2012, p. 632) recommended that researchers who wish to test hypotheses must "preregister their studies and indicate in advance the analyses they intend to conduct."

We offer two observations in response to this one-size-fits-all thread that has been woven into the evidentiary value movement. The first is that because research questions vary considerably

across research domains, so too do optimal research methodologies, and these methodologies should be evaluated according to how well they help to answer the relevant research question (Cronbach, 1957). For example, the best test of the hypothesis that a threat to one's moral purity induces the need to cleanse oneself could be a laboratory experiment in which participants are randomly assigned to receive such a threat or not (Zhong & Liljenquest, 2006). In contrast, the best test of the hypothesis that the infant-mother relationship is linked to conflict management skills in infants' romantic relationships 20 years later could be a decades-long longitudinal study involving observer ratings of infants' behavior (Simpson, Collins, Tran, & Haydon, 2007). These methods are sufficiently distinct that regulations designed to encourage researchers to conduct the optimal laboratory experiment may be nonsensical, cumbersome, or even counterproductive when applied to intensive and/or longitudinal studies, and vice versa. For example, as we elaborate later in this article, procrustean preregistration requirements may be reasonable for one-shot surveys or easy-to-conduct laboratory experiments, but not for many intensive and/or longitudinal studies (e.g., the field would have to disavow all hypothesis tests conducted on existing data from the General Social Survey, a massive, decades-long study that cannot be preregistered at this point).

The second observation is that the structure and nature of scientific communication—the dissemination of research findings in the scientific literature—must remain sufficiently flexible to encourage scholars across research domains to pursue those methods that are best suited to answering the relevant research question. The field has long shown flexibility along these lines, but such flexibility could be imperiled by overzealous implementation of some of the norms and policies proposed within the evidentiary value movement. Indeed, the pace with which the major professional societies, journals, and funding agencies have implemented policy changes is a potential concern not only because the decision-makers seem not to have accounted sufficiently

for false-negative rates, but also because these changes may marginalizing certain research domains, especially those in which optimal methods deviate from the field's prototype.

In short, as the field works to harness the evidentiary value movement for the betterment of our science, policymakers must be sensitive to variation in optimal methodology across research domains. There is peril in one-size-fits-all solutions. Of course, although advocating for the broad principle of flexibility is easy, discerning how to implement this principle is not. The devil is, as always, in the details.

A Case Study: Relationship Science

We take a first step toward such implementation by illustrating how these epistemological and pragmatic considerations play out in relationship science. We provide this illustration not because relationship science is a special case, but rather because the illustration exemplifies the sort of practical, actionable analysis that will help scholars capitalize upon the evidentiary value movement to develop optimal research practices. It can provide an initial template for how scholars can pursue such investigations in other research domains. Indeed, focused analyses of practices in diverse research domains will be required before the field can make well-informed decisions about which new rules and norms to implement in light of the FPR and the EB approaches to increasing the evidentiary value of our science.

Research Practices in Relationship Science

Relationship science is an empirical method of understanding human relationships, with a strong focus on romantic relationships (Berscheid, 1999). Because behavioral science research methods evolved largely for the study of individuals, relationship scientists have had to develop methods better suited to the study of dyads and their development over time (Reis, 2012; Reis, Collins, & Berscheid, 2000). These methods are frequently resource-intensive. Many studies involve the recruitment of couples who meet stringent inclusion criteria (e.g., couples who have

been married for less than six months), laborious coding of the interaction between the two partners, diary or experience-sampling assessments of the partners' thoughts and feelings, multi-year prospective analysis of relationship dynamics, and so forth. A prototypical study might examine newlywed couples every six months for four years, incorporate diverse laboratory procedures and an experience-sampling protocol, and encompass 40 hours of questionnaires.

It is this resource-intensiveness—in terms of money, researcher hours, lag from the beginning to the end of data collection, and so forth—that complicates the intersection of the evidentiary value movement, especially the FPR approach, with relationship science. Indeed, due to the resource-intensiveness of such studies, relationship scientists typically design them to test a relatively large number of distinct ideas. Consequently, such scholars measure a broad range of constructs, many of them at multiple points in time and with multiple methods (e.g., self-reports, partner-reports, and observational codings). This extensive measurement is crucial not only because it allows the researcher to test multiple ideas, thereby using resources sensibly, but also because it frequently enables them to address issues raised during the journal review process. Given the widespread tendency for editors and reviewers to raise alternative explanations and request additional data analysis, and given the impracticality of simply rerunning such resource-intensive studies, relationship scientists frequently assess a broad range of variables the first time around. In some cases, they do so even when they are not certain whether (or how) they will use a given variable in a future empirical article.

Because relationship science methods frequently deviate from the sorts of methods that are particularly emphasized within the evidentiary value movement (laboratory experiments and one-shot surveys) it provides a good illustration of how focused case studies of a specific research domain can inform the broader discussion regarding optimal research practices in the field at large. Indeed, the present case study has implications that reverberate throughout much of the

field, especially throughout those subfield—including large swaths of organizational, developmental, clinical, personality, and health psychology—that rely heavily on intensive and/or longitudinal research methods. More generally, the present case study is likely to reverberate well beyond psychology's borders throughout other disciplines that rely heavily on such methods, including large swaths of education, sociology, economics, political science, epidemiology, and public health.

The False-Positives-Reduction Approach, the Error Balance Approach, and Relationship Science

The present case study addresses six topics that reside at the intersection of the FPR approach, the EB approach, and relationship science. Five of these topics—preregistration, prepublication sharing of materials, and postpublication sharing of data, close replication, and increasing sample size—address central recommendations offered by the evidentiary value movement. The remaining topic—avoiding piecemeal publication—addresses an issue that has been largely neglected but that can substantially increase false-positive rates, especially in subfields that employ intensive and/or longitudinal methods. For all six topics, we discuss (a) what problems each recommendation seeks to solve, (b) how implementing the recommendation in relationship science raises new challenges and affords new opportunities, (c) the implications of the recommendation for best practices in relationship science, and (d) how insights from this case study can inform the discussion surrounding the FPR and EB approaches.

The first three of the six topics—preregistration, prepublication sharing of materials, and sharing data—fit under the umbrella term *open science*, the principle that optimal scientific conduct is facilitated by maximal transparency among scientists. When adhering to this principle, scholars disclose to the scientific community how many studies they ran relevant to a particular

research question, what the procedures and measures were for all of those studies, which statistical analyses they conducted, and so forth.

1. Open science 1—preregistration. In addressing the problems that preregistration seeks to solve, it is useful to distinguish among three distinct types of preregistration: preregistration of theoretical propositions one plans to investigate, preregistration of the studies one plans to conduct, and preregistration of the statistical analyses one plans to perform. Because the evidentiary value movement has focused almost exclusively on the preregistration of studies and data-analytic plans (rather than on the preregistration of theoretical propositions), our discussion starts with those two types of preregistration. We revisit the preregistration of theoretical propositions below.

Preregistration of studies, including the procedures to be employed and the measures to be assessed, helps to solve the pervasive *file-drawer problem*—a phenomenon in which scholars conduct a study but never publish the results, instead burying the findings in a literal or metaphorical file drawer (Rosenthal, 1979; Scargle, 2000). Studies that are file-drawered are not a random sample of the studies that scholars have conducted because of *publication bias*—the much greater likelihood of a finding being published if it is statistically significant, even controlling for the quality of the research design and execution (Fanelli, 2012; Ferguson & Heene, 2012; Francis, 2012; Kühberger, Fritz, & Scherndl, 2014). Publication bias, which results from both editorial decision-making and author decisions regarding which findings to submit for publication, systematically increases the proportion of false-positives in the published literature and distorts perceptions of the robustness and size of an effect (e.g., Ioannidis, 2005; Schooler, 2011). Preregistration helps to ensure that other scholars are aware that a researcher has conducted a particular study (or at least *intended* to conduct such a study), and those scholars can either track down the subsequently reported results or contact the researcher to request information about the

results. If journals agree to accept preregistered studies regardless of the outcome of the data analysis, or if the use of alternative outlets for publishing the results of these studies becomes mainstream, then the magnitude of the file-drawer problem will be reduced.

Preregistration of the specific statistical tests to be conducted—that is, publicly posting a "preanalysis plan" in advance of performing data analysis to "bind [one's] hands against data mining" (Casey, Glennerster, & Miguel, 2012, p. 1755; also see Humphreys et al., 2013; Wagenmakers et al., 2012)—helps to address the practice of "Hypothesizing After the Results are Known," or HARKing (Kerr, 1998). When HARKing, scholars deviate from logical empiricism's hypothetico-deductive approach, in which scientists propose a falsifiable hypothesis in advance of designing the study testing the hypothesis. Scholars adopt an exploratory approach to data analysis but then report their findings as if the statistically significant results from this exploratory process had been hypothesized (and derived from theory) in advance. Norbert L. Kerr (1998) has observed that, relative to traditional hypothetico-deductive methods, HARKing (a) produces a scientific literature littered with effects that are less likely to replicate, (b) risks elevating the status of a false-positive finding by formalizing it in theoretical terms, and (c) undermines Popper's (1959/2002) criterion of disconfirmability because it is impossible to disconfirm a hypothesis that was derived from already-known findings (see de Groot, 1969/2014; Wagenmakers et al., 2012). In other words, statistically significant findings are more likely to be true rather than false positives to the extent that they are confirmatory rather than HARKed (Kerr, 1998; Wagenmakers et al., 2012). Preregistration encourages researchers to increase the extent to which their research is confirmatory, which makes their hypothesis tests more convincing. In addition, if preregistration increases the tendency for researchers to pay more attention to, and to report, the extent to which their analyses are confirmatory versus exploratory, it will help consumers of scientific reports

make more-informed judgments about how persuasively a given hypothesis test supports a given conclusion.

The field has taken large strides in recent years to increase the practice of preregistration. For example, the Center for Open Science launched in 2013, and major journals within psychology have advanced initiatives to foster preregistration. *Psychological Science* has begun honoring some articles with a "Preregistered badge," which "is earned for having a preregistered design and analysis plan for the reported research and reporting results according to that plan" (Eich, 2014, p. 3). *PPS* now "fortifies the foundation of psychological science" by publishing a new type of article called "registered replication reports" that involve multi-lab replications that are preregistered and conducted precisely in accord with the preregistered plan (Association for Psychological Science, 2014; also see Simons et al., 2014). A recent special issue of the journal *Social Psychology* was devoted to preregistered replication attempts (Nosek & Lakens, 2014; for details, see Open Science Framework, 2014-b). *JESP* has devoted a special issue to three types of preregistered studies: (a) replications, (b) fully confirmatory studies, and (c) "exploratory / confirmatory" blends in which the authors followed up at least one study that was not preregistered with at least one that was (Brandt, Crawford, & Giner-Sorolla, 2014). In addition to these changes to established journals, the European Association of Social Psychology and the Society of Australasian Social Psychologists have launched a new journal, *Comprehensive Results in Social Psychology* (*CRSP*), for which the review process involves authors submitting the theoretical background, hypotheses, methods, and statistical analyses prior to data collection; as long as authors follow the approved protocol, their article will be published regardless of the results of the statistical analyses.

Preregistration and relationship science. Relationship scientists can easily preregister their studies when they conduct laboratory experiments and simple surveys, but preregistration becomes much more fraught with regard to intensive and/or longitudinal studies. As noted

previously, these two categories of methods differ markedly in terms of the time elapsed between the beginning and the end of data collection—perhaps a few months versus many years. During these extra years of data collection, the field will advance, and the scientists conducting the research will have new ideas. In many cases, such developments will yield new hypotheses that scientists can test with data from their ongoing or completed longitudinal studies. Such cases are common in relationship science, and preregistration of a data analysis plan is certainly possible in such cases. But preregistration *prior to the collection of data* is by definition impossible with existing datasets, and such a requirement would preclude relationships researchers from participating in new research endeavors, including eligibility of publishing in the new journal *CRSP*, that require the procedure itself to be peer-reviewed and approved before data collection. New policies should not discourage relationship scientists from developing a new hypothesis after the commencement of data collection—not from snooping around in their dataset, but from theory or from learning about new findings from other studies—and using the existing dataset to test it.

Standard data analytic practice with large, intensive datasets like those used in relationship science typically involves a blend of confirmatory and exploratory procedures, which frequently produces are more substantial scientific yield—in terms of discovery and validity—than procedures that are exclusively confirmatory. For example, imagine that researchers hypothesize *a priori*, and subsequently find evidence for, a theoretically sensible effect in one portion of their dataset (e.g., perceived similarity on intelligence predicts initial attraction). They will typically proceed to examine the boundaries of this effect by conducting additional convergent validity (e.g., perceived similarity for good career prospects also predicts initial attraction) and discriminant validity (e.g., *actual* similarity for intelligence does not predict initial attraction) tests. Additional data analytic approaches (e.g., computing similarity in a new way) could be used to probe the robustness of the phenomenon, and theoretically important moderators (e.g., length of

acquaintance) could be tested if they have been assessed. If the dataset contains diary, longitudinal, or behavioral components, the hypothesis could be reconceptualized for testing in these other segments of the dataset depending on the measures available and manuscript length restrictions. The researcher ultimately reports the analyses that tell the complete story of the data, with all theoretically important convergent, discriminant, and moderational tests. Frequently, the story that emerges (e.g., across traits, perceived but not actual similarity predicts attraction; see Tidwell, Eastwick, & Finkel, 2013) is different from the one that the researcher hypothesized (e.g., similarity in intelligence predicts attraction), even if the original hypothesis test was supported. Indeed, in such cases, the failure to engage in exploratory supplementary analyses beyond the initially supported confirmatory test would lead to an incomplete, even inaccurate, understanding of the topic under investigation.

This perspective on data exploration is a far cry from "data torturing" perspective characterizing a certain strain within the evidentiary value movement (e.g., Wagenmakers et al., 2012, Figure 1), which frequently suggests or implies that deviations from a strict preregistered plan are reliably oriented toward *p*-hacking a nonsignificant finding into a statistically significant one. There is no question that some data exploration involves *p*-hacking, but much of it involves a scientifically pure attempt to understand the complete story the data are telling vis-à-vis an *a priori* hypothesis. In such cases, the exploratory analyses are downstream from the crucial hypothesis tests, and the researcher is conducting them to discern the contours of the already-supported effect. She is not particularly invested whether the results are statistically significant; she is simply listening to her data.

Intensive datasets are rich and messy, and we do not know how the blend of confirmatory and exploratory practices illustrated in our similarity–attraction example balances the potential for false-positive and false-negative errors. We do know that reviewers have long exhibited wariness

of false positives in such datasets: It is uncommon for a paper from such a large dataset to be published if it consists of a single hypothesis test without additional analyses that demonstrate the boundaries and robustness of the phenomenon. Exploratory elements in relationship science—that is, elements that are less than strictly confirmatory—will remain essential insofar as they balance *a priori* theoretical, methodological, and statistical components with nondevious deviations to accommodate unanticipated nuances in the data or the conceptual framework (e.g., lack of variation on a given scale item). After all, the most appropriate analyses frequently follow successful or unsuccessful confirmatory tests, and these exploratory elements are vital for developing the most scientifically valuable understanding of the empirical evidence.

Nevertheless, playing fast and loose with researcher degrees of freedom allows scholars to construct an artificially strong case for a phenomenon by selectively omitting variables and analyses that run counter to their preferred conclusions. Preregistration should help to clarify exactly which hypothesis tests were conceived prior to the examination of the data. In general, researchers should disclose which analyses deviated from the confirmatory plan (and *how* they deviated from the plan), and exploratory procedures will continue to inform judgments about the extent to which hypotheses were supported versus refuted for scholars working with large datasets. Scholars must not confuse this latter type of exploration with *p*-hacking.

Preregistration—implications for best practices in relationship science. Widespread adoption of preregistration practices would yield a major advance in addressing the problems of file-drawering and *p*-hacking. Insofar as preregistration increasingly becomes a standard component of scientific practice in psychology, and insofar as the preregistration process becomes increasingly efficient, we encourage relationship scientists to make a good-faith effort to adhere to this practice, even vis-à-vis intensive and/or longitudinal studies. Specifically, as the field increasingly achieves these two qualities vis-à-vis preregistration (widespread adoption and efficiency), we recommend

that relationship scientists incorporate two new procedures into their standard research practice. First, for studies that are either concluded or ongoing, we encourage relationship scientists to preregister the precise theoretical propositions and data analyses they intend to perform for any ideas they intend to test with the data. For example, if they have a new idea they wish to test in an extant dataset, they would preregister their precise hypothesis and analysis plan before beginning data analysis. Second, for studies for which data collection has not commenced, we encourage relationship scientists to preregister the major theoretical propositions they intend to prioritize with the data from the study, along with a general summary of the procedures, measures, and data analyses they intend to implement. In many cases, scholars can get a long way down this road by simply cutting-and-pasting text from a grant proposal or an institutional review board (IRB) submission.[4] In most studies, scholars will end up deviating from the preregistered plan, of course—jettisoning items that reduced the reliability of a new scale in the early waves of data collection; conducting additional discriminant validity, convergent validity, or moderational tests; and so forth—but that is not a reason to avoid preregistering the main ideas and materials. When reporting the results, the authors can simply note the ways in which the final procedures deviated from the preregistered plan. They can also employ a sequential procedure in which they preregister each new idea and data-analytic procedure before conducting the relevant analyses.

Preregistration—implications for psychological science in general. We recommend that policymakers exhibit flexibility to accommodate the particular issues that emerge regarding the preregistration of intensive and/or longitudinal studies. Relationship scientists have already amassed, or are in the process of amassing, a large corpus of studies that can shed light on relationship dynamics over time, and it is impossible to preregister these study procedures

[4] The relevance of IRB proposals might decline sharply in the coming years, pending the extent to which the recent recommendations from the National Research Council influence policy (S. T. Fiske, 2014). If those recommendations are widely adopted, IRB proposals are likely to become relatively rare in psychology.

retroactively. These datasets were costly to collect and represent a treasure trove of potential insights regarding relationship dynamics. In addition, that the datasets already exist does not mean that the research is exploratory or that the reported results were HARKed. Even for new studies that begin after preregistration practices have become prevalent, it is perhaps ill-advised to devalue the testing of theoretical ideas that occur to researchers after they have begun collecting data.

We recommend that policymakers develop a system that allows relationship scientists to participate in preregistration as much as possible—by preregistering the data analysis plan, for example—to ensure that such scholars have just as strong an incentive to pursue preregistration as scholars in other subfields (e.g., eligibility for research integrity badges). Along these lines, we were delighted to learn that *Psychological Science*, in collaboration with the Open Science Framework (2013), views such procedures as eligible for the Preregistered badge, albeit with a special "DE" (data exist) notation indicating that although the study was not preregistered, the data analysis plan was. *Psychological Science* articles remain eligible for this badge even if the analysis plan deviates from the preregistered plan, albeit with special "TC" (transparent changes) notation indicating that the researchers have disclosed all such deviations. To the degree that policymakers reliably exhibit this sort of flexibility, rather than the sorts of procrustean policies adopted by CRSP, our concerns that new norms and policies might inadvertently marginalize disciplines like relationship science will be substantially mitigated.

Psychological science will benefit from a careful consideration of the circumstances under which, or the ways in which, deviations from strictly confirmatory procedures may be good or bad for science. Scholars might consider conducting simulations that approximate the kinds of analyses that researchers working with these datasets typically perform (e.g., calculating the false positive rate when two IVs show an effect on three of five DVs, and four out of these six significant effects are moderated by attachment anxiety). Ideally, such simulations would also

account for (a) variation in the preregistration of theoretical propositions and (b) existing empirical evidence relevant to that proposition and the present operationalizations. Estimates of false positive and false negative rates in situations like these would go a long way toward helping scholars who work with large datasets to refine their confirmatory and exploratory hypothesis testing practices to optimize the balance between false-positive and false-negative error rates.

2. Open science 2—prepublication sharing of materials. For every Method or Results section, scientists make choices about what information to include. In psychology, certain information is nearly always included (e.g., the scale endpoints of self-report measures), whereas other information is nearly always excluded (e.g., whether there was a window in the testing room). In principle, the goal has been for scholars to share the information that, in their estimation, would allow the reader to understand the findings and would allow an independent research team to replicate them. Nevertheless, such estimates are imperfect, and important information is often omitted (Brown et al., in press; Kashy, Donnellan, Ackerman, & Russell, 2009).

One element of the evidentiary value movement is a push for scholars to make *all* their procedures and materials available during the review process (Funder et al., 2014; Miguel et al., 2014). This broader sharing of materials is likely to reduce rates of false positives in the published literature because it will arm reviewers with information that increases their ability to identify spurious effects (LeBel et al., 2013). Given that many decisions about which information is relevant versus irrelevant to a given research question are open to wide discretion, scholars may be able to capitalize on this ambiguity to make it appear as though a failed prediction is actually a significant finding, thus generating a false positive. Simmons et al. (2011) used the example of a researcher presenting analyses on only one of two relevant dependent variables that correlate at $r =$

.50, demonstrating that this undisclosed flexibility inflates α from a nominal .050 to a *de facto* .095.[5]

Prepublication sharing of materials and relationship science. When relationship scientists conduct laboratory experiments or one-shot surveys, complete sharing of materials is straightforward and is no more challenging than in other areas of psychology. However, with intensive and/or longitudinal datasets, the number and complexity of questionnaires, manuals, and procedures may be vast. It is not uncommon for graduate students to work under the auspices of a relationship scientist for a year or more before they can comfortably navigate one of these datasets, which can include observational, diary, longitudinal, and physiological components, along with hundreds of pages of documents. For this reason, relationship scientists may react with incredulity to the requirement that researchers submit all materials in a user-friendly format that describes all the data management conventions used by a particular laboratory. Transforming this skepticism into enthusiastic compliance will be a challenge.

How have relationship scientists traditionally policed the temptation to cherry-pick statistically significant findings? Most relationship scientists recognize that a given Method or Results section reveals only a portion of the entire procedure: the portion that the researchers judged to be relevant to the hypotheses being tested. Thus, if reviewers have misgivings about a particular set of findings (perhaps because they suspect that the authors have surreptitiously omitted information that might reduce their odds of a positive publication decision), they usually ask for additional evidence from the same dataset—rather than asking for a new study, which may be an undue

[5] Some scholars have argued that all materials should be made available not only to editors and reviewers during the review process, but also to the *general public* following publication. For example, for submissions to *PSPB*, SPSP's (2013) new policy requires that "authors are required to submit in a separate file stimulus materials, including the verbatim wording (translated if necessary) of all independent and dependent variable instructions, manipulations, and measures. If the article is published, this appendix will be made available on-line." If this online appendix must contain *all* procedures and materials in the study (e.g., measures set aside for other manuscripts) rather than just the subset that is relevant to the published report, then some of the issues we raise below regarding postpublication sharing of data would also apply to this requirement.

burden. For example, if a study includes a measure of relationship satisfaction, reviewers might ask whether similar conclusions emerge with other measures of relationship well-being, such as love or commitment, if such effects are theoretically sensible. Alternatively, reviewers might ask if a finding is moderated by commonly assessed individual differences, such as attachment style or self-esteem, if such moderation is theoretically sensible. For high-impact journals, it would not be unusual for reviewers to ask to see a particular finding replicated in a separate component of the dataset. For example, if a particular association emerged in a diary portion of a study, the reviewer might ask if the finding can be conceptually replicated in either a longitudinal or observational portion of the study (e.g., Neff & Geers, 2013). The onus then shifts to the study authors to address these additional predictions with a combination of additional data and theory; if the initial finding was a false positive, it will often fail to withstand this portion of the review process.

There is no question that reviewers could generate more precise and informed suggestions for additional hypothesis tests if they had access to the entire suite of measures available to the authors. Nevertheless, it is crucial that reviewers generate *theory-derived* objections and alternative hypotheses. If new norms or policies (intentionally or unintentionally) reduce the tendency for reviewers use theory to derive objections, the potential for cherry-picking from the available measures merely shifts from author to reviewer, and the reduction in false positives might be more than offset by the increase in false negatives. In short, the sheer number of variables involved in some relationship science datasets raises complexities that are not characteristic of most laboratory or one-shot survey studies. As we note shortly, these complexities can be addressed—but they should not be ignored or trivialized.

Prepublication sharing of materials—implications for best practices in relationship science. It will often be useful for editors and reviewers to have access to all information about a given study—all procedures, materials, and so forth—when evaluating the scholarly contribution of that

study. For basic laboratory experiments, it is frequently straightforward to submit materials when authors submit their manuscript. For more complex studies, relationship scientists will need to be prepared to upload whatever documentation they have that describes the methods, materials, and procedures that they implemented. To facilitate this process, ScholarOne Manuscripts—the system currently in use at, for example, *Psychological Science* and *PSPB*—permits the author to upload a large number of supplementary documents of various file types (e.g., .doc, .xls, .pdf). When relationship scientists submit a manuscript to these journals, they may need to set aside extra time to prepare and upload such files. The inclusion of an overview document containing a brief description of each supplementary file would help editors and reviewers make sense of the corpus of files. Presumably, these same documents can be uploaded for subsequent submissions containing data from the study in question, so the sharing process should become less onerous over time. Such openness even has the potential to protect authors from accusations of cherry-picking, especially with respect to counterintuitive findings, as it ensures that editors and reviewers are aware that authors have indeed examined their data from all reasonable angles.

Prepublication sharing of materials—implications for psychological science in general. Given the complexity of sharing materials for intensive longitudinal datasets, it is crucial that the guidelines for doing so be extremely clear. To date, this has not always the case. For example, we interpret the *PSPB* mandate to upload "all independent and dependent variable instructions, manipulations, and measures" literally: Scholars must upload *all* their measures and procedures, however many (perhaps hundreds of) pages are required. But this is one interpretation, and other researchers with whom we have spoken have interpreted the mandate differently (e.g., upload the questionnaire items only from concurrent waves of data collection). As far as we have seen, scholars in the evidentiary value movement have devoted little attention to determining precisely *which* materials authors are required to upload, and to what end. Yet this is perhaps the first

question that enters relationship scientists' minds when told that they must upload all of their materials. More importantly, it is not clear *a priori* what information will actually aid editors and reviewers in evaluating a manuscript. A deluge of questionnaires or training procedures for a coding manual may actually be more burdensome than helpful, and it might be more sensible for authors to upload only those questionnaires and details about coding that contain the measures analyzed in the current manuscript. In addition, until intellectual property issues have been considered more deeply (see Discussion section), policymakers should seek to ensure that the study materials do not reside in perpetuity on various editors' and reviewers' computers.

As prepublication sharing of materials starts to become widespread, we suggest that policymakers, including editors, exhibit flexibility regarding the wide variety of complex study designs that characterize much of relationship science. Frequently, relationship scientists can, with minimal exertion, compile a list of questionnaire items. However, it is an altogether different matter to provide documentation sufficient for a naïve researcher to comprehend all elements of the study.[6] This documentation is typically designed for training graduate students and for reminding oneself of all the intricacies of the procedures, not for broad dissemination and comprehensibility as if it were an actual manuscript. As prepublication sharing of materials becomes a priority in our field, implementing this practice might require that editors to reach out to authors to clarify precisely which supplementary documentation, if any, the authors must submit. In many cases involving intensive and/or longitudinal datasets, the editor and the

[6] We encourage those uninitiated with large, complex datasets to peruse the documentation for the Add Health dataset (http://www.cpc.unc.edu/projects/addhealth/data/guides). Add Health is a 4-wave longitudinal study used by many social scientists, and these (excellent) documents were written to help other researchers engage with and understand the Add Health procedures. We hasten to note, however, that the creation of these documents involved the work of 46 different scholars over many years and spans more than 350 pages. Although close relationships researchers' datasets do not have as many participants as Add Health, the level of complexity of the procedures is comparable, and most relationship scientists have but a tiny fraction of the Add Health grant funding. Thus, in today's funding climate, it may not be realistic to expect that most relationship scientists will be able to produce documentation designed for general use like the Add Health documentation.

corresponding author frequently will need to engage in considerable back-and-forth correspondence before the editor develops a deep understanding of all components of the study.

Finally, the current review process in relationship science has achieved a stable, if less than fully open, détente between the reviewer and the author. Changes to the review process should preserve the priority that both the author and the reviewer must place on theory when arguing on behalf of, or against, a given research finding. If prepublication sharing of materials merely makes it harder to publish, the tradeoff between the decrease in false positives and the increase in false negatives (including theoretical false negatives) might yield poor value, especially in controversial areas where reviewers may be motivated to sink a manuscript.

3. Open science 3—postpublication sharing of data. APA guidelines have long required that scholars share the data used to generate published results with "other competent professionals who seek to verify the substantive claims through reanalysis" (American Psychological Association, 8.14, 2010). This guideline was reaffirmed by SPSP in 2013 in their data sharing policy for *PSPB* and *Personality and Social Psychology Review* (*PSPR*). However, there are several reasons to believe that a stronger data sharing policy—one that requires researchers to post their data publicly immediately upon publication—would have many benefits. First, cases of fraud and fabrication are much easier to detect in the raw data than in summary statistics typically reported in journal articles (Dafoe, 2014; Simonsohn, 2013). Second, researchers frequently report having difficulty locating their raw data even just a year after the relevant paper is published (Wicherts et al., 2011; Wicherts, Borsboom, Kats, & Molenaar, 2006); posting the data concurrently with the publication would reduce this loss considerably, especially in cases where the initial corresponding author is no longer reachable (due to death, career change, etc.). Third, if other scholars have access to the raw data, they may be able to detect and correct errors and gather additional information for use in a subsequent meta-analysis (Asendorpf et al., 2013; Bakker & Wicherts, 2011). Fourth, sharing

data opens up the opportunity for other scholars to perform and publish secondary analyses that the original authors might not have the time or inclination to pursue (Wicherts & Bakker, 2012).

To encourage data sharing, *Psychological Science* now offers an Open Data badge, *PLOS One* requires that data be shared upon acceptance for all publications, and the Social Psychology Network offers 10 gigabytes of space to all profile holders for sharing data as well as accompanying codebooks, manuals, and questionnaires. Postpublication sharing of data should enable other scholars (secondary researchers) to examine and publish their own analyses of the data collected by the original researcher (primary researcher). In many cases, these new publications will consist of critiques of the primary researcher's work (e.g., Simonsohn, 2013), although these publications could consist of novel analyses (e.g., Wicherts & Bakker, 2012). For all secondary publications including novel analyses, the primary researchers presumably would retain the rights to be authors if they wished, as "structuring the experimental design" connotes authorship according to the APA ethical guidelines for research.

This discussion of the myriad benefits of postpublication sharing of data, however, requires one major caveat: These benefits are, under some circumstances, trumped by confidentiality considerations. Consider the infamous 2008 case in which researchers published "anonymized" Facebook data from a university in the U.S. that was, within a week, identified by Internet sleuths to be Harvard (Zimmer, 2010). The sleuths identified the university simply by cross-referencing information from the codebook—that there were 819 male and 821 female participants, the presence of one student who self-identified as Albanian, the constellations of academic majors declared by the students, and so forth—with publicly available information about university enrollment and course offerings. With this information in hand, it was a simple task to determine which row of data applied to the Albanian student, for example, and confidentiality was lost. Such

loss of confidentiality is a very serious issue, one that will, under many circumstances, single-handedly outweigh all the benefits of postpublication sharing of data enumerated above.

To be sure, various entities—including the Inter-university Consortium for Political and Social Research (ICPSR), an international association of more than 700 academic institutions and research organizations—are working on data sharing standards for sensitive data, and we applaud these efforts. However, given the frequency of data breaches at major corporations, governmental agencies, and other organizations that have huge incentives to keep data private, we urge great caution before requiring that researchers make confidential information publicly available. Data breaches can cause enormous damage, not only to the individuals whose privacy has been compromised, but also to the social scientific enterprise more broadly.

Postpublication sharing of data and relationship science. Many studies conducted in relationship science will be compatible with the new norm of sharing data immediately upon publication. However, there is one type of relationship science study that may never be publicly sharable: studies with couples. The confidentiality risk here is immense. To a much greater extent than in most other research domains, people are likely to be highly motivated to crack through researchers' efforts to anonymize the data. If a husband, who might be highly motivated to learn private information about his wife, knows a few pieces of information about her (e.g., her age, race, and income), he could likely identify her responses in the raw data and learn things that harm their relationship (e.g., that she is maritally dissatisfied, that she is having an extramarital affair). Even if only a subset of the data were publicly available (e.g., responses to rating scales only), he would simply need to recall his own responses to a handful of rating scale items. With this knowledge, he plausibly could identify his own row in the dataset, and then he could use the couple-indicator to find his wife's row. A breach of this sort could be fatal for relationship science.

Thus, although data from studies of couples can be shared with other competent professionals, perhaps subject to IRB approval, it might be practically impossible to share such data publicly.

Even in cases in which confidentiality issues can be resolved, relationship scientists might wonder whether they are required to share their *entire* dataset upon publishing a first finding from that dataset. A consensus appears to be emerging that scholars should be required to share only those data (and relevant materials; see Footnote 5 above) that would allow secondary researchers to reproduce the results reported in the published article. For example, that degree of sharing is sufficient to earn *Psychological Science*'s Open Data badge. Indeed, those guidelines explicitly state that "Data from the same project that are not needed to reproduce the reported results can be kept private without losing eligibility for the Open Data badge." *PLOS One* now requires the posting of data for all published papers, but the dataset is a "minimal dataset" that includes "the data that are relevant to the specific analysis presented in the paper" (Silva, 2014).

We believe that this emerging consensus strikes a sensible balance because requiring that scholars who conduct intensive and/or longitudinal studies share their entire datasets upon publication would create thorny problems. One set of major problems revolves around the issue of disincentivizing the sort of resource-intensive research that allows for compelling tests of hypotheses that are particularly difficult to test (e.g., Simpson et al., 2007). When scholars publish an initial manuscript from a new large dataset, they are likely drawing from a limited portion of that dataset, and they typically have analysis plans for other portions of the dataset. These plans could take many years to implement in their entirety (especially if they wish to gather additional evidence to corroborate or extend the findings from this study), and, as discussed previously, other ideas that are testable with these data may occur to relationship scientists after the data are collected. Indeed, one of the major incentives for performing this sort of labor-intensive work is that the researchers will have a reservoir of data to devote to different projects over many years.

Relationship scientists typically maintain written or mental lists of how different elements of their datasets are currently allocated to different collaborators, and they work hard to ensure that these projects overlap minimally so that no conflicts emerge during the writing and review process. When manuscripts are not yet fully written, it can be a challenge to prevent overlap among collaborations, especially once the number of projects deriving from a single dataset proliferates. This challenge becomes much more difficult if the broader scientific community can also publish articles based on these data. If researchers cannot effectively manage how their data are allocated across projects and laboratories, problems associated with piecemeal publication will emerge (see "avoiding piecemeal publication" section below).

Postpublication sharing of data—implications for best practices in relationship science. As stipulated by APA guidelines, and to the extent allowed by IRBs and confidentiality considerations, relationship scientists should (continue to) share data from their published reports with competent professionals. In addition, when it is legally and ethically appropriate, relationship scientists should default to the strategy of publicly sharing the relevant variables from their datasets, although they must be extremely cautious about doing so when couples are involved.

Postpublication sharing of data—implications for psychological science in general. To the extent that confidentiality and legal issues can be fully addressed for a given dataset, postpublication sharing of data should become a standard part of the research process. To facilitate such sharing, a growing selection of security-oriented data repositories has emerged in recent years, including DataBib, Datacite, and Re3Data. In addition, the Open Science Framework is working to connect repositories so that there is a single location where researchers can find and deposit data in a repository best fit for their data. This system supports options like GitHub, Amazon S3, Dataverse, Figshare, and Dropbox. As such systems become increasingly user-friendly and secure, policymakers will have to make difficult decisions regarding which private

data must be posted online, who will have access to the data, and so forth. That said, the frequency of major hacking incidents makes us wary of any perspective that downplays the likelihood that scientific data repositories can be hacked, and we urge great caution regarding this security threat.

In terms of repurposing published data, we support the trend to allow secondary researchers to use posted data to critique, evaluate, or confirm a primary researcher's original published piece. Nevertheless, such repurposing presents two major challenges that have, to our knowledge, been widely neglected in the evidentiary value movement. First, scholars seeking to critique the initial research often have a strong motivation to debunk the initial findings, a motivation that might be considerably stronger than the initial researchers' motivation to find evidence for the hypothesis in the first place (Christakis & Zimmerman, 2013, pp. 2499-2500). Second, with intensive datasets, problems arise if secondary researchers can, without consulting with the primary researcher, use publicly available data to test hypotheses that are distinct from the primary researcher's original article. To ensure that the conduct of ambitious studies remains incentivized, science is best served by having the primary researchers retain the ability to allocate data to different projects, and they retain authorship rights on all papers using the posted data to test novel hypotheses, as discussed above. If the secondary researchers are using the publicly posted data to test novel hypotheses, the cover letter accompanying the journal submission should indicate that the primary researcher has approved the use of the data for this project and was given the opportunity to be an author. Without such safeguards, piecemeal publication may become a major concern (see additional discussion below), and the reduced payoffs associated with collecting one's own large intensive datasets could push some relationship scientists away from prioritizing these excellent methods.

4. Close replication. As noted by Galak et al., (2012), Popper (1959/2002, pp. 23–24) defined a scientifically true effect as one that "can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed." Few scholars would disagree with this

definition—indeed, the ability to replicate effects across laboratories is arguably the *sine qua non* of science—but what constitutes a replication, and how replications should be integrated into established scientific knowledge, are more open to debate.

A widely discussed distinction contrasts *direct replication*, in which researchers exactly reproduce an empirical procedure, with *conceptual replication*, in which researchers test the same hypothesis with a different empirical procedure (Schmidt, 2009). Direct replication seeks to solve a problem that is fundamental to the evidentiary value movement, especially within the FPR approach: determining which statistically significant findings in the published literature are true rather than false positives. Nevertheless, many scholars believe that direct replications are impossible in the human sciences—Strack and Stroebe (2014) call them "an illusion"—because certain factors, such as a moment in historical time or the precise conditions under which a sample was obtained and tested, that may have contributed to a result can never be reproduced identically (Brandt et al., 2014; Lykken, 1968; Rosenthal, 1990; Tsang & Kwan, 1999). To address this conundrum, Cook (1990) introduced the concept of "heterogeneity of irrelevancies"—variations in sampling, procedure, or measurement that in principle are immaterial to the focal constructs of a theory or hypothesis, and therefore should not influence a finding. If direct replications are ever to be pursued, researchers must allow for the existence of irrelevant variations (Brandt et al., 2014; Simons, 2014). The difficulty is in determining *a priori* just which facets of a protocol are immaterial and which are material (Schmidt, 2009).[7]

[7] This problem is not unique to psychological research. In response to growing concerns about the frequency of failures to reproduce results in experimental cellular biology, the noted biologist Mina J. Bissell (2013, p. 334) wrote in *Nature* that "it is sometimes much easier not to replicate than to replicate studies, because the techniques and reagents are sophisticated, time-consuming and difficult to master." Bissell places the blame on the complexity of biological research protocols—"The slightest shift in their microenvironment can alter the results—something a newcomer might not spot." It is plausible that the thinking, feeling, contextually sensitive humans who participate in our research are at least as likely, if not more so, to be influenced by small shifts in the research environment.

Psychology has a robust tradition of conceptual replication, but, until the evidentiary value movement reached full strength in recent years, attempts to conduct replications that hew as closely as possible to the original study—what Brandt et al. (2014) have called *close replications*—were rare and have historically been extremely difficult to publish in major journals. As such, scholars had virtually no career-related incentives to conduct close replications; indeed, doing so would have taken resources away from research testing new ideas or employing new procedures, which had a much better chance of being published. The evidentiary value movement has rapidly increased the status of close replications, however, and major journals now explicitly seek to publish them. *PPS'* registered replication reports (Simons et al., 2014) represent a particularly ambitious new endeavor oriented toward close replication of major findings in the field. For example, for the first of these reports, 31 independent laboratories employed substantial sample sizes to replicate (successfully) the same 1990 effect (Alogna et al., 2014). After all, even though it may be impossible to achieve a truly direct replication, there is little doubt that high-powered close replications typically provide compelling tests of the extent to which a finding is a true rather than a false positive (Brandt et al., 2014; Open Science Framework, 2014-a). Such replications should take place in one or more research labs that are independent of the lab that produced the initial effect—albeit in good-faith coordination with that lab, where possible, to ensure the most exact replication possible (Kahneman, in press)—as this independence will reduce the likelihood that some unrecognized aspect of the initial lab's procedures (e.g., the size of the testing cubicles) was crucial in causing the initial effect to emerge.

Close replication and relationship science. The resource-intensiveness of many relationship science studies, particularly longitudinal studies of dyads, means that close replications take on a different character than close replications of laboratory experiments or one-shot surveys (including those conducted within relationship science). In the latter, researchers can seek to conduct a close

replication with relatively little extra effort and expense. In contrast, seeking to conduct a close replication of a 4-year longitudinal study of newlyweds is likely to be very costly in terms of study duration, resources, and researcher commitment. Moreover, it may be impossible to achieve similar exactness in conducting a close replication of many relationship science studies, and the opportunity costs of doing so are especially high. For example, a new study is likely to involve recruitment changes (e.g., participant recruitment via Facebook and Craigslist versus through flyers and newspaper ads) and temporal shifts (e.g., studying marital conflict during a flourishing versus a floundering economy). Consequently, the discrepancies between an original study and its close replication will often be larger than in many other subfields, and the cause of a replication failure will be even more ambiguous than usual (e.g., a false positive in the initial report versus an unknown methodological moderator). In some cases, meta-analytic procedures can help to determine whether methodological variations across studies, such as sample characteristics or year of data collection, moderate an effect (for a recent example, see Eastwick, Luchies, Finkel, & Hunt, 2014; Eastwick, Neff, Finkel, Luchies, & Hunt, 2014).

Close replication—implications for best practices in relationship science. Close replication attempts are crucial for establishing the robustness of a particular finding. How can relationship scientists reconcile the particular challenges of conducting close replications of intensive and/or longitudinal studies with the importance of ensuring that findings from such studies are also subject to such replication attempts? Although it will rarely, if ever, be a good use of resources to conduct an entire investigation that is nothing more than a close replication, we recommend that relationship scientists devote some components of their intensive and/or longitudinal studies to close replications of one or more published findings. Indeed, there is a sense in which close replications are actually easier in intensive relationship science studies than in laboratory studies or one-shot surveys precisely because such studies are so intensive—they can afford to incorporate

close replication components without having those components dominate the study. Furthermore, because many of the self-report instruments in relationship science are standardized and widely used, it will be relatively easy for scholars to include the identical measures of central constructs like relationship satisfaction, commitment, or trust. Such replications are necessary procedures that allow us, as a field, to learn crucial information relevant to an effect's robustness (or lack thereof).

Relationship scientists sometimes perform close or conceptual replications before—rather than after—the publication of an initial finding by seeking to incorporate data from another laboratory. Doing so allows them to publish the initial finding in conjunction with one or more replications (e.g., DeWall et al., 2011; Finkel et al., 2012). To date, scholars have typically used this strategy to publish conceptual rather than close replications (but see Mikulincer, Shaver, Gillath, & Nitzberg, 2005). However, as journals, including *JPSP*, become more receptive to publishing close replications, relationship scientists can capitalize upon across-laboratory collaborations to conduct such replications for publication with the initial finding.

Close replication—implications for psychological science in general. The evidentiary value movement has substantially elevated the status of, and the ability to publish, close replications. In general, these are changes for the better. However, they are not without complication, and we discuss three such complications here. First, given that direct replications are, in a literal sense, impossible, the field requires a robust discussion of what constitutes a close replication (e.g., Brandt et al., 2014). Toward this end, it is instructive to revisit David T. Lykken (1968)'s classic discussion of operational replication, which occupies a middle ground between direct replication (which Lykken calls "literal replication") and conceptual replication (which Lykken calls "constructive replication"). According to Lykken (1968, p. 155), *operational replication* is a procedure in which "one strives to duplicate exactly just the sampling and experimental procedures given in the first author's report of his research" "to test whether the investigator's

'experimental recipe'—the conditions and procedures he considered salient enough to be listed in the 'Methods' section of his report—will in other hands produce the results that he obtained" (p. 155).

Second, however close replications are defined, the evidentiary value movement may have (implicitly) overestimated the degree to which it is equally feasible to conduct close replications across all subfields of psychological science. If decision-makers wish to avoid implementing policies that inadvertently marginalize those subfields for which such replications are especially challenging, it may be necessary to develop flexible standards for determining what constitutes a close replication. Because close replications of many relationship science studies are likely to take longer to conduct (assuming that the relevant data do not already exist in another lab), fewer such replications will be available for a given article. Consequently, in requesting or reviewing close replications, we recommend that editors and reviewers consider the costs required to conduct the replication (time, effort, money, etc.). In addition, compared to most laboratory experiments, close replications of intensive and/or longitudinal studies will generally involve more methodological discrepancies between the original article and the replication attempts. Even though such discrepancies result in some level of additional uncertainty about the extent to which a failed replication is due to a false positive in the initial report or methodological variation across studies, we recommend that the divergences not be considered a basis for discounting a close replication unless there is a particularly compelling argument for doing so.

Third, the field requires a serious discussion about what close replications and conceptual replications can and cannot achieve. Close replications are especially useful for establishing the robustness and magnitude of a very specific effect in a specific context—a particular set of empirical operationalizations. That said, unless we assume that most or all effects in the published literature are false, a polemical assertion by Ioannidis (2005) that we see little reason to accept as

true (e.g., Goodman & Greenland, 2007; Jager & Leek, 2014), conceptual replication will frequently have further-reaching implications for the *theoretical proposition* under investigation. As noted by Donald W. Fiske Donald T. Campbell and (1959), the evidence for a theoretical proposition is bolstered to the extent that it exhibits a theoretically sensible pattern of convergent and discriminant validity across diverse methods and operationalizations. Jacob Westfall, Charles M. Judd, and David A. Kenny (in press) recently argued that replications of studies in which participants respond to a set of experimental stimuli should use a distinct set of stimuli rather than the identical stimuli as the original study. After all, using the same stimuli in the replication studies not only renders conclusions susceptible to idiosyncrasies of the original stimuli, but it also depresses statistical power even as sample size approaches infinity, especially to the extent that the stimulus set is relatively small. In short, the new emphasis on close replications is welcome, but such replications cannot achieve some of the most important features of conceptual replications.

5. Avoiding piecemeal publication. The APA's publication manual defines piecemeal publication as "the unnecessary splitting of the findings from one research effort into multiple articles" (2010, p. 13). Piecemeal publication can produce a situation in which similar effects from the same dataset are published in distinct journal articles, thereby yielding the impression that the evidence for the robustness of a given effect is stronger than it actually is. Of course, the APA manual allows for the possibility that effective communication might require multiple independent reports; in such instances, the overlap with other publications should be noted in both the article and the cover letter.

Avoiding piecemeal publication and relationship science. On its face, the goal of avoiding piecemeal publication does not appear to mesh well with the realities of relationship science and methodologically similar disciplines, in which researchers frequently investigate multiple questions within a single study. Fine and Kurdek (1994) argued that there are two instances in

which publication of multiple empirical reports from a single data set are justified in relationship science: (a) when the findings of multiple articles cannot be integrated into a single article (e.g., because of space limitations or because the conceptual arguments cannot be readily integrated) and (b) when each article has a distinct purpose (e.g., when there is no clear overlap in the relevant literatures). Consistent with the APA policy, Fine and Kurdek suggested that authors make the case for reporting results in independent articles to the editor, a practice that relationship scientists have often neglected.

Avoiding piecemeal publication—implications for best practices in relationship science. Given the above considerations, relationship scientists appear to be at larger-than-typical risk for inappropriate piecemeal publication. After all, many relationship science studies encompass diverse assessments of related constructs, and decisions about when constructs are or are not conceptually related—when they can or cannot be straightforwardly integrated in a single article—are often ambiguous. For example, in a given research program, relationship commitment and trust in the partner may be conceptually distinct constructs to one researcher, but not to another. Our view is that researchers should, all else equal, favor combining measures, if the theoretical account permits it, rather than treating each measure separately (Eastwick, Neff, et al., 2014; Fletcher, Simpson, & Thomas, 2000). When there is a compelling rationale for publishing related constructs in separate articles, the authors should report (perhaps in supplemental materials) associations among constructs across articles and, where relevant, whether the effects in the later articles are robust beyond effects from the earlier articles.

Avoiding piecemeal publication—implications for psychological science in general. We share the APA's perspective on piecemeal publishing—scholars should only use multiple articles to publish results relevant to similar constructs from the same dataset if there is a compelling scientific reason for doing so, and they should ensure that editors and readers are fully aware of

the overlap across the articles. One interesting issue is whether the evidentiary value movement's emphasis on data sharing, which includes a discussion of the benefits of allowing independent scholars to publish off of the shared data (Wicherts & Bakker, 2012), might actually exacerbate piecemeal publishing. To illustrate this possibility, consider the case in which relationship scientists intend to use data from a major longitudinal study to test several independent ideas related to relationship satisfaction. If these scientists are required to share their data for anybody's use following the publication of the first of these ideas, it is plausible (depending upon the publication rules for preexisting data) that independent research teams will publish articles that will, perhaps without anybody's awareness, violate piecemeal publishing norms. It also increases the likelihood that the initial research teams' publication of the other ideas they have always planned to publish will, as a result of these independent publications, turn out to violate these norms, too. As such, we reiterate our recommendation that policymakers handle any new requirements about the public sharing of published data in a manner that is sensitive to the publishing norms for research domains that frequently employ intensive procedures that result in multiple publications.

6. Increasing sample size. Statistical power—the probability that a statistical test will find evidence for an effect that is true in the population (a true positive, or 1–β)—has historically been low in psychological science. One recent literature review suggested that the median sample size per study is N = ~40 across several American Psychological Association (APA) journals (Marszalek, Barber, Kohlhart, & Holmes, 2011), which yields statistical power of .35 to detect a medium-sized difference (at $\alpha = .05$) between two experimental conditions (Bakker et al., 2012). The major journals covering personality and social psychology between 2006 and 2010—*JPSP*, *PSPB*, and *JESP*—had a median sample size per study of N = ~90, which yields statistical power

of .65 to detect such an effect (Fraley & Vazire, 2014). These observed power levels fall below the commonly recommended power of .80 (Cohen, 1988), even for such basic analyses.

Most scholars have long known that increasing the sample size of a study (*N*) increases its statistical power, which reduces the likelihood of a false negative (Cohen, 1962, 1992; Sedlmeier & Gigerenzer, 1989). What many scholars did not recognize until recently is that underpowered studies can also inflate false positives (e.g., Button et al., 2013). Recent simulations have illustrated how researchers can capitalize on *N*s in the range of 40 per study to find false-positive results. Effect sizes tend to be unstable when sample sizes are small (Schonbrodt & Perugini, 2014), and if researchers peek at underpowered datasets before deciding whether to run more participants, they will inflate the false-positive error rate (Simmons et al., 2011). Furthermore, if researchers run underpowered studies (N = ~40) and only publish significant results, the meta-analysis that emerges will produce an effect size that is positively biased; in contrast, a literature of larger studies (N = ~200 each) appears to yield no meta-analytic bias (Bakker et al., 2012). Thus, a central element of the false positives movement is an emphasis on the collection of samples that are large enough to reduce false positives.

Increasing sample size and relationship science. How do relationship science practices measure up in terms of sample size? To address this question empirically, we surveyed relationship science studies published in *JPSP* from 2009 to 2013. *A priori*, we determined that a study would be classified as a relationship science study if participants completed a measure about a current or a past (but not a hypothetical) romantic relationship or romantic partner. A search of PsycINFO using the word "relationship" in any search field in *JPSP* between 2009 and 2013 returned 188 articles, 69 of which contained at least one study that met this criterion. In terms of sample size, the 69 articles measured up relatively well, at least by the benchmarks of ~40 (Marszalek et al., 2011) or ~90 (Fraley & Vazire, 2014). These 69 articles consisted of 179

separate studies of relationships; their median sample size was $N = 122$ (mean = 265, standard deviation = 658, range = 13–6,554).

Relationship science frequently involves different forms of nonindependence across observations. Some of these forms of nonindependence will decrease power relative to the full sample size; for example, 68 of the 179 studies (38%) in this sample assessed variables from both members of a dating or married couple. Because many variables are correlated across couple members (e.g., relationship satisfaction), relationship scientists routinely use multilevel modeling that accounts for this nonindependence and thus permits accurate statistical inference. Although designs with couples give researchers the ability to test important dyadic phenomena, these designs frequently decrease power when variables are correlated within-dyad (in some cases, however, couples designs can also increase power; Kenny, Kashy, & Cook, 2006).

On the other hand, relationship scientists frequently employ designs that increase power, such as diary and longitudinal designs. Both of these designs obtain multiple reports from the same participants over time, thereby increasing power on average. Indeed, 23 of the 179 studies (13%) used a diary design and 40 of the 179 studies (22%) used a longitudinal design; 56 of the 179 studies (31%) included one or both of these features. The diary studies averaged 19.8 assessments from participants (typically daily), and longitudinal studies averaged 4.3 assessments from participants (typically annually or semiannually). The power that researchers lose by examining couples can be offset by diary and longitudinal designs, and this tradeoff appears to be reflected in the literature—49% of studies of couples used a diary or longitudinal design, whereas only 21% of studies of individuals used a diary or longitudinal design.

Table 2 illustrates the effect sizes that relationship scientists will be able to detect reliably given the sorts of sample sizes and designs reflected in our survey of this literature. The first row in Table 2 indicates the size of a correlation, one of the simplest hypothesis tests in this literature,

that a study has the power of .80 to detect at $\alpha = .05$ (two-tailed) given sample sizes at the 25th percentile ($N = 80$), 50th percentile ($N = 122$), and 75th percentile ($N = 232$). With no forms of nonindependence in the data, researchers using samples at the 25th percentile have adequate power to detect medium-sized correlations (according to Cohen's, 1988, conventions of small $r = .10$, medium $r = .30$, and large $r = .50$). Researchers using larger samples have the power to detect correlations that vary from small to medium in size.

The precise calculation of statistical power with nonindependent designs (e.g., longitudinal, diary, couples) is complicated and requires more than a rough estimate of effect size. When there are multiple sources of random variability in a design (e.g., couples taking part in a diary study), the most accurate method of determining the power of many common tests requires simulations that capitalize on random effects revealed in real, preexisting data (Bolger & Laurenceau, 2013; Bolger, Stadler, & Laurenceau, 2011). Thus, if researchers are required to perform power analyses with no existing data on hand, calculations that account for nonindependence will be rough approximations—much more so than in designs with independent observations.

Nevertheless, some relatively simple formulae are available that aid in the estimation of power if the design contains only one source of nonindependence (Scherbaum & Ferreter, 2009; Snijders & Bosker, 2012). Snijders and Bosker (2012, p. 24) calculate the "effective sample size," which is typically larger than the number of actual participants in a diary or longitudinal design, as follows:

$$\text{Effective sample size} = Nk/(1 + (k - 1)*ICC) \quad (1)$$

In this formula, N is the original sample size, k is the number of repeated observations, and ICC, the intraclass correlation, is the extent to which dependent measure observations are correlated within-participant over time. This formula is designed for cases in which the independent variable is measured at Level 2 (e.g., an individual difference predictor in a diary design). In the typical daily diary or experience sampling study, Level 1 hypothesis tests (e.g., a time-varying predictor)

tend to be better-powered than Level 2 tests, so this Level 2 effective sample size formula generates a lower-bound estimate of power.

The second and third rows of Table 2 approximate the effect size of a Level 2 variable that can be detected with a power of .80 if researchers implement diary and longitudinal designs. For the diary and longitudinal rows of Table 2, we derived from our *JPSP* survey values for *k* of 20 and 4, respectively, and we estimated *ICC* to be .70, which is a conservative (i.e., high) estimate of the extent to which observations correlate within-participant across time points. As Table 2 illustrates, the use of diary and longitudinal designs lowers the size of the correlation that researchers can detect with a power of .80, but this decrease is fairly modest. (Formulae described by Scherbaum and Ferreter, 2009, produce results that are, to the hundredths place, identical to these estimates.)

The fourth row in Table 2 approximates the effect size that researchers can detect with a power of .80 if they implement a couples design—specifically an Actor-Partner Interdependence Model (APIM) design. The formula (Kenny et al., 2006, p. 180) used to make this adjustment is:

$$\text{Effective sample size} = N/(1 + ICC^2) \qquad (2)$$

In this case, we estimated *ICC* to be .45, which is a conservative (i.e., high) estimate of the extent to which observations correlate within-dyad—a value that Kenny et al. (2006, p. 58) called "consequential nonindependence." Table 2 illustrates that a design with couples raises the size of the correlation that researchers can detect with a power of .80, but this increase is fairly modest.

The bottom half of Table 2 repeats these calculations for effect size *q*, which is a test of the difference between two correlations. This test is akin to the test of a statistical interaction between a categorical variable (e.g., participant sex) and a continuous variable. Effect size *q* can be interpreted similarly to effect size *r* in that, by convention, .10 indicates a small effect, .30 indicates a medium effect, and .50 indicates a large effect (Cohen, 1988). Table 2 reveals that sample sizes at the 50th percentile only have the power of .80 to detect large *q*s; sample sizes

exceeding the 75th percentile are required to detect medium-sized *q*s. Small *q*s may only be detectable in unusually large designs (i.e., thousands of participants) or in a meta-analysis.

Increasing sample size—implications for best practices in relationship science. Compared with previous estimates of power in APA journals (N = ~40; Marszalek et al., 2011) and in personality/social psychological journals (N = ~90; Fraley & Vazire, 2014), relationship scientists are doing pretty well: Most relationships studies in *JPSP* have a power of at least .80 to detect medium-sized effects for simple correlational hypotheses. However, when it comes to N, bigger is better; increasing N is one recommendation of the evidentiary value movement that (within a given sample) decreases rates of both false positives and false negatives.[8] Thus, although relationship scientists are on relatively solid ground, they should pursue even larger samples, especially when testing for small effects or moderational hypotheses.

Increasing sample size—implications for psychological science in general. It is important for policymakers and journal editors to recognize that the calculation of power in nonindependent data is much more complex than estimating an effect size and picking up a copy of Cohen (1988). Complicating matters further, scholars who analyze multilevel data often have good reason to favor unstandardized over standardized regression coefficients (Hox, 2010), so relationship scientists often do not, or cannot, use the same currency as in other areas of psychology (e.g., *d*s and *r*s). New requirements must allow for approximate power estimates in such cases to avoid inadvertently marginalizing subfields like relationship science, which is setting a relatively strong example when it comes to sample size and power.

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

Discussion

In this article, we contrasted the FPR and the EB approaches to maximizing the evidentiary value of our science, and we argued that policymakers seeking to optimize scientific conduct are best served by developing policies that are flexible enough to allow for variation in appropriate methods across research domains. Rather than addressing these topics exclusively in the abstract, we provided a concrete case study (of relationship science) to illustrate the sorts of nuances and complexities that arise when seeking to harness the potential of the evidentiary value movement to optimize scientific practice. In this final section, we summarize our central points and address several broader issues that the field is confronting in light of the evidentiary value movement.

The Error Balance Approach

At an epistemological level, we argued that the evidentiary value movement's dominant focus to date—the reduction of false-positive rates—is too narrow. A discussion of best practices should take account of both false positives and false negatives, and it is ill-advised to alter policies based on consideration of only one type of error. Tradeoff considerations involving false positives and false negatives must also be attentive to the practical value of each type of error in the relevant context (e.g., incorrectly concluding that, versus incorrectly failing to find evidence that, a domestic violence intervention is effective). As discussed previously, many policy changes oriented toward reducing false-positive rates will exacerbate false-negative rates, so it is crucial to consider tradeoffs when evaluating the optimal way to improve scientific conduct in light of the evidentiary value movement.

We are especially concerned about the evidentiary value movement's relative neglect of false negatives because, for at least two major reasons, false negatives are much less likely to be the subject of replication attempts. First, researchers typically lose interest in unsuccessful ideas, preferring to use their resources on more "productive" lines of research (i.e., those that yield

evidence for an effect rather than lack of evidence for an effect). Second, others in the field are unlikely to learn about these failures because null results are rarely published (Greenwald, 1975). As a result, false negatives are unlikely to be corrected by the normal processes of reconsideration and replication. In contrast, false positives appear in the published literature, which means that, under almost all circumstances, they receive more attention than false negatives. Correcting false positive errors is unquestionably desirable, but the consequences of increasingly favoring the detection of false positives relative to the detection of false negatives are more ambiguous.

To be sure, there are cases in which an increasing emphasis on the reduction of false positives can simultaneously reduce the prevalence of false negatives. In particular, as the field increasingly prioritizes larger sample sizes and greater statistical power, the incidence of false negatives will be reduced. In addition, new vehicles for disseminating the results of replications are increasing researcher options for publicizing null results (e.g., American Psychological Association, 2014; Brandt, Crawford, & Giner-Sorolla, 2014; Nosek & Lakens, 2014), thereby making other researchers aware of these null effects and offering the opportunity to conduct the sorts of replications that can help to ferret out false negatives. Nevertheless, false negatives remain less likely to be corrected than false positives.

Pragmatic Considerations

At a pragmatic level, we argued that policies that are well-suited to one research domain are sometimes poorly suited to another. Research domains vary in the questions they ask and, consequently, in the optimal methods for seeking answers to those questions. For example, it might be scientifically and practically sensible in some domains to require that scholars pursue a two-step process in which they first run an exploratory study and then run a follow-up study that deviates not one iota from preregistered methodological and data-analytic plans (Wagenmakers et al., 2012). But, as discussed previously, that approach may be nonsensical when data collection

involves intensive and/or longitudinal methods. Indeed, 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 (Simonsohn et al., 2011).

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. 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. After all, a major purpose of these changes—awarding badges and the like—is to equip scholars with a quick-and-dirty indicator of which studies, and which scholars, have strong research integrity. It is easy to imagine a future in which many scholars assess as scientifically valid only those articles that have been honored with at least one (or maybe all) of the research integrity badges.

Perhaps some scientists would applaud such developments, believing that they would help to bolster subfields with stronger research integrity and marginalize fields with weaker research integrity. Before we, as a field, continue down that road, however, it is crucial for scholars to think deeply about what sorts of topics are important to study and what sorts of methods help us study those topics well. Do we care about understanding which interaction processes early in a marriage place couples at elevated risk for divorce? Do we care about how domestic violence influences

children's cognitive development throughout grade school? Do we care about how couples deal with the deterioration in cognitive and social functioning that occurs when one spouse develops Alzheimer's disease? If the answer to such questions is yes, then policymakers inspired by the evidentiary value movement must exert themselves to ensure that they have considered the reverberations of potential policy changes across research domains before formalizing those changes.

Beyond this general point, it is important to note that intensive and/or longitudinal research may well be a crucial pathway through which the field can reduce false-positive rates. Sanjay Srivastava (2014) has argued that one means of optimizing research practice in light of the evidentiary value movement is to "go intensive." Specifically, he recommends that scholars (a) employ "data-intensive, multilevel designs"; (b) "probe variability of effects across contexts and people"; and (c) "study intraindividual variation and within-person causation." We share the view that, all else equal, statistically significant findings from such studies, relative to those from studies that employ less intensive procedures, are much more likely to be true rather than false positives. Yet these studies will be best analyzed with a mix of confirmatory and exploratory approaches, they will be difficult to replicate directly, they will have some methodological elements that are challenging to understand even when all the researcher's documents have been uploaded, and they are at risk for yielding piecemeal publication if the owners of the dataset cannot manage how the data are allocated to distinct projects. As the evidentiary value movement and relationship science work to find common ground on these issues, both entities stand to reap substantial benefits.

Some Issues that Warrant Robust Discussion

One of our major goals in this article is to raise issues that have been largely neglected, or at least insufficiently addressed, in the conversation emerging from the evidentiary value movement.

Our view is that many of the recommendations emerging from the movement trigger questions that require compelling answers if the movement is going to alter the field in the most constructive manner possible. We address four such questions here.

What is the optimal balance of promotion versus prevention focus in our scientific conduct? The increasing prioritization of false positives over false negatives is likely to influence the "regulatory focus" (Higgins, 1997) scholars adopt when engaged in scientific endeavors—the extent to which they conduct research in a *promotion focused* mindset oriented toward the detection of true effects (even at the expense of an increased likelihood of false positives) versus a *prevention focused* mindset oriented toward reducing the likelihood of mistakenly believing a false effect to be true (even at the expense of an increased likelihood of false negatives). Both a promotion and a prevention focus can be useful and constructive means of goal pursuit, but they engender distinct strategic orientations, and these strategic orientations have implications for which sorts of scientific findings scholars are likely to investigate versus neglect (Higgins, 1992).

To clarify these strategic considerations, we revisit Figure 1. According to regulatory focus theory (Higgins, 1997), individuals in a promotion focus tend to focus on the left side of Figure 1: They tend to pursue true positives (upper-left quadrant) and seek to avoid false negatives (lower-left quadrant). In contrast, individuals in a prevention focus tend to focus on the right side of Figure 1: They tend to pursue true negatives (lower-right quadrant) and seek to avoid false positives (upper-right quadrant). Scientists adopting a promotion focus tend to adopt an *eager* means of goal pursuit oriented toward ensuring that they have discovered every opportunity to learn new truths, whereas scientists adopting a prevention focus tend to adopt a *vigilant* means of goal pursuit oriented toward ensuring that they have not concluded that a false effect is true. As noted by Rosnow and Rosenthal (1989, p. 1278), whereas "Type I [false-positive] errors may be thought of as inferential errors of gullibility or overeagerness, that is, an effect or a relationship is

claimed where none exists" (a promotion-focused error); "Type II [false-negative] errors may be thought of as inferential errors of conservatism or blindness, that is, the existence of an effect or a relationship that does exist is denied" (a prevention-focused error).

If the evidentiary value movement yields a marked shift away from promotion-focused and toward prevention-focused scientific conduct, it is likely that this shift will have substantial implications for the sorts of research questions scholars pursue, how they seek to answer those questions, what sorts of findings they obtain, and what sort of theories they support. Indeed, the adoption of an eager versus a vigilant emphasis has major downstream implications for how people pursue goals in general (for a review, see Molden, Lee, & Higgins, 2008). When applying these ideas to scientific practice, individuals adopting a stronger promotion (vs. prevention) focus are likely to be (a) more sensitive to opportunities to discover new knowledge but less sensitive to opportunities to ensure that existing knowledge is accurate, (b) more likely to consider a broad range of evidence as relevant to a given hypothesis, (c) more likely to prefer change over stasis, (d) more likely to conduct research efficiently rather than diligently, and (e) more likely to adopt a holistic and integrative cognitive mindset and less likely to adopt narrow and detail-oriented mindset. This list is not exhaustive, but it illustrates how profoundly a marked shift from a promotion-focused to a prevention-focused—from a riskier to a more conservative—pursuit of science can alter the nature of scientific inquiry. As noted by the Nobel laureate Peter B. Medawar (1969, p. 7): "The exposure and castigation of error does not propel science forward, though it may clear a number of obstacles from its path." We need to strike an optimal balance between propelling forward and clearing obstacles.

What is the primary function of our empirical journals? This second question is related to the first. Should our journals function as a "repository of the accumulated knowledge of a field" (American Psychological Association, 2010, p. 9), archiving for posterity the accumulated correct

wisdom of our discipline? Or should they function as a medium through which scholars communicate to their colleagues the fruit of their efforts, helping others "to avoid needlessly repeating work that has been done before, to build on existing work, and in turn to contribute something new" (American Psychological Association, 2010, p. 9)? This latter perspective capitalizes on the (contestable) idea of science as a self-correcting enterprise—that sooner or later, weak or erroneous theories or findings are replaced by more accurate ones. Often, this process is initiated when researchers become aware of a finding or interpretation that seems questionable to them, leading them to conduct research that demonstrates why the original account was flawed.

Fiedler and colleagues (2012, p. 667) observe that the dominance of the FPR approach within the evidentiary value movement makes the "context of justification and hypothesis testing" superordinate to "the context of discovery." That is, the reduction of false positives trumps or supersedes the pursuit of more speculative discoveries, which increasingly will go unreported and hence unable to serve as a stimulus for advancing the work of others. To the extent that the field leans more strongly toward the reduction of false positives, journals will publish fewer findings that are intriguing and novel but not (yet) robustly supported. Such changes might carry the unintended side-effect of inhibiting researchers from exploring bold new ideas or from challenging conventional wisdom (Higgins, 1992; Wegner, 1992). To be sure, in the Internet era, other venues exist for disseminating new ideas (Nosek & Bar-Anan, 2012). But as long as publication in leading journals remains the field's primary gauge of scientific contribution—and the basis for most hiring and promotion decisions—conceptualizing the function of these journals more strongly in terms of publishing scientific truths, and less strongly in terms of updating colleagues on a research program in progress, will tilt our science in a more conservative direction. Whether such a tilt is good or bad for our science warrants careful consideration.

To what extent should the pursuit of best practices in psychology account for intellectual property considerations? Thus far, we have largely sidestepped intellectual property, a particularly thorny issue that resides at the intersection of the evidentiary value movement and psychological science. We have sidestepped this issue because, despite considerable thought and discussion, we (the authors of this report) have been unable to develop confident conclusions about how the field should address it. The root of the complexity is that such considerations sometimes pit the best interests of the scientist against the best interests of the science.

The easy response here is to say that no individual scientist matters; only the science matters. From this perspective, if the individual scientist toils for years to produce an impressive dataset, we should not worry if her ideas are scooped—with her own data—before she is able to publish them; the only thing that matters is that the data that exist are used to advance the science. We are wary of certain implications of that response, however, and not only because secondary publications may in fact undermine the science by partially scooping the (perhaps better) ideas she had been planning to publish (see the "postpublication sharing of data" section above). We are also wary because of the personal harm such practices can inflict upon her. It seems unfair—and perhaps even illegal or unethical—that she might be at elevated risk for a negative tenure vote in part because her publication plans were undermined by other scholars' use of her own data. In general, the sorts of 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. Until these issues are addressed, policymakers eager to maximize the openness of our science may wish to adopt a cautious approach to topics that could potentially undermine scientists' intellectual property. Such issues are relevant, for example, when determining whether scientists must make

available not only the measures and data from a published report, but also all other measures or data from the broader study.

To what extent should publication decisions be made by legislative bodies versus by journal editors? A major emphasis in the evidentiary value AFP movement is the implementation of new rules and norms that can bolster the quality of our science. As noted by a reviewer of the initial submission of this article, this emphasis on stricter rules of conduct serves to move some level of decision-making power from editors to legislative bodies. This movement toward stricter rules—and away from subjective editorial decisions—is predicated on noble and compelling principles, but a power shift from editors to legislative bodies may have underappreciated risks, including the possibility that the list of rules becomes long and unwieldy, potentially even counterproductive (for an analogy, consider what has happened with IRBs in recent decades). In the end, expert editors must take responsibility for making subjective judgments about the overall magnitude of contribution offered by a manuscript (King, 2012). In the words of the reviewer: "The idea that we can construct some kind of editorial formulary that will free us of these sorts of responsibilities is an illusion—a possibly dangerous one."

One issue that will be especially difficult to fit into some form of editorial formulary is that any specific finding must be contextualized within the broader theoretical and empirical knowledge base (e.g., Rosnow & Rosenthal, 1989); in isolation, it cannot provide definitive evidence for the veracity or inveracity of a given theoretical proposition. Consider the case in which Schimmack's (2012) "incredibility index" suggests that a series of published studies almost certainly omits at least one instance of file-drawering or *p*-hacking. The extent to which that fact undermines our confidence in the article's conclusions should be lower when the conclusions align with a broadly supported theoretical context than when it does not. Of course, some studies (including those with large sample sizes) are more compelling than others, and those studies

should be given more weight when assessing the overall corpus of evidence regarding a theoretical proposition. But no study is sufficient, on its own, to tell the full story regarding that proposition.

Conclusion

The evidentiary value movement has focused attention on widespread practices in psychological science that increase false-positive error rates and has, with breathtaking pace, opened the door to significant changes in the norms of scientific conduct and in official policies at our field's major journals. On balance, we are enthusiastic about these developments. We suggest that the movement's positive influence will be maximized by an increased emphasis on false-negative error rates and by the adoption of rules and policies that are flexible enough to account for variation across research domains in optimal methodologies. To the extent that it does so, research practices in our field will be better—in terms of scientific discovery and validity—in 2020 than they were in 2010, which will be a great credit to the movement.

References

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

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2014). Journal of Personality and Social Psychology: Instructions to Authors. Retrieved June 21, 2014, from http://www.apa.org/pubs/journals/psp/

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

Association for Psychological Science (2014). *Registered replication reports*. Downloaded 13 September 2014 from http://www.psychologicalscience.org/index.php/replication.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543-554.

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666-678.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Bissell, M. (2013). Reproducibility: The risks of the replication drive. *Nature*, *503*(7476), 333-334.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York: Guilford.

Bolger, N., Stadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies (285-301). In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life*. New York: Guilford.

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

Brandt, M. J., Crawford, M. T., & Giner-Sorolla, R. (2014). *Call for submissions: Special issue on pre-registered research*. Downloaded 19 June 2014 from http://www.journals.elsevier.com/journal-of-experimental-social-psychology/call-for-papers/call-for-submissions-special-issue-on-pre-registered-researc/.

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

Brown, S. D., Furrow, D., Hill, D. F., Gable, L. P., Porter, L. P., & Jacobs, J. (in press). A duty to describe: Better the devil you know that the devil you don't. *Perspectives on Psychological Science*.

Burgess, E. W. (1926). The family as a unity of interacting personalities. *The Family,* 7, 3-9.

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*(5), 365-376.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81-105.

Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *The Quarterly Journal of Economics*, 1755-1812. doi:10.1093/qje/qje027.

Chamberlin, J. (2000). A student publishing tradition: The student-run, student-reviewed Representative Research in Social Psychology celebrates 30 years. *APA Monitor*, *31*, 36.

Christakis, D. A., & Zimmerman, F. J. (2013). Rethinking reanalysis. *JAMA*, *310*, 2499-2500.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145-153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). New York: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, *505*(7485), 612-613.

Cook, T. D. (1990). The generalization of causal connections: multiple theories in search of clear practice. In L. Sechrest, J. Bunker & E. Perrin (Eds.), *Research methodology: Strengthening causal interpretation of non-experimental data* (pp. 9-31). PHS Pub. No. 90-3454. Rockville, MD: Agency for Health Care Policy & Research.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*, 671-684.

Dafoe, A. (2014). Science deserves better: The imperative to share complete replication files. *PS: Political Science & Politics*, *47*, 60-66.

De Groot, A. D. (1969/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta psychologica*, *148*, 188-194.

DeWall, C. N., Lambert, N. M., Slotter, E. B., Deckman, T., Pond, R. D., Finkel, E. J., Luchies, L. B., & Fincham, F. D. (2011). So far away from one's partner, yet so close to romantic alternatives: Avoidant attachment, interest in alternatives, and infidelity. *Journal of Personality and Social Psychology*, *101*, 1302-1316.

Eastwick, P. W., Luchies, L. B., Finkel, E. J, & Hunt, L. L. (2014). The predictive validity of ideal partner preferences: A review and meta-analysis. *Psychological Bulletin, 140,* 623-665.

Eastwick, P. W., Neff, L. A., Finkel, E. J., Luchies, L. B., & Hunt, L. L. (2014). Is a meta-analysis a foundation, or just another brick? Comment on Meltzer, McNulty, Jackson, and Karney. *Journal of Personality and Social Psychology*, *106*, 429-434.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*(1), 3-6.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891-904.

Fazio, R. H., Zanna, M. P., & Cooper, J. (1977). Dissonance and self-perception: An integrative view of each theory's proper domain of application. *Journal of Experimental Social Psychology*, *13*(5), 464-479.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555-561.

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.

Fine, M. A., & Kurdek, L. A. (1994). Publishing multiple journal articles from a single data set: Issues and recommendations. *Journal of Family Psychology*, *8*(4), 371-379.

Finkel, E. J., DeWall, C. N., Slotter, E. B., McNulty, J. K., Pond, R. S., Jr., & Atkins, D. C. (2012). Using I^3 theory to clarify when dispositional aggressiveness predicts intimate partner violence perpetration. *Journal of Personality and Social Psychology*, *102*, 533-549.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fiske, S. T. (2014, October). Change is coming: A new day for human subjects research participation. *APS Observer*, *8*, 5/21.

Fletcher, G. J., Simpson, J. A., & Thomas, G. (2000). The measurement of perceived relationship quality components: A confirmatory factor analytic approach. *Personality and Social Psychology Bulletin*, *26*(3), 340-354.

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.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151-156.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the Dependability of Research in Personality and Social Psychology Recommendations for Research and Educational Practice. *Personality and Social Psychology Review*, *18*, 3-12.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*, 933.

Goodman, S., & Greenland, S. (2007). Why most published research findings are false: problems in the analysis. *PLoS medicine*, *4*(4), e168. DOI:10.1371/journal.pmed.0040168

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

Greenwald, A. G. (1976). An editorial. *Journal of Personality and Social Psychology, 33*, 1-7.

Higgins, E. T. (1992). Increasingly complex but less interesting articles: Scientific progress or regulatory problem? *Personality and Social Psychology Bulletin*, *18*(4), 489-492.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280-1300.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd edition). New York: Psychology Press.

Humphreys, M., de la Sierra, R. S., & Van der Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, *21*, 1-20.

Huston, T. L., Caughlin, J. P., Houts, R. M., Smith, S. E., & George, L. J. (2001). The connubial crucible: newlywed years as predictors of marital delight, distress, and divorce. *Journal of Personality and Social Psychology*, *80*, 237.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. Retrieved from http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19,* 640–648. doi:10.1097/EDE.0b013e31818131e7

Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, *15*, 1-12.

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

Kahneman, D. (2012, September 26). *A proposal to deal with questions about priming effects* [Open letter to social priming researchers].

Kahneman, D. (in press). A new etiquette for replication. *Social Psychology*.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in *PSPB*: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*(9), 1131-1142.

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

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217.

King, L. A. (2012). A Dinosaur Comments on the Coming Apocalypse: Does Anybody Else See That Asteroid? *Psychological Inquiry*, *23*, 274-276.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS one*, *9*, e105825.

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

Lehrer, J. (2010). The truth wears off: Is there something wrong with the scientific method. *The New Yorker*, 52-57.

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

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, *9*, 343-351.

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*(2), 331-348.

Medawar, P. B. (1969). *Induction and intuition in scientific thought.* Philadelphia, PA: American Philosophical Society.

Miguel, E., Camerer, C., Casey, K. , Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R, Simmons, J. P., Simonsohn, U., and Van Der Laan, M. (2014). Promoting transparency in social science research. *Science, 343*, 30-31.

Mikulincer, M., Shaver, P. R., Gillath, O., & Nitzberg, R. A. (2005). Attachment, caregiving, and altruism: boosting attachment security increases compassion and helping. *Journal of personality and social psychology*, *89*(5), 817-839.

Molden, D. C., Lee, A. Y., & Higgins, E. T. (2008). Motivations for promotion and prevention. In J. Shah & W. Gardner (Eds.), *Handbook of motivation science* (pp.169-187). New York: Guilford.

Neff, L. A., & Geers, A. L. (2013). Optimistic expectations in early marriage: a resource or vulnerability for adaptive relationship functioning? *Journal of Personality and Social Psychology*, *105*(1), 38-60.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*(3), 217-243.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Open Science Framework (2013, September 3). *Badges to acknowledge open practices*. Downloaded 24 May 2014 from https://osf.io/tvyxz/wiki/view/.

Open Science Framework (2014-a, May 21). *Open science framework*. Downloaded 24 May 2014 from https://osf.io/4znzp/wiki/home.

Open Science Framework (2014-b, May 31). *Registered reports: A method to increase the credibility of published reports*. Downloaded 1 June 2014 from https://osf.io/zv2cs/.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536.

Popper, K. R. (1959/2002). *The logic of scientific discovery.* New York: Routledge.

Reis, H. T. (2012). A brief history of relationship research in social psychology. In A. W. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 363-382). New York: Psychology Press.

Reis, H. T., Collins, W. A., & Berscheid, E. (2000). The relationship context of human behavior and development. *Psychological bulletin*, *126*(6), 844-872.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638-641.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–777.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276-1284.

Sbarra, D. A. (2014). Forward thinking: An introduction. *Perspectives on Psychological Science*, *9*, 443-444. DOI: 10.1177/1745691614539905

Scargle, J. D. (2000). Publication bias: the "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*(1), 91-106.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*(2), 347-367.

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

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90-100.

Schnall, S. (in press). Clean data: Statistical artefacts wash out replication efforts. *Social Psychology*.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, *54*(2), 93-105.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309-316.

Silva, L. (2014, February 24). PLOS' new data policy: Public access to data. Downloaded 24 May 2014 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.

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

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

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875-1888.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534. DOI: 10.1037/a0033242

Simpson, J. A., Collins, W. A., Tran, S., & Haydon, K. C. (2007). Attachment and the experience and expression of emotions in romantic relationships: a developmental perspective. *Journal of personality and social psychology*, *92*(2), 355-367. doi:10.1037/0022-3514.92.2.355

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd edition).

Srivastava, S. (2014, February). *How can we make psychology less sciencey and more scientific?* Paper presented at the Society for Personality and Social Psychology Annual Meeting, Austin, TX.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*, 108–112.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59-71.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological science in the public interest*, *1*(1), 1-26.

Tanner Jr., W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401.

Terman, L. M. (1938). *Psychological factors in marital happiness*. New York, NY: McGraw-Hill.

Tidwell, N. D., Eastwick, P. W., & Finkel, E. J. (2013). Perceived, not actual, similarity predicts initial attraction in a live romantic context: Evidence from the speed-dating paradigm. *Personal Relationships, 20*(2), 199-215.

Tsang, E.W., & Kwan, K.M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review, 24*, 759–780.

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*(6), 632-638.

Wegner, D. M. (1992). The premature demise of the solo experiment. *Personality and Social Psychology Bulletin*, *18*, 504-508.

Westfall, J., Judd, C. M., & Kenny, D. A. (in press). Replicating studies in which samples of participants respond to samples of stimuli. *Current Opinion in Psychological Science*.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40*(2), 73-76.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6*, e26828. doi:10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726-728.

Zhong, C-B., & Liljenquest, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451-1452. DOI: 10.1126/science.1130726.

Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology*, *12*(4), 313-325.

Author Note

Eli J. Finkel, Department of Psychology and Department of Management and Organizations, Northwestern University; Paul W. Eastwick, Department of Human Development and Family Sciences, University of Texas at Austin; Harry T. Reis, Department of Clinical and Social Sciences in Psychology, The University of Rochester.

We thank Galen Bodenhausen, Jim Coan, Lydia Emery, Lisa Neff, Leif Nelson, Brian Nosek, and Tessa West for their insightful feedback on an earlier draft of this manuscript. These generous scholars do not bear any responsibility for the content or recommendations in this article, but their input helped us think more deeply about the issues at play.

Correspondence should be addressed to Eli J. Finkel, Northwestern University, 2029 Sheridan Road, Swift Hall Rm. 102, Evanston, IL, 60208-2710. E-mail: finkel@northwestern.edu.

Table 1. The error balance approach: Key tenets

Tenet Number	Tenet
1	Both false positives and false negatives undermine the superordinate goals of science, which are discovery and validity.
2	Neither type of error is uniformly a greater threat to validity than the other type.
3	Any serious consideration of optimal scientific practice must contend with both types of error simultaneously.

Table 2. Detectable effect sizes at α = .05 (two-tailed) and power = .80 given relationship science sample sizes in *JPSP*.

Effect size type	Form of nonindependence	25th Percentile N = 80	50th Percentile N = 122	75th Percentile N = 232
Correlation (*r*)	None (independent)	.30	.25	.18
	Longitudinal	.27	.22	.16
	Diary	.26	.21	.15
	Couples	.33	.27	.20
Difference between correlations (*q*)	None (independent)	.65	.52	.37
	Longitudinal	.57	.45	.33
	Diary	.54	.44	.31
	Couples	.72	.57	.41

The columns representing the 25th, 50th (median), and 75th percentile are conditioned on the relationship science sample sizes in the *JPSP* literature based on our 2009-2013 survey (see main text). *N* refers to the number of individual participants in the study (i.e., the 50th percentile is equal to *N* = 61 couples). Longitudinal and diary power calculations were conducted using the average length of longitudinal and diary designs in the literature (4 waves and 20 days, respectively). Interdependence (conservatively) estimated at *ICC* = .70 in longitudinal/diary designs (i.e., the correlation between a participant's reports at two time points) and *ICC* = .45 between couple members (i.e., the correlation between two partners' reports on the same variable at the same time point).

Figure 1. *A signal detection analysis of the logic underlying hypothesis testing.*

	H_1 is True in the Population	H_1 is False in the Population
Study Yields Statistically Significant Evidence for H_1	True Positive Correct Rejection of H_0	False Positive Incorrect Rejection of H_0
Study Does Not Yield Statistically Significant Evidence for H_1	False Negative Incorrect Nonrejection of H_0	True Negative Correct Nonrejection of H_0

Note. H_0 refers to the null hypothesis that there is no effect in the population. H_1 refers to the alternative hypothesis that there is an effect in the population. Other names for false positive are "α-error" and "Type I error"; other names for false negatives are "β-error" and "Type II error."