PSYCHOMETRIKA
DOI: 10.1007/s11336-015-9447-z

THINKING ABOUT DATA, RESEARCH METHODS, AND STATISTICAL ANALYSES: COMMENTARY ON SIJTSMA'S (2014) "PLAYING WITH DATA"

IRWIN D. WALDMAN AND SCOTT O. LILIENFELD

EMORY UNIVERSITY

We comment on Sijtsma's (2014) thought-provoking essay on how to minimize questionable research practices (QRPs) in psychology. We agree with Sijtsma that proactive measures to decrease the risk of QRPs will ultimately be more productive than efforts to target individual researchers and their work. In particular, we concur that encouraging researchers to make their data and research materials public is the best institutional antidote against QRPs, although we are concerned that Sijtsma's proposal to delegate more responsibility to statistical and methodological consultants could inadvertently reinforce the dichotomy between the substantive and statistical aspects of research. We also discuss sources of false-positive findings and replication failures in psychological research, and outline potential remedies for these problems. We conclude that replicability is the best metric of the minimization of QRPs and their adverse effects on psychological research.

Key words: questionable research practices, replicability, minimizing false positives.

"Averages and relationships and trends and graphs are not always what they seem. There may be more in them than meets the eye, and there may be a good deal less."

Darrell Huff, "How to Lie with Statistics"

We are honored to comment on Klaas Sijtsma's provocative article, "Playing with Data—Or How to Discourage Questionable Research Practices and Stimulate Researchers to Do Things Right." Sijtsma's article is timely, not only in view of recent concerns in psychology and many other fields regarding the validity and trustworthiness of scientific findings (e.g., Asendorpf et al., 2013; Pashler, & Wagenmakers, 2012), but also because it marks the anniversaries of three classic publications that are integral to the issues he—and we—raise.

The first is the 60th anniversary of Darrell Huff's (1954) delightful little book "How to Lie with Statistics." In the words of the author and befitting the title, "This book is a sort of primer in ways to use statistics to deceive… The crooks already know these tricks; honest men must learn them in self-defense (p. 9)." The second is the 40th anniversary of the publication of Amos Tversky's and Daniel Kahneman's (1974) seminal paper in *Science*, "Judgment under uncertainty: heuristics and biases." In this work, the authors described and provided examples of numerous cognitive biases to which humans fall prey when drawing causal inferences under conditions of uncertainty. As Tversky and Kahneman recognized, these biases, many of which reflect the inappropriate application of otherwise useful heuristics (e.g., representativeness, availability, anchoring and insufficient adjustment), are not relegated to layperson's decision-making in everyday life. Instead, these biases apply in equal measure to scientists' inferences in their research, a point that has become all-too-evident in the recent psychological literature and many other scientific domains (Ioannidis, 2005; Kahneman, 2011). The third is the 25th anniversary of the publication of Peter Sedlmeier's and Gerd Gigerenzer's (1989) paper in *Psychological Bulletin* entitled "Do studies of statistical power have an effect on the power of studies?" This third paper updated Jacob

Correspondence should be made to Irwin D. Waldman and Scott O. Lilienfeld, Department of Psychology, Emory University, 475 PAIS Building, 36 Eagle Row, Atlanta, GA 30322 USA. Email: psyiw@emory.edu and slilien@emory.edu

Published online: 28 March 2015

Cohen's (1962) famous article, which documented the low statistical power (.46 on average) of studies published in the 1960 volume of *Journal of Abnormal and Social Psychology*. Rather dismayingly, average statistical power actually appeared to have decreased slightly in the ensuing 24 years, as it was only .37 in the 1984 volume of *Journal of Abnormal Psychology*. As disparate as these three works may appear, they reflect the two central themes raised by Sijtsma, namely, (a) deliberate researcher malfeasance and (b) inadvertent bias or error in the use of methods and statistics, respectively.

In this commentary, we first address Sijtsma's central issue, the prevention of Questionable Research Practices (QRPs) by means of two policy changes: (a) making research materials publicly available and (b) regular consultation with methodological and statistical consultants. We then use these recommendations as a launching point to discuss several other issues that we believe are important impediments to the construction of a replicable, valid, and trustworthy science in psychology. These issues include (a) bridging the gap between the substantive and methodological/statistical roles of the researcher; (b) reasons for the perpetuation of false-positive results (e.g., excessive reliance on p-values and the extant system of incentives governing scientific advancement); and (c) encouraging replication and its importance for creating a trustworthy corpus of scientific knowledge.

1. Uncovering Questionable Research Practices

Coinciding with concerns in psychology and many other fields regarding the validity and trustworthiness of scientific findings is the rise in documentation of various QRPs over the past decade. As Sijtsma highlights, it is likely that most QRPs are the result of suboptimal methodological and statistical training rather than outright fraud or deception. We concur with Sijtsma that many or most QRP's stem from a pervasive academic "culture" in which problematic analytic practices by investigators are tolerated and even encouraged, and in turn transmitted to subsequent generations of students. One need not be an orthodox Skinnerian to recognize that the reinforcement contingencies of academia frequently favor QRPs, especially those involving "p-hacking": analytic practices that bring test statistics below the hallowed .05 threshold of statistical significance (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014).

QRPs include, but are by no means limited to, selective inclusion and exclusion of participants in analyses to achieve statistical significance (John, Loewenstein, & Prelec, 2012), selective "cherry-picking" of outcomes that attain statistical significance (outcome reporting bias; Ovan et al., 2008), analyzing data in multiple ways (e.g., by performing different data transformations) until one of them reaches statistical significance (Simonsohn, Nelson, & Simmons, 2013); use of repeated significance testing ("peeking" or "snooping") to decide on when to stop collecting data (Simmons, Nelson, & Simonsohn 2011), confusion between exploratory and confirmatory modes of analysis (Cumming, 2014), and HARKing (hypothesizing after results are known; Giner-Sorolla, 2012). These practices, either in isolation or in conjunction, almost certainly contribute to the perpetuation of false-positive findings and resultant replication failures (Ioannidis, 2005).

2. Making Research Materials Publicly Available

We agree with Sijtsma on numerous points. In particular, we concur that (a) for addressing QRPs' prevention is a better long-term strategy than is attempting to infer researchers' intentions, (b) preventive measures are better implemented through institutional reform than through targeting individual researchers and their work, and (c) encouraging researchers to make their data and research materials public is the best institutional antidote against QRPs. As Sijtsma points out,

ferreting out researchers' intentions is typically a feckless task given the difficulties of inferring researchers' motives with any confidence and of distinguishing researcher malfeasance from simple mistakes. Indeed, the research literature on the detection of deception offers a sobering reminder that even the most discerning among us are frequently mistaken when it comes to judgments of who is lying and who is telling the truth (Vrij, Granhag, & Porter, 2010). It will be much more profitable for research domains to implement routine, systematic methods for making public not only data, but all other important aspects of the research enterprise.

Sijtsma discusses some of the impediments—both legitimate and illegitimate—to researchers' sharing of their data and research materials. Disappointingly, although perhaps not surprisingly, researchers who are less apt to share their data also tend to make more statistical errors and to report weaker evidence against the null hypothesis than researchers who are willing to share (Sijtsma, this issue; Wicherts, Bakker, & Molenaar, 2011). More warranted concerns are the considerable effort involved in the preparation of data to share with others and a research team's lost opportunities to analyze and publish data that they worked hard to collect before sharing it with others.

Although these are justifiable concerns, they are far outweighed by the benefits of sharing, and they are solvable, particularly if incentives are put in place to make such sharing worthwhile. First, as Sijtsma points out, there is great value in having independent sets of eyes look over one's data and analyses, not only for the typical researcher and research team, but also for researchers with considerable statistical expertise. Across most fields, one cannot help but be struck by the divergent viewpoints regarding alternative methodological and statistical procedures and what represents "best practice." Also, given the reinforcement contingencies prevalent in academia (e.g., institutional rewards for publishing articles at a high rate) it is far too easy for a researcher or small research team to acquire and maintain problematic methodological habits, including QRPs, that are transmitted to more junior members of the research team, such as graduate and undergraduate students and postdoctoral fellows. "Sunlight is the best disinfectant," to paraphrase former U.S. Supreme Court Justice Louis Brandeis, is a quote that is as relevant here as anywhere. Second, it is certainly possible for researchers to share their data sequentially, starting with data they have already analyzed and published, followed by additional data releases in successive waves as further publications emerge. Third, as Sijtsma describes, another related important incentive for data sharing is the opportunity to contribute to and co-author subsequent publications by other researchers with whom one has shared one's data. Thus, rather than viewing data sharing as a means of being 'scooped,' one can often view it as an unanticipated opportunity to increase one's scientific impact and advance scientific knowledge. Fourth, because such data sharing requires time and effort, it is likely that this practice will not become commonplace until it is adopted as the norm within scientific domains. We agree with Sijtsma that various domains within psychology, as well as in other sciences, should move in this direction, as some fields already have.

Even if one agrees with these points, there are other logistical issues to overcome. What are the best ways to share data, balancing appropriate access with necessary safeguards to participants' privacy? Sijtsma describes several possibilities for this arrangement, including data warehouse facilities that provide secure links to researchers' websites, and consortia whose role is to store, manage, and disseminate data to researchers who have made successful application for their use. One successful example of this last option is the Psychiatric Genomics Consortium (Collins & Sullivan, 2013), which comprises working groups corresponding to several major adult and child psychiatric disorders and has fostered a multitude of diverse research projects. One successful example of data sharing are the advances in understanding the molecular genetics of schizophrenia afforded by the pooling of genome-wide association scan data from multiple samples, leading to considerable statistical power for detecting genetic effects of small magnitude (Ripke et al., 2013). Another issue concerns the best ways for researchers to make public their research hypotheses, designs, and planned analyses prior to conducting the study, and to make explicit the analyses

they plan to conduct prior to executing the study. Sijtsma and others (see also Nosek & Lakens, 2014), mention several registries for the former, and it could eventually become standard policy for such research design considerations to be registered prior to initiating a study. Preregistration also would serve the purpose of researchers staking an interim claim to a set of hypotheses as well as the statistical approaches for investigating them. With regard to the latter point, namely making explicit the analyses they conduct in a study, researchers can save the scripts they create to run their analyses, both those that are focal and those that are supplementary. In turn, researchers could provide these scripts as appendices to their publications (either in the text or supplementary materials) and make them available on their websites (ideally through electronic links in their publications). Researchers could also keep records of all of these study features in a digital lab notebook, akin to those that are a mainstay of research in established laboratory sciences, such as chemistry and molecular biology.

3. Increased Involvement of Methodological and Statistical Consultants

Another point highlighted by Sijtsma with which we largely agree is that methodological and statistical "best practices" would increase and QRPs would decrease if increased involvement of methodological and statistical consultants in psychological research became routine. We concur with Sijtsma that a fundamental problem contributing to QRPs is a lack of statistical knowledge and experience in successful applications of statistical reasoning to real-world data analytic problems, that additional graduate statistical coursework is warranted, and that research projects would greatly benefit by enhanced statistical expertise. Like Sijtsma, we are not referring only or primarily to the use of more sophisticated statistical approaches, such as multi-level modeling, factor mixture modeling, or latent growth curve analyses. Rather, we refer also to a deep understanding of fundamental statistical issues, such as what the null hypothesis does and does not test, the role of small sample size not only in overlooking true effects but also in finding "effects" that do not exist (i.e., false positives; see Button et al., 2013), and the superiority of effect sizes and confidence intervals to p-values in describing the results of research (Cumming, 2014).

At the same time, we are not fully persuaded by Sijtsma's proposed solution, namely to routinely include methodological and statistical consultants in most if not all research projects. Although we are solidly in favor of increased methodological and statistical expertise being applied to psychological research, we have concerns about such expertise being provided by consultants who are largely external to the substantive aims and hypotheses of those projects. In our experience, although the inclusion of such consultants often boosts the statistical sophistication and ideally the accuracy of analyses, this model also comes with unanticipated costs.

First, there is the serious logistical hurdle of not having nearly enough good statistical consultants to go around. Moreover, those who are available tend to be concentrated (at least in the U.S.), in a relatively small number of large universities with substantial research commitments and footprints. We fear that researchers at smaller, less funded, schools, including most colleges and more modest universities, will not have such consultants available at their disposal. Even the statistical consultants at premier research universities tend to be woefully overcommitted and often spread across numerous research projects. It is not uncommon to hear Principal Investigators of grants and research projects lament that although they listed such consultants as key personnel on the grant and are paying a portion of their salary or consulting funds, it is difficult if not impossible to get them to actually work on the project or even to meet with them regarding methodological or statistical issues.

Second, we have frequently witnessed a sharp dichotomy on research projects between the Principal Investigator and Co-Investigators with substantive interests and expertise, on the one hand, and statistical consultant(s) with methodological and statistical expertise, on the other. We

believe that this unfortunate dichotomy has been fueled by the prevailing grant culture—in which the inclusion of statistical experts in addition to researchers with substantive expertise has become *de rigeur*—and can inadvertently contribute to a progressive diminution in researchers' statistical expertise. We have often been surprised by the extent to which Principal Investigators are distant from their data and the results of analyses thereof. Upon asking questions about those data or results, one often hears responses such as "I'll have to check with my 'stats guru' about that" or "Hmmm…I wasn't involved in those analyses." Although the Principal Investigator and Co-Investigators with substantive interests and expertise may not be familiar with all the statistical nuances of the analyses, it is essential that they at least be conversant with the details of their data and the integral features of their statistical results. Sadly, based on our experience, this is often not the case.

Moreover, we are convinced that statistical and methodological knowledge helps investigators to better conceptualize the research questions they are posing, as well as the hypotheses they are advancing to test them. Hence, a heightened reliance on statistical consultants could inadvertently lead researchers to think less clearly and deeply about what their data can, and cannot, tell them. It may also encourage them to think less deeply about their central research questions before posing them.

4. Bridging the Gap Between the Substantive and Methodological/Statistical Roles of the Researcher

Related to the aforementioned points, this dichotomy between the substantive and methodological/statistical aspects of research fosters a separation of roles even in the education of graduate students, most of whom are trained on the substantive aspects of research and only a few of whom are trained to become "stats whizzes" whose future careers will focus on supporting substantive researchers. To reduce QRP's and to increase the quality and trustworthiness of research, it will be necessary to rethink graduate education and even faculty members' relationships with the statistical aspects of their research after they obtain their Ph.D.s.

We believe that there are several ways to begin to rectify the current state of affairs. First and most important, we must increase training for all of our graduate (and undergraduate) students in research methods, statistical reasoning, and advanced statistical methods, so that they all will be able to answer questions about their data and the results of their analyses (see also Asendorpf et al., 2013). Some fields, such as Human Genetics, already have made this transition to good effect. Second, just as many faculty members in applied fields involving clinical or educational practice are required to attend workshops and seminars for Continuing Education (CE) credits, we should require all faculty members involved in scientific and educational activities to similarly attend seminars, workshops, and short courses in various areas of methodology and statistics and to document their involvement through an analogous system of CE credits. Third, there should be a greater investment in research activities that bring together the substantive and statistical aspects of research in given domains.

5. The Perpetuation of False-Positive Results and Barriers to Replication

In addition to the assorted QRP's mentioned by Sijtsma and others (e.g., John et al., 2012), we contend that there are two major contributors to the perpetuation of false-positive results and barriers to successful replication, namely the use of small samples and over-emphasis on p-values. Although these are separable problems, their conjunction conspires to contribute to the proliferation of false-positives and failures to replicate findings.

Prior to discussing the contribution of small sample size and over-emphasis on p-values to false-positives and replication failures, one must answer the deceptively complex question, "What does replication mean?" First, it should be clear that one rarely, if ever, replicates a study *per se*, but rather one may replicate all or some of the findings from it. Second, it is probably best to construe replication as lying on a continuum, rather than being represented dichotomously as a "replicated" versus "did not replicate" binary decision (see Lykken, 1968). Third, related to the second point, researchers have placed far too much emphasis on results that fall below the hallowed, but scientifically arbitrary, $p < .05$ threshold in adjudicating whether results have successfully replicated those of a previous study (Rosnow & Rosenthal, 1989). Fourth, even if one uses p-values as a means of deciding on the outcome of a replication, it is clear that p-values are insufficient on their own, and instead need to be considered in conjunction with effect sizes and their accompanying confidence intervals. Fifth, it is a bit artificial and perhaps misleading to speak in terms of an "original" and a "replication" study, given that each of these are merely instantiations of many such studies (Cumming, 2014; Schmidt, 1992). As such, it is better to construe them merely as two studies sampled from a large population of studies, even if no other such studies have yet been conducted. Sixth, we agree with many other researchers that replication should be routinely assessed via meta-analytic techniques, which allow one to estimate an overall effect size, test whether it differs significantly from the null hypothesis value, characterize the magnitude of its heterogeneity across studies, and ideally posit and test various substantive and methodological moderators of the effect sizes across studies (see for example Schmidt et al., 1992). As such, we operationalize replication as the conjunction of an overall effect size that differs significantly from its null hypothesis value (most often 0) and exhibits minimal heterogeneity across studies, as indicated by minimal I^2 and a non-significant Q-statistic, and the study's estimated effect size lying within the confidence interval established by previous studies.

It is both sobering and surprising that 25 years after the publication of Sedlmeier and Gigerenzer's (1989) article, low statistical power remains a serious concern. Every indication is that the continuing problem of low statistical power in psychological research—and many other scientific domains—remains unsolved. Although there are many reasons why this might be so, two stand out as particularly important. First, as we discuss in further detail in the paragraphs to come, researchers have been so focused on p-values as opposed to effect sizes in summarizing the results of studies that it is rare for researchers to generate reasonable expectations for the magnitude of effects in many research domains, or to even be able to articulate what would constitute a small, medium, or large effect in their domain of study. Although researchers routinely rely on the provisional guidelines set out by Cohen (1969, 1994) to quantify small, medium, or large effect sizes in conducting power analyses, these metrics are arbitrary and completely generic, as Cohen himself acknowledged and warned. Regrettably, Cohen's crucial caveat has been more honored in the breech than in the observance. Although the adaptability of effect sizes to any research domain in many ways represents a strength, it also belies the extent to which researchers are unfamiliar with the typical magnitude of effect sizes in their area. It is inconceivable to us that small, medium, or large effect size magnitudes would be uniform, or even very consistent, across markedly different research domains, such as the social psychology of group influence as opposed to the molecular genetics of schizophrenia and other exceedingly complex phenotypes. Second, the mounting pressures on investigators to publish and obtain grants—and all that follows from these pressures (e.g., status, money, tenure, and success)—inevitably play potent roles in driving under-powered studies, as these forces lead many researchers either to terminate data collection as soon as they obtain a significant p-value for their effect of interest (John et al., 2012; Simmons et al., 2011) or to use the minimum sample size that they (most often incorrectly, see Wicherts, 2014) anticipate will be necessary for achieving statistical significance. These strategies are in contrast to using samples large enough to reliably estimate an effect of interest and to minimize its 95% confidence interval (Cumming, 2014).

It is undeniable that null hypothesis significance testing and *p*-values have had an inordinate effect on the research enterprise and on the body of knowledge within psychological science. It is worth noting, however, that this same over-reliance has characterized many other fields and has certainly exerted a pervasive influence in the biomedical fields (see Ioannidis, 2005). The *p*-value controversy has had a long - and depending on one's perspective - storied or sordid history. Although it traces back to the debate between Fisher (1925) versus Neyman and Pearson (1928), it seems to resurface every 15 years or so (see for example, Gigerenzer, 1998; Harlow, Mulaik & Steiger, 1997; Lykken, 1968; Meehl, 1978; Nickerson, 2000; Oakes, 1986; Rozeboom, 1960) and continues largely unabated to this day. We will not recapitulate the arguments for and against significance testing herein, given that this has been so thoroughly and expertly covered by others. Instead, we limit ourselves to a few general comments and specific points regarding how an over-reliance on *p*-values contributes to false positive results and replication failures.

First, many researchers and students still seem confused about what specifically p-values represent (see also Abelson, 1995; Cohen, 1994), as many still seem to believe that they represent any one or more of the following:

1. The probability of the null hypothesis being true,
2. The probability of the alternative hypothesis being true,
3. The probability of the alternative hypothesis being true, given the observed data,
4. The probability of the null hypothesis being true, given the observed data,
5. The probability of observing the data, given the alternative hypothesis,
6. The inverse of the probability that one's results may replicate,
7. The inverse of the probability that the substantive theory that spawned the hypothesis is true, rather than
8. The probability of observing the data, given the null hypothesis.

As readers of this journal are well aware, only the final description is accurate.

It also seems to us that researchers—novice students and seasoned faculty alike—have more of an essentialist than a stochastic perspective on and interpretation of *p*-values. Rather than interpreting *p*-values as the probabilities of observing the present data under the assumption that the null hypothesis is true, which can fluctuate as circumstances change (e.g., as individuals are added to the sample, as multiple samples are examined, given analyses of the original data versus transformations thereof), researchers seem to believe that the "more significant" a result, the "more true"—and the more reflective of what is actually going on in nature—it must be. Even more, some seem to believe that the more significant the result, the more likely the substantive hypothesis is to be true, which is a variant of the well-known logical error of affirming the consequent.

There are other widely held misconceptions regarding statistical significance and *p*-values that we have heard from students and faculty colleagues alike, some of which we confess we ourselves held until several years ago. Among these are that (a) if one obtains statistically significant results in a study, by definition one must have had adequate power to reject the null hypothesis, (b) the only problem with small sample sizes is insufficient power to reject the null hypothesis (as opposed to statistically significant results in small samples also being more likely to represent false-positives than true results; Button et al., 2013), and (c) false-positives are more likely to emerge from studies with larger sample sizes given the increase in statistical power. With respect to (c), a senior (unnamed) colleague at an (unnamed) university noted that he/she was expressly taught to avoid using large samples for this very reason! Another investigator, upon detecting statistically significant but miniscule correlations (e.g., $r = .03$) in an extremely large sample ($N > 50,000$), randomly selected a much smaller subsample from this larger sample to determine whether the original findings were "really" significant. Often accompanying such misconceptions are what we would regard as faulty statistical lore that students and faculty are taught and come to believe.

In addition to these misconceptions, statistical significance is a poor metric of a research domain's replicability (Cumming, 2014). Indeed, one can view the development of meta-analysis and the shift in emphasis from *p*-values to effect sizes and confidence intervals as expressly designed to overcome this problem (Smith, Glass, & Miller, 1980; Hedges & Olkin, 1985; Hunter & Schmitt, 1990; Rosenthal, 2001). Despite the fact that meta-analysis has existed for over 35 years and that many books, statistical programs, workshops, and short courses have been developed to facilitate student and faculty researchers' learning such methods, one still sees many narrative literature reviews that base their assessments of the replication of findings largely or exclusively on statistical significance. We have encountered many misconceptions regarding meta-analysis, with some of our colleagues' maintaining that meta-analyses cannot be used in their research domain because the sample sizes are small. Although there is probably no single cure for researchers' addiction to null hypothesis statistical testing, we strongly recommend Geoff Cumming's informative and entertaining simulation and animation, *The Dance of the p values* (Cumming, 2014; www.tiny.cc/dancepvals). For those unfamiliar with this demonstration, Cumming illustrates that although effect sizes and their 95% confidence intervals differ across replication studies, the magnitude of *p*-values differs much more dramatically, with the result that it would be very hard to draw any systematic conclusion from them regarding the degree of replication across studies.

We also maintain that the biggest problems follow from the conjunction of small sample size and an over-emphasis on *p*-values, rather than from either in isolation. Although small sample size results in low statistical power (i.e., increased Type II error) and an increased rate of false-positives, if one estimates the overall effect size from a set of 15 or 20 studies with small samples, one should obtain an unbiased estimate that converges on the actual effect size in the real world. Indeed, this is what would occur if research domains developed without the constraints of relying on null hypothesis statistical testing and the necessity of surpassing arbitrary *p*-value thresholds. In contrast, when sample sizes are small and statistical significance is a prerequisite to a study's publication, it is virtually inevitable that the effect sizes of published studies will be inflated—sometimes dramatically so—a phenomenon referred to as "the winner's curse" (Button et al., 2013). Given the inflated effect size estimates from such studies, eventually more and more failures to replicate (as defined by the failure to detect a statistically significant association) will accumulate. Hence, one is likely to observe a "decline effect" (Lehrer, 2010) or "law of initial results" in which the effect sizes reported by studies decrease over time. One can imagine an alternative scenario, however, in which at least the preliminary studies in a domain are performed in large unselected samples representative of the population in which the focus is on accurate estimation of effect sizes and their 95% confidence intervals, without regard to null hypothesis statistical tests and *p*-values. Although we agree with researchers who have advocated this approach (most recently Cumming, 2014), it might be difficult and even undesirable to abandon null hypothesis statistical tests and *p*-values altogether. Nevertheless, as a research domain develops over time, it may be desirable to have null hypothesis statistical tests follow, rather than precede, knowledge of the magnitude of expected effect sizes and their confidence intervals.

6. Encouraging and Incentivizing Replication

Despite all of us learning in our Research Methods classes the critical importance of replication for building a valid and trustworthy corpus of scientific knowledge, replication attempts are few and far between in some psychological domains. Will the recent and manifold focus on the crisis in psychological research and on the prevalence of QRP's fuel a renewed commitment to replication efforts? We certainly hope so. It is useful to consider not only some of the impediments

to researchers conducting replications of others' work but also to having the results of their own research replicated. Perhaps the greatest obstacles to replicating others' results are insufficient specification of the methods used and the results obtained in published and unpublished studies. Many researchers continue to present only *p*-values, rather than effect sizes and standard errors or confidence intervals, to describe the results of their analyses (Fidler, Thomason, Cumming, Finch, & Leeman, 2004).

Indeed, in the meta-analyses that we have conducted to assess the replicability and generalizability of results across studies in a given domain, we have consistently been forced to exclude a sizable proportion of studies due to insufficient presentation of results. This reporting failure is at times reinforced inadvertently by journal editors who encourage authors to present only *p*-values and, compounding the problem, rounding these values off based on conventional cut-offs (i.e., $p < .05$, .01, or .001). Replication attempts will be enhanced by journals and editors requiring more detailed results, as some journals (e.g., *Psychological Science*) have begun to do. Analogously, researchers can increase the probability that others will attempt to replicate their research by providing such information in their publications. Although these suggestions are exceedingly straightforward, our difficulty in including all available studies in the meta-analyses we have conducted is testament to the importance of encouraging—or perhaps imploring - researchers to routinely provide the details of their analyses.

It may be even more useful to propose mechanisms for encouraging and incentivizing replication. Two major possibilities come to mind in this regard. First, replication can be built into both undergraduate and graduate curricula in psychology, as others have suggested. Frank and Saxe (2012) described including the direct replication of published studies as a central focus of an undergraduate research methods course that she and her colleagues teach. Although this didactic approach is time- and labor-intensive, and potentially challenging for instructors and students, it is much more rewarding both for the students and the field than more typical approaches to this class. Nosek and colleagues (2012) suggested that Master's theses should routinely consist of an attempt to replicate and ideally extend previous findings.

Second, replication attempts will be enhanced by according them greater status and by increasing the incentives for researchers to conduct them. One mechanism for accomplishing this may be to create research metrics akin to *h* and others (Hersh, 2005; Harzing, 2011) to quantify the extent to which researchers attempt to replicate others' findings, as well as the extent to which their own work is replicated by others. Granted, it is easier to suggest such indices than to specify exactly how they should be determined and calculated, especially given that (as noted earlier) replication lies on a continuum. Developing algorithms for tracking replication attempts and their outcomes surely will be more difficult and involved than the algorithms that have been developed for tracking how often researchers' work is cited by other researchers. Other issues that will need to be considered are whether replication attempts are direct as opposed to constructive/conceptual (Lykken, 1968; Simons, 2014), how close a researcher's study must be to another researcher's study to constitute a replication attempt, and what it means for a study to be successfully replicated. With regard to this last point, how does one quantify the scenario in which some findings are replicated while others are not, or if only the central finding of a study is replicated but the other findings are not? And how does one determine whether a finding or set of findings is replicated given the difficulties inherent in the concept of replication discussed earlier?

As a final consideration, we are not advocating that "replication metrics" replace the researcher metrics that have become commonplace in many fields, but instead that they complement and augment them. This addition to the armamentarium of metrics would help ensure that replication may come to achieve a status that balances that of novelty or "surprisingness," which has lately come to be regarded as the *sine qua non* of scientific impact in our field (e.g., Gray, & Wegner, 2013). Although we would not want to stifle researchers' efforts at making novel discoveries or at addressing risky scientific questions, we believe (see also Bertamini & Munafò,

2012) that these virtues have become over-emphasized at the expense of replication. Hence, we urge the field to begin to implement institutional mechanisms and incentives to restore a more healthy balance.

From a Kuhnian (1970) standpoint, this balance also ensures that adequate emphasis be accorded to "normal" science, which involves the solving of puzzles and the detection of anomalies. Although normal science is almost always less "sexy" and exciting than is revolutionary science, it is at least equally important, especially in sciences—including some domains of psychology—that are still gaining their footing. Replication lies at the heart of a healthy science, and is ultimately the best metric of the minimization of QRPs and their adverse effects on psychological research.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

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

Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science*, *7*, 67–71.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

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

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Collins, A. L., & Sullivan, P. F. (2013). Genome-wide association studies in psychiatry: What have we learned. *British Journal of Psychiatry*, *202*, 1–4.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, *25*, 7–29.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidenceintervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–126.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 600–604.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*, 562–571.

Gray, K., & Wegner, D. M. (2013). Six guidelines for interesting research. *Perspectives on Psychological Science*, *8*, 549–553.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Harzing, A. W. (2011). *The publish or perish book: A guide to the software*. London: Tarma Software Research.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hersh, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 16569–16572.

Huff, D. (1954). *How to Lie with Statistics*. New York: W.W. Norton & Company.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124.

Ioannidis, J., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*, 235–241.

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

Kahneman, D. (2011). *Thinking: Fast and slow*. New York: Farrar, Straus and Giroux.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Lehrer, J. (2010). The truth wears off. *The New Yorker*, 52–57.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 1151–1159.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biomettika*, *20A*, 175–240, 263–294.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*, 137–141.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.
Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on replicability in psychological science. A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.
Ripke, S., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*, 150–159.
Rosenthal, R. (2001). *Meta-analytic procedures for social research* (2nd ed.). Beverly Hills, CA: Sage Publications.
Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.
Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.
Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, *105*(2), 309–316.
Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80.
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.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-Curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–537.
Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, *11*, 89–121.
Wicherts, J.M. (2014, May 24). The power paradox and the myth of the failed study. Paper presented at the Annual Convention of the Association for Psychological Science, San Francisco, California.
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.