A Vast Graveyard of Undead Theories : Publication Bias and Psychological Science's Aversion to the Null

Christopher J. Ferguson and Moritz Heene

Perspectives on Psychological Science 2012 7: 555
DOI: 10.1177/1745691612459059

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 555–561

DOI: 10.1177/1745691612459059
http://pps.sagepub.com
SAGE

A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null

Christopher J. Ferguson[1] and Moritz Heene[2]
[1]Texas A&M International University and [2]Ludwig Maximilian University Munich, Germany

Abstract

Publication bias remains a controversial issue in psychological science. The tendency of psychological science to avoid publishing null results produces a situation that limits the replicability assumption of science, as replication cannot be meaningful without the potential acknowledgment of failed replications. We argue that the field often constructs arguments to block the publication and interpretation of null results and that null results may be further extinguished through questionable researcher practices. Given that science is dependent on the process of falsification, we argue that these problems reduce psychological science's capability to have a proper mechanism for theory falsification, thus resulting in the promulgation of numerous "undead" theories that are ideologically popular but have little basis in fact.

Keywords

publication bias, falsification, null hypothesis significance testing, meta-analyses, fail-safe number

For psychology to truly adhere to the principles of science, concepts such as the falsifiability of theory and the need for replication of research results are important issues to consider. However, this observation, though reasonable in the abstract, is complicated by a lack of agreed criteria regarding whether a theory has been falsified or even how exactly one might do such a thing (e.g., Trafimow, 2009; Wallach & Wallach, 2010) or how consistent replication efforts must be for a particular data point to be considered replicated or not. Other scholars have noted the comparative rarity of replication research in psychology and have attributed this in large part to the focus of journals on new and innovative research rather than replications (Drotar, 2010). Although we agree with this assessment, we note other potential issues that limit true replication, and thus falsifiability, in the psychological sciences—namely, psychology's aversion to null results and failure to publish them. Replication in the psychological sciences is meaningless unless failed replications are published as enthusiastically as are successful replications. We do not believe this to be the case. We argue that the phenomenon of publication bias remains a significant problem in many subfields of psychology, reducing opportunity for replication through equal publication of successful and failed replications and thus the credibility of the process of psychological science.

The Phenomenon of Publication Bias

Publication bias refers to the tendency for statistically significant findings to be published over nonsignificant findings (Rosenthal, 1979). To some degree, this bias is due to well-understood limitations of null-hypothesis significance testing (NHST; Cohen, 1994; C. J. Ferguson, 2009) in which null results are considered difficult to interpret. Most scholars who have submitted null results for publication are likely to have received the comment from editors or reviewers that null results are difficult to interpret, that they may be Type II error, or that they may even be the result of not trying hard enough to find significant results.

The aversion to null results is a particularly important issue in light of recent discussions of the methodological flexibility problem (LeBel & Peters, 2011; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; see also Shea, 2011). Data management within psychological science sometimes allows for considerable alternatives in hypothesis testing such that scholars, despite acting in good faith, may be prone to selecting analysis strategies that confirm their preexisting hypotheses and disfavor analyses that do not (LeBel & Peters, 2011). This issue was brought to the forefront, in part, by controversy over an article (Bem, 2011) purporting to claim that humans had supernatural abilities to foretell the future or read minds (see LeBel & Peters, 2011 or Wagenmakers et al., 2011 for discussion). When scholars are under pressure to publish (Fanelli, 2010a) and publication bias

Corresponding Author:
Christopher J. Ferguson, Department of Behavioral Sciences, Texas A&M International University, 5201 University Blvd., Laredo, TX 78041
E-mail: CJFerguson1111@aol.com

exists against null findings, we may find a situation in which the replication process is undone by a neglect of unsuccessful replications and the flexibility to convert unsuccessful replications to successful ones by "cleaning" data and rerunning analyses until expected results are achieved or by running simplistic analyses that favor one's own hypotheses (Heene, Lorenzi, & Schonemann, 2010; Schonemann & Heene, 2009; Simmons et al., 2011).

In addition to publication bias at the journal level, publication bias also likely occurs at the individual scholar level, insofar as scholars decline to submit null results for publication either because such results conflict with beloved theoretical models in which they are invested or simply because they believe null results will not be published (Coursol & Wagner, 1986; Greenwald, 1975). Indeed publication bias may be more pernicious at the level of the individual scholar than it is at the journal level.

These phenomena all contribute to publication bias, which exists as an 800-lb gorilla in psychology's living room. Although many publication bias measures exist, they often suffer from problems of low power and, conversely, may overidentify cases of publication bias, particularly when between study heterogeneity in effect sizes is high (Ioannidis & Trikalinos, 2007). One common debate regarding the meaningfulness of significant publication bias findings (a finding that indicates publication bias exists within a field) is that significant publication bias findings might be due to small study effects (differences between methodologies of smaller and larger studies that introduce potential bias but are distinct from publication bias). Small study effects may indeed cause positive bias findings, but we argue here that such small study effects may be equally problematic for meta-analyses and the impetus should be on authors of meta-analyses to explore and explain findings that potentially indicate publication bias rather than assume that small study effects are the root cause and are thus unimportant to the interpretation of meta-analytic results (Sterne, Gravaghan, & Egger, 2000). Despite the arguments regarding small study effects, Levine, Asada, and Carpenter (2009) argue that true publication bias and aversion to the null are the most probable explanations for significant bias findings.

C. J. Ferguson and Brannick (2012) proposed a tandem method for identifying publication bias designed to reduce Type I errors by combining multiple methods into a decision paradigm. By assessing a number of published meta-analyses in top ranked psychology journals, it was found that approximately 25% of meta-analyses were at risk for publication bias. We argue this number may in fact be too low given power problems with publication bias techniques in general. Levine, Asada, and Carpenter (2009) noted that effect size and sample size are negatively correlated in 80% of meta-analyses and suggested that the phenomenon of publication bias may in fact be far more common.

Further, one of the common methods for reducing publication bias, the inclusion of unpublished studies, appears to be highly problematic, at least in practice (C. J. Ferguson & Brannick, 2012). Although it is often taken as an article of faith that including unpublished studies would rather obviously decrease publication bias, there are several problematic issues that reduce our confidence in this belief. First, meta-analytic authors may not always conduct a fully diligent search for unpublished studies. Ferguson and Brannick found that a plurality of meta-analyses including unpublished studies had only a token number of them (most often less than 5% of all included studies).

Second, searches for unpublished studies may themselves be subject to selection bias. Searches for nonindexed studies (those that cannot be found within the PsycINFO, ERIC, Medline, or Dissertation Abstracts databases, for instance) almost certainly are biased toward scholars who are well-represented in a particular area rather than those who are not. Results of C. J. Ferguson and Brannick (2012) find that meta-analytic authors themselves are more than twice as represented in unpublished articles as published, confirming the existence of selection bias. To some degree, selection bias is due to events beyond the meta-analytic authors' control: most unpublished studies are not indexed, some unpublished datasets may be forgotten even by their creators, some scholars may actively suppress null results, and some null results may have been converted to statistically significant results due to questionable researcher practices (LeBel & Peters, 2011; Simmons et al., 2011). However, some of the problems appear to be within the purview of meta-analytic authors, such as the failure to recruit widely from other scholars (particularly from groups ideologically opposed to the meta-analytic scholars themselves) and a tendency to favor one's own studies. Ferguson and Brannick found that inclusion of nonindexed unpublished studies tended to unwittingly increase rather than decrease bias in most cases. This is not to say that the ideal search for unpublished studies is inherently problematic, rather that this ideal search is so rare that, in practice, such efforts may do more harm than good.

Rothstein and Bushman (2012) acknowledged some of these issues but argued that the risk of selection bias in meta-analysis is overestimated by C. J. Ferguson and Brannick (2012) and that meta-analytic scholars should code for "best practices" and study quality. However, one of their own meta-analyses reveals the limited applicability of these approaches. In their recent analysis of video game violence (Anderson et al., 2010), the authors were found to have selectively included unpublished studies of their own, as well as those of close colleagues and collaborators, while failing to solicit unpublished studies from groups with a differing perspective on video game effects, despite sometimes being in contact with those other groups on other matters (C. J. Ferguson & Kilburn, 2010). Similarly, their coding of methodological quality selectively missed the important issue of measurement standardization and quality that is known to spuriously inflate effect size estimates in this field (C. J. Ferguson & Kilburn, 2009). This helps to explain why their meta-analysis does not comport

with the meta-analyses and conclusions of other groups (C. J. Ferguson & Kilburn, 2009; Sherry, 2007). Thus, the considerable problems with Rothstein and Bushman's own meta-analysis highlight the limited applicability of their suggestions. The intent is not to be unduly critical of Rothstein and Bushman, who undoubtedly have spoken in good faith. Rather, we argue that this gulf between the ideal of including unpublished studies in meta-analysis and the actual damage done to replicability in meta-analysis due to the flawed implementation of such practices is more the norm than the exception. Our argument is that it is time to consider how to improve actual practice.

Is Publication Bias an Overstated Problem in Meta-Analysis?

In the previous section, we argued that publication bias not only exists, but is common and is related in some respects to questionable journal practices, as well as methodological flexibility issues within the science. The problem of validity-threatening publication bias in meta-analytic results drew early attention in the development of meta-analytic methods in the 1970s and has since then shown to be a widespread problem (cf. Atkinson, Furlong, & Wampold, 1982; Coursol & Wagner, 1986; Dickersin & Min, 1993; Levine, Asada, & Carpenter, 2009; Scherer, Langenberg, & Von Elm, 2007; Sterling, Rosenbaum, & Weinkam, 1995; Thornton & Lee, 2000). To illustrate the extent of the problem, consider Fanelli (2010b, p. 4) who found that ". . . the odds of reporting a positive result were around five times higher for papers published in Psychology and Psychiatry and Economics and Business than in Space Science." Furthermore, Fritz, Scherndl, and Kühberger (2012) reported a correlation between sample size and effect size from 395 randomly selected psychological studies of $r = -.46$. As effect sizes from studies with small sample sizes are more strongly affected by sampling error and are thus more likely to give more extreme effect size misestimates, this negative correlation implies a strong tendency in psychology to selectively report positive results and overestimated effect sizes from small sample size studies that are typical for psychology (Fritz et al., 2012). Statistically, one would expect a zero correlation between study-sample size and effect size. (A simulation syntax written in R illustrating the latter point can be obtained from Moritz Heene.)

As published studies obviously do not constitute a random sample of all studies conducted, methods were developed to estimate the extent of bias in meta-analyses induced by unpublished studies that do not yield a significant result. Rosenthal's fail-safe number (FSN; Rosenthal, 1979; Rosenthal & Rubin, 1978) represents an early and still widely used attempt to estimate the number of unpublished studies, averaging null results, that are required to bring the meta-analytic mean Z value of effect sizes down to an insignificant level. Usually, the FSN estimate turns out to be high (cf., E. Ferguson & Bibby, 2011; Hsu, 2002; Martins, Ramalho, & Morin, 2010; Prinzie, Stams, Deković, Reijntjes, & Belsky, 2009; Rosenthal & Rubin, 1978; Voyer, 2011), suggesting statistically stable and trustworthy results and thus implicitly indicating that publication bias is, in most instances, an overstated problem. To illustrate this point, let us define N_{filed} as the number of filed studies, k as the number of published studies, and $\bar{Z}_k$ as the mean Z value of the k published studies. According to Rosenthal (1979), for an alpha-level of 5%, the FSN is given as $N_{\text{filed}} = (k/2.706)[k(\bar{Z}_k)^2 - 2.706]$. Now let us assume we assembled 50 studies in a meta-analysis with a mean Z value of 2.0. In this case, the FSN would yield a value of 6,854. Thus, 6,854 studies averaging null results would be needed to bring the mean Z value down to an insignificant level. Even assuming that one would have found a smaller mean Z value of 0.90 (for example), one gets an FSN of 1,313. In both cases, such a large number of unpublished studies would suggest that the file drawer hypothesis (i.e., that the combined results were due to sampling bias) can safely be ruled out.

However, the FSN treats the file drawer of unpublished studies as unbiased by assuming that their average Z value is zero. This wrong assumption appears mostly not to be recognized by researchers who use the FSN to demonstrate the stability of their results. In fact, if only 5% of studies that show Type I errors were published, the mean Z value of the remaining unpublished studies cannot be zero but must be negative. Without making this computational error, the FSN turns out to be a gross overestimate of the number of unpublished studies required to bring the mean Z value of published studies to an insignificant level. The FSN thus gives the meta-analytic researcher a false sense of security.

Although this fundamental flaw had been spotted early (Elsahoff, 1978; Iyengar & Greenhouse, 1988a, 1988b), the number of applications of the FSN has grown exponentially since its publication (Heene, 2010). Ironically, getting critiques of the FSN published was far from an easy task (Schonemann & Scargle, 2008, see also http://www.schonemann.de/pdf/91_FSN_reviews.pdf for the mentioned reviews in this source). So, the question still remains to what extent meta-analytic results are threatened by publication bias. In order to address the question about the stability of meta-analytic results in the presence of publication bias, Scargle (2000) extended the results by Iyengar and Greenhouse (1988a) on an improved version of the FSN. He defined $\text{FSN} \equiv N_{\text{filed studies}}/N_{\text{published studies}}$ and investigated it in relation to the significance level z_0 and the rejection probability $S0$. (Note that, for the sake of clarity, $S0$ represents the probability that a study will be published if its Z value is smaller than the significance level Z_0, hence, if it is insignificant). In contrast to Rosenthal and Rubin (1978) and Rosenthal (1979), he found out that the FSN is large ($\log_{10}\text{FSN} >> 1$) only if the significance level corresponding to z_0 of published studies is ≥ 2 and the publication bias probability $S0$ is close to zero (Scargle, 2000, p. 101, Figure 3; see also Schonemann & Scargle, 2008 for a generalization of Scargle's model). Hence, the true FSN is almost never as large as Rosenthal's FSN. Consequently, "apparently significant, but actually

spurious, results can arise from publication bias, with only a modest number of unpublished studies" (Scargle, 2000, p. 102).

It cannot be emphasized too strongly that the most important point is that this finding goes beyond a mere technical critique of the FSN. Scargle's results demonstrate the instability of meta-analytic results in the presence of even a small publication bias.

Do Meta-Analyses Work Against Replication and Falsifiability?

As argued earlier in this article, attempts to identify publication bias using the FSN have resulted in overconfidence and, in general, psychological science has not yet been efficient in addressing this issue. In essence, what we observe is the considerable gap between how meta-analyses should work and how they actually do work. Including unpublished studies in meta-analyses should reduce publication bias, but in practice, including such studies is imprecise and does not adequately reduce publication bias (the exception potentially being doctoral dissertations, which are indexed and thus less likely to experience selection bias). Meta-analyses should be more objective arbiters of review for a field than are narrative reviews, but we argue that this is not the case in practice. The failures of these ideals, as argued earlier, are due to multiple issues. Meta-analyses are known to suffer from the "junk in/junk out" phenomenon (which is unlikely to be fixed by "best practices" efforts, when scholars may simply value their own junk higher than the junk of others). The selection and interpretation of effect sizes from individual studies requires decisions that may be susceptible to researcher biases.

It is thus not surprising that we have seldom seen a meta-analysis resolve a controversial debate in a field. Typically, the antagonists simply decry the meta-analysis as fundamentally flawed or produce a competing meta-analysis of their own (see Twenge, Konrath, Foster, Campbell, & Bushman, 2008, vs. Trzesniewski, Donnellan, & Robins, 2008; or Gerber & Wheeler, 2009, vs. Baumeister, DeWall, & Vohs, 2009; or perhaps most famously Rind, Tromovitch, & Bauserman, 1998, vs. Dallam et al., 2001, vs. Lilienfeld, 2002). As Greenwald (2012) noted, empirical results such as meta-analyses seen as debate-ending by one side of a controversy are typically viewed as fundamentally flawed by the other. It is not our intent to take a position on any of the debates cited above—rather, we observe that the notion that meta-analyses are arbiters of data-driven debates does not appear to hold true.

But to further the point, meta-analyses may be used in such debates to essentially confound the process of replication and falsification. This may be due in part to the common misuse of meta-analyses and the implication that the "average effect size wins!". Simply summing up individual studies and getting an average effect size that is statistically significant (in our observation, owing to their inherent power, most meta-analyses are statistically significant no matter how trivial the average effect size may be) is not evidence of a replicated finding. In fact, focusing on the average effect size may be used to, in effect, brush the issue of failed replications under the theoretical rug or into the file drawer.

But more pernicious is that the average effect size may be largely meaningless and spurious due to the avoidance of null findings in the published literature. This aversion to the null is arguably one of the most pernicious and unscientific aspects of modern social science.

Where Are All the Null Results?

So from our discussion thus far we argue that, although meta-analysis is certainly of great potential value, the improper use (even in good faith) of meta-analysis can work against the replicability principle in science. Consistent with ideas expressed by Ioannidis (2005), we suggest that publication bias is more likely in fields that are newer, politicized, or ideologically rigid, where small groups of researchers have invested heavily in a particular theoretical model, where pressures to publish exist (Fanelli, 2010a), or where scholars have taken to making extreme claims about the alleged consistency of research in their field (see C. J. Ferguson, San Miguel, Garza, & Jerabeck, 2012, for a discussion of how this occurred in the field of video game violence research with researchers favorably comparing their own research to that on smoking and lung cancer, global warming, or evolution and how this leads to a replication problem in this area). This aversion to the null functions to protect existing theoretical models from any true replication or falsification by explicitly rejecting any efforts to do so as being without value.

The main argument for the rejection of null results is that they are difficult to interpret or may be due to Type II error (i.e., a larger sample might have produced statistically significant results). For example, the first author of this article once received a comment from an editor that a result may have been due to Type II error, despite having a sample size of 150 and an effect size of exactly $r = .00$ (the article was subsequently published elsewhere as C. J. Ferguson, Munoz, Contreras, & Velasquez, 2011). We suspect such comments by editors and reviewers are depressingly common. Such phenomena may encourage scholars to effectively *chase the significant*—that is, to increase their sample sizes until statistical significance is achieved without regard for the triviality of resultant effect sizes (and adding to publication bias in the process). It is perhaps no surprise then that the publishing of null results has actually diminished in the social sciences over time (Fanelli, 2012).

More critically, this argument is misguided. If results that fall below the arbitrary $\alpha = .05$ line are meaningless, than so are results that fall above it, which are as likely due to Type I error (particularly when effect sizes are low) as null effects are due to Type II error. This problem with potentially spurious findings is particularly true when the assumptions of parametric statistics, including random sampling, are not met in the

involved databases. NHST is too often interpreted in a formulaic, unsophisticated manner, particular with respect to the reliability of replicability of statistically significant results (Cummings, 2008; Cummings & Maillardet, 2006). Furthermore, such views neglect the fact that guidelines have been suggested for the interpretation of null effects based either on sophisticated further analyses (Levine, Weber, Park, & Hullett, 2008) or by the careful interpretation of effect sizes rather than statistical significance (C. J. Ferguson, 2009).

This pernicious aversion to the null promotes publication bias and is as much the product of questionable researcher practices (Simmons et al., 2011) as questionable journal practices. Furthermore, this aversion to the null may not merely involve a naive preference for the excitement of statistical significance, but the active effort by scholars in ideologically driven fields to protect theories in which they are heavily invested. But further, if null results are summarily rejected, notions of replication and falsification are mere mockeries of what they should be in a fully functional science. What is the point of replication if all the failed rejections are dismissed out of hand?

The Invincibility of Psychological Theories

In this article thus far, we have argued for the prevalence of publication bias in meta-analysis, which stems from aversion to null results and which ultimately works against true replicability of findings in psychological science, as failed replications are largely ignored. We understand that our comments will upset many psychologists given our observation that psychology has driven itself to believe it is just as good, if not better, than other sciences such as medicine (e.g., Meyer et al., 2001; Rosnow & Rosenthal, 2003, although statistics used to make this conclusion are now known to be flawed, see C. J. Ferguson, 2009). Indeed, we suspect this narrative is a large part of psychology's aversion to the null and its common denial of publication bias problems, given how such issues would inevitably crumble the façade of psychology as a purely objective science. To be fair, we are well aware that there are some meta-analytic authors and other scholars who take these issues very seriously. Similarly, we do not mean to imply that psychological science is unique in experiencing these problems (see, e.g., discussion of replication problems in cancer research by Begley & Ellis, 2012) nor that psychological science is any more uniquely bad than it is uniquely good. But by contrast, we also perceive a counter trend within the field to, in essence, remain in denial of these important issues.

Nonetheless, the aversion to the null and the persistence of publication bias and denial of the same, renders a situation in which psychological theories are virtually unkillable. Instead of rigid adherence to an objective process of replication and falsification, debates within psychology too easily degenerate into ideological snowball fights, the end result of which is to allow poor quality theories to survive indefinitely. Proponents of a theory may, in effect, reverse the burden of proof, insisting that their theory is true unless skeptics can prove it false (a fruitless invitation, as any falsifying data would certainly be rejected as flawed were it even able to pass through the null-aversive peer review process described above).

In the absence of a true process of replication and falsification, it becomes a rather moot point to argue whether individual theories within psychology are falsifiable (Wallach & Wallach, 2010) as, in effect, the entire discipline risks a slide toward the unfalsifiable. This is a systemic discipline-wide problem in the way that theory-disconfirmatory data is managed. In such an environment many theories, particular perhaps those tied to politicized or "hot" topics, are not subjected to rigorous evaluation and, thus, are allowed to survive in a semi-scientific status long past their utility. This is our use of the term *undead theory*, a theory that continues in use, having resisted attempts at falsification, ignored disconfirmatory data, negated failed replications through the dubious use of meta-analysis or having simply maintained itself in a fluid state with shifting implicit assumptions such that falsification is not possible. Fanelli (2010b) found that theory supportive results are far more prevalent in psychology and psychiatry than in the "hard" sciences (91.5% versus 70.2% in the space sciences, for instance). Although it may be true that psychologists are almost always right about their theories, we find it more plausible to suggest that the fluidity and flexibility of social science merely makes it easy for scholars, even those acting in good faith, to appear to be right. We suspect a good number of theories in popular use within psychology likely fit within this category; theories that explain better how scholars wish the world to be than how it actually is.

The only way forth is for psychological science to take seriously the major limitations in how psychological data are handled. Previous efforts have attempted to address these issues (e.g., Wilkinson & Task Force on Statistical Inference, 1999), but although such efforts are obviously well intentioned, there is little evidence that they have made substantial effect on the day-to-day way in which psychologists handle data or answer questions. Psychological science will benefit greatly from increased efforts to improve rigor in meta-analyses and by ending the culture in which null results are aversely treated. Otherwise psychology risks never rising above being little more than opinions with numbers.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review. *Psychological Bulletin*, *136*, 151–173. doi:10.1037/a0018251

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is

there a (statistically) significant relationship? *Journal of Counseling Psychology*, *29*, 189–194. doi:10.1037/0022-0167.29.2.189

Baumeister, R. F., DeWall, C., & Vohs, K. D. (2009). Social rejection, control, numbness, and emotion: How not to be fooled by Gerber and Wheeler (2009). *Perspectives On Psychological Science*, *4*, 489–493. doi:10.1111/j.1745-6924.2009.01159.x

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.

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. doi:10.1037/a0021524

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

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, *17*, 136–137. doi:10.1037/0735-7028.17.2.136

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*, 217–227. doi:10.1037/1082-989X.11.3.217

Dallam, S. J., Gleaves, D. H., Cepeda-Benito, A., Silberg, J. L., Kraemer, H. C., & Spiegel, D. (2001). The effects of child sexual abuse: Comment on Rind, Tromovitch, and Bauserman (1998). *Psychological Bulletin*, *127*, 715–733. doi:10.1037/0033-2909.127.6.715

Dickersin, K., & Min, Y. I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences*, *703*, 135–148. doi:10.1111/j.1749-6632.1993.tb26343.x

Drotar, D. (2010). Editorial: A call for replications of research in pediatric psychology and guidance for authors. *Journal of Pediatric Psychology*, *35*, 801–805. doi:10.1093/jpepsy/jsq049

Elsahoff, J. D. (1978). Commentary. *Behavioral and Brain Sciences*, *1*, 392.

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US States data. *PLoS ONE*, *5*, e10271. doi:10.1371/journal.pone.0010271

Fanelli, D. (2010b). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*, e10068. doi:10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. doi:10.1007/s11192-011-0494-7

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120–128. doi:10.1037/a0024445

Ferguson, C. J., & Kilburn, J. (2009). The public health risks of media violence: A meta-analytic review. *Journal of Pediatrics*, *154*, 759–763.

Ferguson, C. J., & Kilburn, J. (2010). Much ado about nothing: The misestimation and over interpretation of violent video game effects in Eastern and Western nations: Comment on Anderson et al. (2010). *Psychological Bulletin*, *136*, 174–178.

Ferguson, C. J., Munoz, M. E., Contreras, C., & Velasquez, K. (2011). Mirror, mirror on the wall: Peer competition, television influences and body image dissatisfaction. *Journal of Social & Clinical Psychology*, *30*, 458–483.

Ferguson, C. J., San Miguel, C., Garza, A., & Jerabeck, J. (2012). A longitudinal test of video game violence effects on dating violence, aggression and bullying: A 3-year longitudinal study of adolescents. *Journal of Psychiatric Research*, *46*, 141–146.

Ferguson, E., & Bibby, P. A. (2011). Openness to experience and all-cause mortality: A meta-analysis and requivalent from risk ratios and odds ratios. *British Journal of Health Psychology*. doi:10.1111/j.2044-8287.2011.02055.x

Fritz, A., Scherndl, T., & Kühberger, A. (2012, April). *Correlation between effect size and sample size in psychological research: Sources, consequences, and remedies*. Paper presented at the 10th Conference of the Austrian Psychological Society, Graz, Austria.

Gerber, J., & Wheeler, L. (2009). On being rejected: A meta-analysis of experimental research on rejection. *Perspectives on Psychological Science*, *4*, 468–488. doi:10.1111/j.1745-6924.2009.01158.x

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–12.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*, 99–108. doi:10.1177/1745691611434210

Heene, M. (2010). A brief history of the fail safe number in applied research. *Arxiv preprint arXiv:1010.2326*, 1–8. Retrieved from http://arxiv.org/ftp/arxiv/papers/1010/1010.2326.pdf

Heene, M., Lorenzi, P., & Schonemann, P. H., (2010). *How valid is the General Record Examination really?* Retrieved from http://demonstrations.wolfram.com/HowValidIsTheGeneralRecordExaminationReally/

Hsu, L. M. (2002). Fail-safe ns for one- versus two-tailed tests lead to different conclusions about publication bias. *Understanding Statistics*, *1*, 85–100. doi:10.1207/S15328031US0102_02

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

Ioannidis, J. P., & Trikalinos, T. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176*i(8). Retrieved from http://www.cmaj.ca/content/176/8/1091.full

Iyengar, S., & Greenhouse, J. (1988a). Rejoinder. *Statistical Science*, *3*, 133–135. doi:10.1214/ss/1177013019

Iyengar, S., & Greenhouse, J. B. (1988b). Selection models and the file drawer problem. *Statistical Science*, *3*, 109–117. doi:10.1214/ss/1177013012

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

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76*, 286–302. doi:10.1080/03637750903074685

Levine, T. R., Weber, R., Park, H., & Hullett, C. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*, 188–209. doi:10.1111/j.1468-2958.2008.00318.x

Lilienfeld, S. O. (2002). When worlds collide: Social science, politics, and the Rind et al. (1998) child sexual abuse meta-analysis. *American Psychologist*, *57*, 176–188. doi:10.1037/0003-066X.57.3.176

Martins, A., Ramalho, N., & Morin, E. (2010). A comprehensive meta-analysis of the relationship between emotional intelligence and health. *Personality and Individual Differences*, *49*, 554–564. doi:10.1016/j.paid.2010.05.029

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist*, *56*, 128–165. doi:10.1037/0003-066X.56.2.128

Prinzie, P., Stams, G. J. J., Deković, M., Reijntjes, A. H., & Belsky, J. (2009). The relations between parents' Big Five personality factors and parenting: A meta-analytic review. *Journal of Personality and Social Psychology*, *97*, 351-362. doi:10.1037/a0015823

Rind, B., Tromovitch, P., & Bauserman, R. (1998). A meta-analytic examination of assumed properties of child sexual abuse using college samples. *Psychological Bulletin*, *124*, 22–53. doi:10.1037/0033-2909.124.1.22

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. doi:10.1037/0033-2909.86.3.638

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *1*, 377–386. doi:10.1017/S0140525X00075506

Rosnow, R., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, *57*, 221–237.

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods*, *17*, 129–136. doi:10.1037/a0027128

Scargle, J. D. (2000). Publication bias: The "File-Drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*, 91–106. Retrieved from http://www.scientificexploration.org/journal/jse_14_1_scargle.pdf

Scherer, R. W., Langenberg, P., & Von Elm, E. (2007). Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, *2*, 158–162.

Schonemann, P. H., & Heene, M. (2009). Predictive validities: Figures of merit or veils of deception? *Psychology Science Quarterly*, *51*, 195–215. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/2-2009/06_Heene.pdf

Schonemann, P. H., & Scargle, J. D. (2008). A generalized publication bias model. *Chinese Journal of Psychology*, *50*, 21–29. Retrieved from http://www.schonemann.de/pdf/91_Schonemann_Scargle.pdf

Shea, C. (2011). Fraud scandal fuels debate over practices of social psychology: Even legitimate researchers cut corners, some admit. *Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/As-Dutch-Research-Scandal/129746/

Sherry, J. (2007). Violent video games and aggression: Why can't we find links? In R. Preiss, B. Gayle, N. Burrell, M. Allen, & J. Bryant, (Eds.), *Mass media effects research: Advances through meta-analysis* (pp. 231–248). Mahwah, NJ: Erlbaum.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*, 1359–1366.

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. *American Statistician*, *49*, 108–112. doi:10.2307/2684823

Sterne, J., Gravaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*, 1119–1129.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*, 207–216.

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology*, *19*, 501–518. doi:10.1177/0959354309336319

Trzesniewski, K. H., Donnellan, M., & Robins, R. W. (2008). Do today's young people really think they are so extraordinary? An examination of secular trends in narcissism and self-enhancement. *Psychological Science*, *19*, 181–188. doi:10.1111/j.1467-9280.2008.02065.x

Twenge, J. M., Konrath, S., Foster, J. D., Campbell, W., & Bushman, B. J. (2008). Egos inflating over time: A cross-temporal meta-analysis of the Narcissistic Personality Inventory. *Journal of Personality*, *76*, 875–902. doi:10.1111/j.1467-6494.2008.00507.x

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychonomic Bulletin & Review*, *18*, 1–11. doi:10.3758/s13423-010-0042-0

Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. doi:10.1037/a0022790

Wallach, L., & Wallach, M. A. (2010). Some theories are unfalsifiable: A comment on Trafimow. *Theory & Psychology*, *20*, 703–706. doi:10.1177/0959354310373676

Wilkinson & Task Force on Statistical Inference. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.