Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science

John Kitchener Sakaluk, M.A.
Alexander Jonathan Williams, M.A.
Monica Biernat, Ph.D.

Accepted for publication at *Perspectives on Psychological Science* (August 6, 2014)

Abstract

We propose analytic review as a solution to the problem of misreporting statistical results in psychological science. Analytic review requires authors submitting manuscripts for publication to also submit the data file and syntax used during analyses. Regular reviewers or statistical experts then review reported analyses, in order to verify that the analyses reported were actually conducted, and that the statistical values are accurately reported. We begin by describing the problem of misreporting in psychology, and then introduce the basic analytic review process. We then highlight both primary and secondary benefits of adopting analytic review, and describe different permutations of the analytic review system, each with its own strengths and limitations. We conclude by attempting to dispel three anticipated concerns about analytic review, namely: analytic review will increase the workload placed on scholars, analytic review will infringe on the traditional peer-review process, and analytic review will hurt the image of the discipline of psychology. Although implementing analytic review will add one more step to the bureaucratic publication process, we believe it can be implemented in an efficient manner that would greatly assist in decreasing the frequency and impact of misreporting, while also providing secondary benefits in other domains of scientific integrity.

Key words: Data analysis; misreporting; peer-review; scientific accuracy

Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science

The use of quantitative data is the cornerstone underlying a majority of psychological science. When using quantitative data, psychologists must report an array of numerical values including descriptive statistics, parameter estimates, inferential statistics, degrees of freedom, *p*-values, effect sizes, and confidence intervals (American Psychological Association, 2010, p. 32-34). However, as recent incidents in economics (Reinhart & Rogoff, 2010) and cardiology (Le et al., 2011) have plainly demonstrated, the scholarly exercise of reporting statistical results is subject to human error (see Herndon, Ash, & Pollin, 2013, and retraction in *Journal of American College of Cardiology 59* (21), respectively). It should come as no surprise that misreporting occurs within psychology as well (e.g., Office of Research Integrity, 2012,).

Reporting mistakes may occur with respect to the description of analyses conducted (e.g., reporting that an ANOVA was conducted, when in actuality, an ANCOVA was conducted), or with respect to the reporting of values from analyses (e.g., reporting a $p = .12$ as $p < .05$). The nature of reporting errors can be further roughly categorized into three types: (1) Psychologists may make *accidental* reporting mistakes because of typos, or misreading output from analyses; (2) Psychologists may, alternatively, make *innocent* reporting mistakes, without thinking they are being deceptive, because they truly believe particular misreporting practices are common, justifiable, or otherwise legitimate. Also in the *innocent* misreporting category, normal forms of motivated reasoning may cause psychologists to be less thorough when data confirm their hypotheses, and more thorough when they do not; (3) Psychologists may even, on occasion, knowingly make *deliberate* reporting mistakes so as to purposefully misrepresent their data (e.g., Marc Hauser, ORI, 2012). Although these three types of mistakes are not morally equal, they are equivalent in terms of practical outcome: populating the scientific literature with inaccuracies about the psychological phenomena under study.

Recent studies suggest that misreporting of statistical results in psychology (Bakker & Wicherts, 2011; Wicherts, Bakker, & Molenaar, 2011) is disturbingly common. In the present paper, we propose a process—Analytic Review—to ensure the accurate reporting of statistical results in psychological science. We begin by describing the problem of misreporting and its consequences in greater detail. Next, we introduce the core components of Analytic Review, and describe the primary and secondary benefits that Analytic Review would provide to the discipline of psychology. We then elaborate on four possible permutations of Analytic Review that we think are worth consideration for adoption, and highlight the strengths and limitations of each approach. Finally, we attempt to rebut three anticipated concerns that psychologists may share about Analytic Review.

The Problem of Misreporting

A reporting error occurs whenever a psychologist accidently, innocently, or deliberately writes an inaccurate report of what analyses were conducted and/or what values those analyses produced. Inaccurate descriptions of analyses that were conducted are potentially less common than inaccuracy in data values and statistical results (see Levine & Hullet, 2002), but we are not aware of any empirical research documenting the frequency of these descriptive mistakes. The frequency of misreporting statistical values, alternatively, has become a topic of interest for recent meta-scientific investigations, including some within psychology. Herein, we briefly review the literature on this kind of misreporting, and describe some of the major consequences of this problem.

How Common is Misreporting?

Investigations into the frequency of misreporting statistical values have yielded varying estimates of the prevalence of misreporting (e.g., 1.0-2.0% in Fanelli, 2009; 22.0-23.3% in John et al., 2012). Given the use of self-report measures of misreporting in these studies, it is likely that participants underreported their frequency of misreporting, either because misreporting is frowned upon (i.e., if deliberate), and/or because they would not be aware when they misreported (i.e., if accidental or innocent).

Bakker and Wicherts (2011) recently adopted a different approach: sampling articles to study misreporting. In total, the authors reviewed over 4000 statistical results from over 300 empirical articles from both high and low impact psychology journals. Nearly 10% of all statistical results were misreported; 1.5% of these were considered *gross errors*, which "affected the statistical decision on the basis of the nominal significance level of .05" (p. 669). Surprisingly, misreported results were present in more than 50% of all articles reviewed, with 18% containing at least one gross error. The authors' subsequent examination of a random sample of articles from PsycINFO produced similar estimates of misreporting.

In summary, Bakker and Wicherts's (2011) investigation of misreporting suggests that between one-third to one-half of all published articles contained at least one reporting error. Their investigation, however, was limited to reporting of inferential statistics. To the best of our knowledge, the frequency of misreporting other statistical values (e.g., descriptive statistics) has yet to be studied.

What Are the Consequences of Misreporting?

Misreported statistical results mislead readers. The extent to which readers are misled may range from trivial to severe. A misreported mean of 3.36, for example, when the correct value of the mean is 3.63, might not seem a terribly egregious error. Misreported inferential statistics, alternatively, may be more seriously misleading, particularly when tests are misreported as statistically significant when they are not, or vice versa (cf. gross errors, Bakker & Wicherts, 2011). Still, both have the potential to seriously affect the outcome of subsequent meta-analyses. In one examination of misreporting, for example, the estimation of Cohen's *d* was biased by more than 0.10 in 23% to 41% of misreported cases (Bakker & Wicherts, 2011). Taken together with the estimate of frequency of misreporting, between 8% and 21% of published articles contain at least one error that could have a "profound" (p. 669) impact on the outcome of meta-analyses (Bakker & Wicherts, 2011). This is particularly troubling given the renewed emphasis in our field on meta-analytic review (e.g., Cumming, 2014; Funder et al., 2014). Indeed, the extent of effect size bias introduced by misreporting, by some scholarly standards, could mean the difference between considering an effect negligible or meaningful (e.g., Petersen & Hyde, 2010).

Considerations Regarding Accidental and Innocent v. Deliberate Misreporting

Misreporting and questionable research practices, such as *p*-hacking (Simmons et al., 2011), may be distinct methodological concerns, perhaps requiring separate solutions. Nevertheless, one study provides preliminary evidence that the issues of misreporting and questionable research practices may, in fact, be related (Wicherts, Bakker, & Molenaar, 2011). In over 95% of the instances of misreporting the authors encountered, the reported *p*-values were actually smaller than correctly recalculated *p*-values. Even more concerning, authors who were less likely to share data were more likely to have made a reporting error. This association was particularly true of cases where misreporting affected interpretations of statistical significance.

One explanation is that researchers who are more organized with their data are also more generally attentive when reporting results from their analyses (Wicherts et al., 2011). This account suggests that misreporting is mostly accidental or innocent in nature. The alternative and more regrettable possibility is that some researchers may deliberately misreport their findings, and then may subsequently be less willing to share their data because of the questionable nature of their analytical practices (cf. John et al., 2012; Simmons et al., 2011). As Stroebe (2013) described, "[I]t is highly likely that fraudsters would have been particularly unwilling to comply with such a request [data sharing]" (p. 7). Indeed, some research practices (e.g., the use of null hypothesis significance testing) and editorial policies (e.g., valuing novel and significant effects over carefully conducted studies) have come under renewed scrutiny, in part because many think that they incentivize questionable research practices, such as deliberate misreporting (Cumming, 2014; Nelson, Simmons, & Simonsohn, 2012; Nosek, Spies, Motyl, 2012).

The problem of misreporting of statistical values in psychological science requires further empirical study, but given the association between rates of misreporting and honoring dataset requests, (Wicherts et al., 2011), we suspect that misreported results enter the psychological literature through legitimate accidents, innocent practices, and deliberate actions. In the case of the latter mechanism, the link between misreporting and questionable research practices suggests that solutions to the former may help to partially address the latter—an idea we return to later in the paper.

Based on this brief review of misreporting, it seems that it is a problem in psychological science requiring a solution (Bakker & Wicherts, 2011; Wicherts et al., 2011). We are proposing Analytic Review as a new feature of the peer-review process, primarily as a means to addressing the problem of misreporting, which may also have secondary benefits to other methodological problems in the field (e.g., questionable research practices).

Analytic Review

We now turn to describing the core features of our proposed Analytic Review (AR) system for preventing misreported statistical results from becoming published in the psychological literature. We then describe the primary and secondary benefits that we anticipate the field will enjoy from adopting AR.

What Is Analytic Review?

In its simplest form, AR occurs at the journal level, and involves having a reviewer examine a submitted manuscript for the specific purposes of (1) verifying that the analyses reported in-text were conducted, and (2) verifying that the reported values (e.g., descriptive statistics, inferential statistics, degrees of freedom, *p*-values, effect sizes, etc.) are accurate. Although new to psychology, systems like AR are being adopted within other disciplines (see McNutt, 2014). For manuscripts in which an exorbitantly large number of values are reported, a smaller subset of values could be sampled for AR—the matter of how they are sampled is discussed later.

In order to facilitate AR, aspiring authors would need to submit their aggregate data file, as well as the syntax for their analyses, from whatever analysis software they used. Editors would select a reviewer responsible for AR—either a typical reviewer or a statistical expert (this choice is discussed later)—who would execute the syntax on the data file for the analyses under review, and then compare what is reported to what the syntax produced. The reviewer responsible for AR would then note where descriptions of analyses departed from what was actually done (e.g., "The authors reported that they conducted a between-subjects ANOVA (p. 22), but their syntax revealed that they controlled for age, so it would be more appropriate to report that they conducted a between-subjects ANCOVA"), and similarly, where statistical values were misreported (e.g., "For the comparison between the experimental and control conditions, $p = .03$ was reported, but the syntax provided actually produces $p = .045$”). These notes would be provided to the journal editor, who would distribute them to the author along with other reviews, and the AR process would be completed.

In making decisions on whether to publish a given article, the AR would be factored in along with other, traditional reviews. An editor might weigh whether statistical analyses were chronically misreported throughout a paper, or systemically misreported in a way that supported the author(s)’s conclusions; misreporting of key effects might even be cause for editors to send an article back out for review. And of course, the editor could ask for corrections to any misreporting should a revision be invited or acceptance granted. For manuscripts that reach publication, a notation could be added indicating that the article—or a subset of its analyses—had passed AR.

Primary Benefits of Analytic Review

Instituting AR would benefit psychological science in two primary ways. First, and most obviously, AR would decrease the prevalence of misreported statistical results in the psychological literature. As a result, scholars would have greater confidence in the accuracy of individual psychological reports and meta-analyses. Instituting AR would therefore not only help psychologists carry out their mission of conducting and reporting high quality science, but it would also help reestablish the credibility of our discipline with the general public following events that have called its integrity into question (e.g., Carey, 2011).

An additional, albeit less obvious benefit of instituting AR would be facilitating accurate estimation of the frequency of misreporting, broadly defined, as well as specific misreporting practices (e.g., sample means v. *p*-values). This would require that journals create a database of analytic reviews, ideally in de-identified form. Ongoing collection of misreporting data could be valuable for numerous purposes. For example, such data could be used to determine whether instituting particular editorial policies has an effect on rates of misreporting, or on rates of submission and acceptance. Ongoing data collection could also prove useful in domain-specific examinations of misreporting practices; it may be interesting to examine whether rates of misreporting are more endemic to particular sub-fields of psychology or psychological literatures. This would require that journals share information about their AR outcomes, with the goal of creating a normative database that could inform further practices at the journal-level.

Secondary Benefits of Analytic Review

Implementing AR could further benefit our field in a number of other ways: by enforcing greater conformity to reporting requirements, demanding better data archiving practices by researchers, facilitating easier and more frequent data sharing, and serving as a dissuader of questionable research practices.

More conformity to reporting requirements. Journals adopting APA standards for reporting the results of statistical analyses require that articles, whenever possible, include exact *p*-values, confidence intervals, and measures of effect size (American Psychological Association, 2010). Even so, in many cases, researchers do not meet these requirements. In one investigation, for example, fewer than 75% of articles reviewed reported exact *p*-values (Fidler et al., 2005). Further, standardized effect sizes were reported in fewer than 50% of all articles, and confidence intervals in fewer than 20% (see also Cumming et al., 2007).

As there is a movement among some psychology journals to increase the field’s emphasis on effect sizes and confidence intervals (see Cumming, 2014; Funder et al., 2014), it is especially important to enforce the APA requirements. Investigations into the frequency of reporting effect sizes and confidence intervals clearly demonstrate that the status quo is ineffective (Cumming, 2007; Fidler et al., 2005), and thus new solutions to this problem are needed.

In the current peer-review process, reviewers may elect to focus on "big picture" issues, such as determining whether appropriate designs or methods have been used to provide a strong test of the author's hypotheses. In doing so, they may overlook whether a given paper has conformed to reporting requirements, considering the issue more minutiae than substantive concern. Reviewers responsible for AR, in addition to ensuring that analyses are correctly described and accurately reported, could also evaluate whether authors are reporting all relevant values, thereby helping to ensure that journal reporting standards are met.

Data archiving and sharing practices. Because AR requires researchers to submit their data and syntax files for review along with their manuscript, they will need to ensure that these files are clear, organized, and archived in a coherent manner. Variables that are cryptically labeled and incoherent syntax prevent others from replicating analyses, and could hold up AR and, in turn, thereby publication. The increased organization prompted by AR may also increase the likelihood of data sharing, perhaps encouraging scholars to take advantage of tools such as The Open Science Framework (see Miguele et al., 2014; Nosek & Bar-Anan, 2012).

Requests for scholars to share their data are typically unsuccessful, despite authors declaring their Certification of Compliance with APA Ethical Principles, which mandates, "researchers must make their data available to permit other qualified professionals to confirm the analyses and results" (American Psychological Association, 2010, p. 12). In one investigation, for example, data was requested from the authors of 141 papers, and only 25.7% of requested datasets were sent (Wicherts, Borsboom, Kats, & Molenaar, 2006). The authors speculated that the reasons for this disappointing rate were that (1) sharing data and syntax requires considerable effort, and (2) there is little incentive, beyond compliance with the APA's standards, for authors to go through the trouble of preparing datasets and syntax in order to share them with requesters.

Implementing AR will likely improve rates of data sharing because it addresses both of these reasons. Little effort will be required of authors to get their data into shape to share, as this step will have already been taken and indeed the data may have already been posted to a public repository. And because passing AR is a requirement of publication, authors will be sufficiently incentivized to go through the trouble of making their data and syntax sharable. In this way, AR would likely promote the goal of open access to data, for which many others have been advocating (e.g., Crocker & Cooper, 2011; Simonsohn, 2013; Stroebe, 2013)

Questionable research practices and fraud. The relationship between investigators' unwillingness to share data and their beneficial misreporting of statistics may be explained by a lack of care in archiving data (Wicherts et al., 2011). Alternatively, it may be a result of questionable research practices (see John et al., 2012).

The transparency of AR should discourage questionable research practices. Under the status quo, a researcher might "forget" to mention that she controlled for variables when reporting that she/he obtained a particular *p*-value. Not so with AR. If a researcher is required to provide a clearly labeled data file and syntax, she/he will have to include the covariates in her/his syntax. Otherwise, the results from the AR will be incongruent with her reported findings.

Importantly, AR promises a higher degree of certainty that one's analyses will be checked for accuracy than what a general open access to data system would offer (e.g., Crocker & Cooper, 2011). AR also clearly delineates who is responsible for checking the accuracy of reported analyses, whereas an open access to data approach (i.e., without AR) might be vulnerable to diffusion of responsibility; much of these data may never be re-examined.

Combined, these features should translate to a system that is highly dissuasive of many types of questionable research practices.

Of course, AR cannot ascertain whether a psychologist crafted a data set out of thin air. An investigator could use simulations to fabricate data and proceed undetected through AR, as long as the results were reproducible, thus requiring alternative methods of detecting fraud (e.g., Simonsohn, 2013). However, fashioning a complete, convincing dataset in this manner is an effortful process, and may therefore be unappealing for would-be fraudsters (Stroebe, 2013). The AR system would still safeguard against less elaborate means of fraud, such as simply making up results or creating only a partial data set. Stroebe (2013) notes that this latter method may have been the one employed by of Diederik Stapel (Tilburg University, 2011). Even if AR cannot stop the most sophisticated fabulists, halting the less creative would-be fraudsters of the academy is of substantial benefit.

Possible Permutations of Analytic Review

Thus far, we have offered but a general description of how AR should function. Many important logistical decisions about the form AR takes remain. We think many of these choices are best left to the editors of journals where AR is employed, in order to suit the needs of the particular editor, journal, and readership. Below, we describe four major permutations we foresee for AR, and highlight the major benefits or limitations we anticipate for each option.

Pre-Acceptance v. Post-Acceptance Analytic Review

Arguably the most pressing decision to be made about AR is establishing at what stage of the review process it should be conducted. Legitimate arguments could be made for AR at either pre-acceptance or post-acceptance of an article. Conducting AR on every submitted article could add substantial workload to the peer-review process, and for journals with very high rejection rates, much of this could be seen as wasted effort. Further, should a revision be requested, many analyses might be changed or added, thereby undoing previously conducted AR. Additionally, should a paper undergo AR at one journal, but suffer a rejection, it may unnecessarily undergo AR again upon submission to another journal. An author might therefore be subjected to multiple rounds of AR for a single paper, and we suspect the return of incremental gains in reporting accuracy would diminish rapidly from round-to-round.

Conducting AR upon acceptance may therefore appear to be the easy choice. Adopting this strategy would ensure that AR only need be conducted on whatever final analyses the author(s) and editor have negotiated, although the author(s) would be required to submit data and syntax for subsequent revisions of the article in which analyses changed. This approach to AR is therefore likely the most efficient. Even so, this approach entails several limitations. First, AR applied only to manuscripts accepted for publication would mean less ability for journals, societies, or sub-fields to collect data on the broader frequency of misreporting in the discipline.

Second, if only accepted articles undergo AR, then unpublished articles will be more likely to contain reporting errors. The consequence of this differential reporting accuracy is that meta-analyses may be biased by the inclusion of unpublished studies (cf. Rosenthal, 1979). This limitation, however, may be addressed by forthcoming developments in meta-analytic techniques that can accurately estimate effect sizes from published studies alone (i.e., *p*-curve, see Simonsohn, Nelson, & Simmons, in press; Nelson, Simonsohn, & Simmons, in prep). In light of these considerations, we suspect that most journals will opt to conduct AR on accepted articles.

Inclusion of Articles and Analyses in AR

Journals may opt to submit every submitted (or accepted) article to AR, or may opt to submit a subset of articles, or studies within articles, or analyses within studies to AR. How best to select which analyses, studies, or articles should be subject to AR? Randomly selecting analyses or articles for AR would facilitate collecting a representative sample of results; hence, a generalizable estimate of the frequency of misreporting could be attained.

Analyses or articles, alternatively, could be selected on the basis of their theoretical or applied importance. For example, in a multi-study paper (e.g., on the effects of mortality salience, Burke, Martens, & Faucher, 2010, the effectiveness of a particular therapy, Hofmann & Smits, 2007, or gender differences in scholastic achievement, Voyer & Voyer, 2014), AR could focus on the important statistics related to primary statistical comparison of interest, as opposed to reviewing covariate effects of secondary interest. Similarly, papers might be selected for AR on the basis of the level of extraordinary claims made by authors. Under such an approach, Bem's (2011) infamous paper on precognition would likely be a more suitable candidate for AR, compared to a paper examining a less extraordinary claim. Adopting this more purposive approach to selecting analyses or articles for AR, compared to a random selection, would better facilitate addressing the problem of misreporting—whether innocent or deliberate—as authors are most likely to misreport statistics from analyses related to their central claims (Wicherts et al., 2011).

However, the purposeful approach to AR is not without drawbacks. Analyzing only the most important statistics in articles may incentivize the misreporting of findings of secondary, albeit substantive importance. Further, disproportionately targeting extraordinary claims risks biasing publication of intuitive findings. Still, the weaknesses of the purposeful approach may be tolerated by a journal because of its advantages in targeting topics of importance.

"Ordinary" v. "Specialty" Analytic Reviewers

Who should editors seek out to perform the AR? We think most scholars, including graduate students, will be able to carry out AR in most cases. We believe this is possible because many manuscripts report analyses with which many will be familiar (e.g., ANOVA, regression, etc.,), thereby enabling regular reviewers to conduct AR.

In a smaller number of instances, however, the methods of analysis used may be so specialized and/or complicated that a regular reviewer might not be able to carry out AR. A reviewer, for example, might not have access to the specialized software required to re-execute analyses, or alternatively, might simply not know where to find a particular reported value within the hundreds of pages of output that some programs produce. In cases such as these, a "specialty" reviewer may be required to effectively conduct AR.

Our primary concern with the "specialty" AR solution is that scholars with specialized knowledge may be requested to perform AR more frequently than other scholars who are fit to perform AR on more typical research articles. We think that as long as scholars are given the freedom to decline carrying out AR—just as they are free to decline serving as conventional peer reviewers—then the scope of this problem might be limited. However, this may make it more difficult for editors to find reviewers to conduct AR on articles with specialized or complex analyses. Commercializing AR might eliminate this problem altogether.

Commercializing Analytic Review

Perhaps the most unusual permutation of AR that we have considered is the possibility of commercializing the AR process entirely. One option for commercializing AR could involve journals hiring dedicated, statistically knowledgeable professionals to review articles being subject to AR. As one of our reviewers suggested, "Journals are currently motivated to prove their worth, and meanwhile, plenty of stats-savvy PhD students struggle to find work after they graduate." This version of commercialized AR would come with the benefit of essentially standardizing the AR process at a given journal, as the same select few individuals would perform all ARs.

AR, alternatively, could be 'outsourced' to other agencies concerned with scientific integrity. For example, in the future, the mission of The Center for Open Science could be expanded to include facilitating and conducting AR for psychology periodicals. Journals wanting to take advantage of this AR service might subsequently pay into Center for Open Science to facilitate the hiring of full-time AR staff, who would be familiar with both typical and specialized kinds of analyses.

A benefit of a commercialized approach to AR is that most psychologists would not have to be bothered by personally performing AR. Additionally, a small job market would be created for graduates with sufficient statistical acumen to conduct AR, either at specific journals or agencies concerned with scientific integrity. However, the limitations of such an approach are unknown; commercializing an aspect of peer-review is relatively uncharted territory.

Anticipated Concerns about Analytic Review and Rebuttals

Analytic Reviews Will Create an Unjustifiable Amount of Additional Work for Scholars

Perhaps the most likely response to our proposal is that AR will require scholars to spend more time jumping through bureaucratic hoops. We cannot deny that AR will add to the to-do lists of researchers. But none of the requirements for AR are arduous, and they have more benefits than costs. For new submissions, scholars should be able to easily locate their data files, and with data submission becoming a requirement at more journals, this process will become normalized. AR adds only the additional requirement of providing syntax as well. Revisions required after AR may cost the researcher some time, but the benefit is increased accuracy in data reporting. Indeed, we believe that instituting AR could actually *save* time for the field of psychology, by eliminating time wasted on misreported data. In some cases, misreported data could lead researchers to expand on research findings that were misstated to begin with. This pursuit of dead ends does not just stymie researchers' careers; it stymies progress in the field.

Increased workload for reviewers is also a concern, depending on the specific AR system put into place, but we suggest that reviewers will quickly become integrated into AR without much visible change in workload. Indeed, many of the permutations of AR that we have mused about (e.g., post-acceptance AR, sampling articles or analyses for AR, commercializing AR) would help to significantly reduce the burden placed by AR on the academic community. As a discipline, prioritizing quality and depth of articles in publication, as opposed to volume (see Nelson, Simmons, & Simonsohn, 2012), could help to further mitigate concerns about workload.

Analytic Reviews Will Intrude upon the Peer-review System

Some may worry that AR will infringes on many of the responsibilities traditionally reserved for conventional peer-reviewers. But in the system we have in mind, conventional reviewers would continue to comment on manuscripts in the typical manner. Analytic reviewers, on the other hand, would focus on using the provided data and syntax files to verify the conduct and reporting accuracy of the analyses. Currently, conventional reviewers often accept in good faith that analyses were completed and reported correctly; only obvious mistakes might be noticed. By replicating reported analyses, analytic reviewers would be able to catch the mistakes that conventional peer-reviewers typically cannot.

There are also functions that conventional peer-reviewers serve which analytic reviewers *should not*. Analytic reviewers would not be expected to have expertise in or evaluate the theoretical frameworks employed in the manuscript. We need conventional peer-reviewers to assess the quality of the theoretical arguments and conclusions contained in submitted research.

Further, we do not mean to suggest that conventional peer-reviewers should be restricted from commenting on reported analyses. Conventional reviewers should continue to be able to request changes to analyses, both large and small. All we propose analytic reviewers do is confirm the conduct and accuracy of reported analyses. Editors would have discretion over what use to make of the AR, just as they determine the weight to give to conventional reviews.

Analytic Reviews Will Make the Discipline of Psychology Look Bad

It is possible that AR would give the appearance that psychologists cannot be trusted. An alternative framing of AR is that it signals that psychologists care about scientific integrity. Just as theory, method, interpretation, and level of contribution are critiqued and vetted in the conventional review process, the accuracy of reporting would now receive focused—rather than occasional—attention. Recommendations to share data have long been a part of APA guidelines, and efforts to encourage depositing of data are on the rise. Like data-sharing practices, AR may become normalized, accepted as a standard, and even helpful, feature of peer review.

A related concern, that the field's image may be tarnished if the frequency of misreporting is reported publically, rests on the assumptions that many errors would be detected, and that data regarding errors would be publicized. The former may be true given recent investigations (Bakker & Wicherts, 2011; Wicherts et al., 2011). But this ignores the possibility that AR will serve as a deterrent. In advance of AR, researchers may be more likely to re-check their reporting, leading to fewer errors at the point of submission. The second assumption may or may not be the case. It is rare for information about most journal review outcomes to be made public (though see Petty, Fleming, & Fabrigar, 1999). But even if high levels of misreporting rates did make news, we suggest that the value of this information is greater than any reputational cost. As a field, we could gain insight into patterns that could inform practices ranging from student training to editorial policy. The public may even applaud psychologists for deeper scrutiny of their scholarly activities.

Conclusion

We recognize that our proposal may be met with a lack of enthusiasm; psychologists do not often support changes to the publishing process (Fuchs, Jenny, & Fiedler, 2012). However, our idea is not radical: all we are suggesting is that before research is published, a third-party must first replicate reported analyses and verify that values from these analyses were reported accurately. Indeed, it may be relatively easy to implement AR as more journals begin to require scholars to deposit their data and analysis syntax (cf. Crocker & Cooper, 2011).

AR, as we have described it, may not go far enough for others. Indeed, some of our reviewers hoped for a version of AR in which statistical experts vetted the appropriateness of employed data analysis techniques. At this early stage of visioning, however, we have kept the focus of AR on addressing the particular issue of misreporting.

Regardless of how AR is enacted, it should not be terribly onerous to meet its requirements. And while we acknowledge that the workload for psychologists may increase, we think that it can be kept manageable, and that the payoff—more accuracy in psychological science—is well worth the adjustments it would take to implement AR.

References

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

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678. doi: 10.3758/s13428-011-0089-5

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

Burke, B. L. Martens, A., & Faucher, E. H. (2010). Two decades of terror management theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review*, *14*, 155-195. doi: 10.1177/1088868309352321

Carey, B. (November, 2011). *Fraud case seen as a red flag for psychology research*. Retrieved from http://www.nytimes.com.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science*, *334*, 1182. doi: 10.1126/science.1216775

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29. doi: 10.1177/0956797613504966

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleing, A., … Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230-232. doi: 10.1111/j.1467-9280.2007.01881.x

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*, e5738. doi: 10.1371/journal.pone.0005738

Fidler, F., Cumming G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., … Schmitt, R. (2005). Toward improving statistical reporting in the journal of consulting and clinical psychology. *Journal of Consulting and Clinical Psychology*, *73*, 136-143. doi: 10.1037/0022-006X.73.1.136

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, *7*, 639-642. doi: 10.1177/1745691612459521

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. doi: 10.1177/1088868313507536

Herndon, T., Ash, M., & Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, *38*, 1-23.

Hofmann, S. G., & Smits, J. A. J. (2008). Cognitive-behavioral therapy for adult anxiety disorders: Meta-analysis of randomized placebo-controlled trials. *Journal of Clinical Psychiatry*, *69*, 621-632. doi: 10.4088/JCP.v69n0415.

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. doi: 10.1177/0956797611430953

Le, R. J., Fenstad, E. R., Maradit-Kremers, H., McCully, R. B., Frantz, R. P., McGoon, M. D., & Kane, G. C. (2011). Syncope in adults with pulmonary arterial hypertension. *Journal of American College of Cardiology*, *59*, 863-867. doi: 10.1016/j.jacc.2011.04.026

Levine, T. R., & Hullet, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*, 612-625. doi: 10.1111/j.1468-2958.2002.tb00828.x

McNutt, M. (2014). Reproducibility. *Science*, *343*, 229. doi: 10.1126/science.1250475

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., …Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*, 30-31. doi: 10.1126/science.1245317

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

Nelson, L. D., Simonsohn, U., & Simmons, J. P. (in prep). P-curve fixes publication bias: Obtaining unbiased effect size estimates from published studies alone. Available at SSRN: http://ssrn.com/abstract=2377290

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

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. doi: 10.1177/1745691612459058

Office of Research Integrity (2012). *Case summary: Hauser, Marc.* Retrieved from http://ori.hhs.gov.

Petersen, J. L., & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993-2007. *Psychological Bulletin*, *136*, 21-38. doi: 10.1037/a0017504

Petty, R. E., Fleming, M. A., & Fabrigar, L. R. (1999). The review process at PSPB: Correlates of interreviewier agreement and manuscript acceptance. *Personality and Social Psychology Bulletin*, *25*, 188-203. doi: 10.1177/0146167299025002005

Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review: Papers and Proceedings*, *100*, 573-578. doi: 10.1257/aer.100.2.573

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

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. doi: 10.1177/0956797611417632

Simonsohn, U. (2013). Just post it: The lesson learned from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875-1888. doi: 10.1177/0956797613480366

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (in press). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*.

Stroebe, W. (2013). Scientific fraud: Lessons to be learnt. *European Bulletin of Social Psychology*, *25*, 5-12.

Tilburg University. (2011). *Interim report regarding the breach of scientific integrity committed by Prof. D. A. Stapel*. Retrieved from http://www.tilburguniversity.edu

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*, 1174-1204. doi: 10.1037/a0036620

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of 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*, 726-728. doi: 10.1037/0003-066X.61.7.726