Peer Review and Publication of Negative Results: Encouraging Rigor, Reproducibility, and the Integrity of the Scientific Record

Rochelle E. Tractenberg, PhD, MPH, PhD, PStat®, FASA
Collaborative for Research on Outcomes and –Metrics, Washington, D.C.
Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine; Georgetown University Medical Center, Washington, D.C.

Key Words: Negative results; peer review; reporting guidelines; reproducibility; publication bias; ASA Ethical Guidelines for Statistical Practice.

Abstract
Reproducibility and *rigor* are currently surging as features to be assessed in publications and grant proposals submitted for peer review. Publication bias, resulting from the submission or acceptance of submissions that report "statistically significant" results - and no other reports - *limits* both rigor and reproducibility. Negative results are essential contributions to the research record in every scientific discipline, but not if they are badly done, incompletely reported, or otherwise ineffectively reviewed. Effective reviewing of negative results can be complicated by the inaccurate perception that studies presenting "positive" results are actually **correct**; since many studies presenting "positive" results may or do turn out to be irreproducible, actually-negative results are already a large part of the research record. There are additional impediments to the promotion of effective reviewing of all research that includes analyses, whether qualitative, quantitative, or mixed methods. Statisticians and data scientists are uniquely prepared to provide such evaluations, thereby positively influencing the integrity of the scientific record. The Ethical Guidelines for Statistical Practice urge "the ethical statistician" to apply these principles to what the statistician produces as well as what they review. This paper outlines definitions of "negative results" and highlights key features of the planning, execution, analysis, and write up of studies that should be considered in the writing of informative reviews of submitted manuscripts that describe negative (or positive) results. All quantitative scientists are encouraged to review actively and informatively, and to provide specific input and advice to journal editors on the viability of *all* research results. By carefully reviewing all submissions, including those with "negative" results, statisticians and data scientists can fulfill their ethical and professional obligations while advocating for the integrity of the scientific record.

1. Introduction

Reproducibility and *rigor* are currently surging as features to be assessed in publications and grant proposals submitted for peer review. Wikipedia defines reproducibility as: "the ability of an entire analysis of an experiment or study to be duplicated, either by the same researcher or by someone else working independently." This consensus-based definition continues, "in science, a very well reproduced result is one that can be confirmed using as many different experimental setups as possible and as many lines of evidence as possible." The National Institutes of Health (NIH) in the United States defines rigor as: "the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results." and, "In published papers, full transparency in reporting experimental details is crucial for others to assess, reproduce, and extend the findings."

In fact, it has long been acknowledged, if not argued, that it is *unethical* for investigators and scientists to publish (or submit for publication) work that is poorly done/incorrectly and incompletely reported, and otherwise neither reproducible nor rigorous (see, e.g., Altman, 1998; Fidler et al. 2004; Finch et al. 2001). What is current about this "surge" of interest in assessing reproducibility and rigor is the uptick in number as well as breadth of efforts being expended to more formally promote these attributes in research that is published, ranging from the active (e.g., Moher & Altman, 2015; Collins & Tabak, 2014; McNutt 2014) to the provocative (e.g., Ioannidis, 2005; Hayden 2013).

A wide range of efforts are underway internationally to improve the actual science that underpins manuscripts being submitted for publication, as well as to create standards for authors to use in their reporting (see, e.g., Lang 2010; see Aczel et al. 2017 for an alternative perspective on reporting standards). This includes a robust conversation about the definitions of "reproducibility" and "replicability" (e.g., Pusztai et al, 2013; Kennet & Schmueli, 2015; Begley & Ioannidis, 2015; Flier 2017) and how to integrate such ideas into how we train those who will use statistics in their research (but who are not themselves statisticians, e.g., Henson et al. 2010; Casadevall et al. 2016; Curran-Everitt, 2016; Szucs & Ioannidis, 2017).

Instead, this article focuses on *reviewing*: specifically, on the specific contributions that should be made by statisticians, data scientists, and others who have the unique capabilities to competently and coherently review and evaluate submissions for publication as peer reviewers. Many critics of the lack of reproducibility of research would agree that it is the misuse (or misunderstanding) of statistics, and not necessarily that science (or statistics) itself "doesn't work", at the heart of what some consider a "reproducibility crisis"; thus, reviewing of the submissions for publication must be lacking at least some attention to the use or reporting of statistics (probably to both).

Although insufficient statistical literacy among reviewers is problematic generally (see Tractenberg 2017), in this article particular emphasis is placed on providing informative reviews of "negative results". However, the recommendations in this article are intended to be relevant for *all* peer reviewing (manuscripts as well as grant proposals). The emphasis on reviewing by statisticians, data scientists, and other members of the statistics community and profession is warranted for many reasons; most journals have difficulty securing competent statistical evaluations of manuscripts sent out for peer review and some grant agencies specifically direct statistical reviewers to *only* read and review the components of a proposal labeled "statistical methods". In their scoping review of what

characteristics are required by journal editors, those who solicit and summarize the reviews of the manuscripts that are later published (or rejected), Galipeau et al. (2016) report that, of 203 unique statements (across 225 published documents relating to what journal editor competencies might be) expressing or describing the knowledge, skills, behaviors, or tasks that the editors are asked to have/carry out, only *three* related to "ensure thorough statistical review". By contrast, 20 references were made to being able to manage one's time effectively (as a scientific journal editor). In September 2017, a consensus statement outlining the core competencies for scientific editors for a biomedical journal (Moher et al. 2017) included one item that should be mentioned here, "4.2 Assess the appropriateness of the research design and methods described in research manuscripts, as well as the validity of findings and conclusions, in relation to the stated research question." (p. 6 of 10). This competency is characterized under the general section of "Editorial principles and processes" – which includes the item "Identify and use trustworthy resources" but does not discus whether or how reviewers should be encouraged to follow item 4.2 themselves. While it is essential for journal editors to exhibit this ability, given the enormous volume of manuscripts that are sent out for peer review, even at the most selective journals, it is impossible for the editor to assess the appropriateness of every single paper that is accepted for publication.

Whereas the appropriate use and reporting of the statistical methodology should be considered an essential feature of any written report or research proposal, experimental and study design are truly crucial *foundations* for reproducible and rigorous science. If a publication or proposal contains incorrect data analysis methods, or they are not reported fully or appropriately, these deficiencies *could* be remediated in revisions. However, if the study design itself is flawed, then no amount of revising can make the write-up publishable. If statistical reviewers are directed *not to read the entire document* (as has happened to the author on multiple grant review panels), these foundational flaws are unlikely to be detected. If the work proposed in such a grant is then carried out and/or a manuscript is submitted for peer review, without competent statistical reviewing, an article that contains un-reproducible results is likely to enter the scientific record. Statisticians might agree on the paramount importance of statistics/statistical or analytic considerations from the very inception of an experiment or research project; the purpose of this article, however, is to encourage expert statistical review for the purposes of advising journal editors (and authors) about what should (and should not) become part of the scientific literature.

2. Ethical obligations around reviewing (and writing) research

The Ethical Guidelines for Statistical Practice of the American Statistical Association (ASA, 2016), revised in 2016, offer specific guidance on how the ethical statistician (and data scientist) can promote reproducible and rigorous work. From the preamble to these Guidelines: "Because society depends on informed judgments supported by statistical methods, all practitioners of statistics, regardless of training and occupation or job title, have an obligation to work in a professional, competent, and ethical manner and to discourage any type of professional and scientific misconduct." As argued by Altman (1998) and more recently explicated by Nicholls et al. (2016), the complete/correct reporting of research results is an ethical obligation, so that failing in this must be considered a form of misconduct. The growing recognition that publishing weak science is a waste of resources, and only endangers future science (e.g., Ioannidis, 2017) underscores the conclusion that inappropriate, incomplete, or otherwise ineffectual reviewing of such work – which might otherwise prevent its eventual publication – is a

contributor to this waste and endangerment, and can therefore also be labeled "scientific misconduct".

The American Statistical Association (ASA) Ethical Guidelines (http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx) comprise 8 core principles, which entail 49 specific elements (See Appendix):

A. Professional Integrity & Accountability (6)
B. Integrity of data and methods (10)
C. Responsibilities to Science/Public/Funder/Client (5)
D. Responsibilities to Research Subjects (6)
E. Responsibilities to Research Team Colleagues (4)
F. Responsibilities to Other Statisticians or Statistics Practitioners (5)
G. Responsibilities Regarding Allegations of Misconduct (6)
H. Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners (7)

The Guidelines are not limited in their scope to the *analysis* that "the ethical statistician" is actually doing themselves; these Principles also pertain to the provision of competent reviewing. Specifically, Principles A, B, and C are directive:

A. The ethical statistician uses methodology and data that are relevant and appropriate, without favoritism or prejudice, and in a manner intended to produce valid, interpretable, and reproducible results. (Professional Integrity and Accountability)
B. The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may impact the integrity or reliability of the statistical analysis. (Integrity of Data and Methods)
C. The ethical statistician supports valid inferences, transparency, and good science in general, keeping the interests of the public, funder, client, or customer in mind (as well as professional colleagues, patients, the public, and the scientific community) (Responsibilities to Science/Public/Funder/Client)

The ethical scientist must report their work correctly and completely; similarly, the ethical statistician has an obligation to ensure that whatever they are reviewing, as well as the work they execute themselves, meets these Principles. Principles D and F are also slightly less directly informative about the obligations to review fully; Principle D ("Responsibilities to Research Subjects") points to the obligation to utilize, and not waste, the effort and contributions of all research subjects, while Principle F ("Responsibilities to Other Statisticians or Statistics Practitioners") may be interpreted as particularly relevant in supporting best practices by all those who analyze –and report on- data. Hirst & Altman (2012) also discussed the obligations of reviewers to follow best practices in their *reviews*. The majority of emphasis over the past five years especially has been on creating journal guidelines or requirements for *authors* to meet (see Lang, 2010), with possibly insufficient attention to how or whether *reviewers* are similarly charged with ensuring – or at least assessing whether or not – those reporting guidelines are met.

3. Statistical expertise is essential for competent peer review

Altman (1998) discussed a wide variety of challenges for the statistical reviewer; at that time he also commented on the increasing use of statistics in biomedical research reports. Since that time, both the number of articles requiring review and the statistical and methodological complexities have increased. Unfortunately, the level of statistical literacy of reviewers of these articles has not increased. Tractenberg (2017) describes this problem as a mismatch between the level of statistical abilities most scientists achieve (through sometimes no required courses in statistics, sometimes a single course), and the level that is required for competent reviewing of argumentation that involves data or statistical methodology. Most "introductory statistics" courses, whether for undergraduate or graduate students, require *remembering*, *describing*, or *applying* the "one correct method" that is the current topic of the course at that point. These represent low cognitive complexity according to Bloom's taxonomy (Bloom et al. 1956). To review scientific papers (and grant proposals), however, the highest level of cognitive complexity in Bloom's taxonomy, *evaluation and judgment*, are required. The next highest level of complexity, *synthesis* and *creation*, are also required to identify alternative analytic options that might have led to different results; recommending that tests of these plausible alternatives be incorporated in revisions (as sensitivity analyses, for example) is therefore extremely unlikely for someone who has not been trained to reason with data and statistics at these levels. Cumming et al. (2007) discuss the use of error bars in the graphics that are included in experimental biology reports. This is a relatively low-Bloom's complexity construct ("document the variability in your sample"), but Cumming et al note that their article explaining this is necessary because these are incorrectly or infrequently included even though they are important descriptive features of the summarization of results in this domain. Papers that are published without, or with incorrect versions of, this information reflect an absence of even this low-level Bloom's cognitive activity (recognize when a graphic is missing an essential piece of information for your own comprehension) *in the reviewers* of these journals. The statistical literacy of these reviewers is insufficient for their task.

Casadevall et al. (2016) describe six consensus-based recommendations to improve the quality of research in the biological sciences; item 1 on this list is a call for "rigorous and comprehensive evaluation criteria to recognize and reward high-quality scientific research", while item 2 focuses on ensuring that biological scientists are appropriately trained in experimental design, analysis, and interpretation. In their 2010 article, Henson et al. (2010) articulated the essential nature of statistics in education research, calling for a "collective quantitative proficiency" and encouraging faculty to create and strengthen a culture, at least throughout doctoral education, within which statistical methods were prioritized in the design of research as well as in the consideration of this research and its impact. The point here is that it is essential to achieve these recommendations across the sciences, but also that these very training outcomes must also be *integrated*, bringing appropriate levels of knowledge of experimental design, analysis, and interpretation, to the rigorous and comprehensive evaluation of all research – i.e., competent statistical review.

Weissgerber et al. (2016) correctly articulate that—and the myriad empirical arguments why—basic scientists *need* training in statistics. This argument is also not new – that scientists who are publishing may not be capable of self-reviewing their statistical and/or methodological arguments (see e.g., Vaux 2012; Hayden, 2013; Vaux 2014; Baker, 2016; see also Tractenberg 2017; Kruger & Dunning, 1999; Kim et al. 2015). Thus, the challenges outlined by Altman in 1998 are almost certainly both still present and probably much greater now (nearly 20 years later). Altman (1998) and many other

authors, consortia, and journals have published reporting guidelines specifically to outline what must be included in reports of different types (see Moher et al. (2011) for a then-complete list of 81 reporting guidelines). Recent contributions to this collection include the Statistical Analysis and Methods in the Published Literature (SAMPL) Guidelines (Lang & Altman, 2013) and a book with specific requirements for a variety of types of analytic methods for reviewers (Hancock & Mueller, 2010) which is currently (August 2017) in revision for its second edition –with new chapters (i.e., methodologies) being added (Hancock, personal communication 30 July 2017). While guidance for the *reporting* of study statistics/results/outcomes has proliferated, only 19 of 116 recently surveyed journals in the domain of "health research" specifically directed peer reviewers to these guidelines (Hirst & Altman, 2012). Incredibly, only 41/116 of these journals provided online instructions for peer reviewers at all. The 19 of 116 as identified by Hirst & Altman (2012) directed reviewers to use the journal's guidelines on reporting for, or as part of, their reviews. Again, the emphasis has been on articulating requirements that authors are not apparently equipped to meet – while virtually ignoring the crucial role reviewers play in the determination of what should become part of the scientific record, specifically, their ability to provide, or suggest the need for, competent statistical review.

4. Negative results

Publication bias, resulting from the submission or acceptance of submissions that report "statistically significant" results - and no other reports – (see, e.g., Egger et al. 1997) is yet another limiting factor on both rigor and reproducibility of work being published as part of the scientific literature. The most common type of "negative result" is a failure to find statistically significant results. However, there are many other types of "negative" results, for example, when expected outcomes are not observed (i.e., failures to replicate); when weaknesses in a study design are identified by a pilot study or other preliminary work; or when results critique a dominant paradigm or theory. All of these can be considered "negative", and that may impact willingness by editors to propel a manuscript on to peer review; when such papers do go out for peer review, the perception that they are "negative" may lower or limit the attention and scrutiny that the report might be given. Since so little of currently-published (i.e., reporting "significant" results) research ends up being reproducible (e.g., Ioannidis, 2005; Nature, 2016), the main difference between manuscripts reporting "positive" and "negative" results may ultimately prove to be *just the narrative*. This article argues that this fact should lead to more consistency (i.e., at a high level of scrutiny and competence) in reviewing and ultimately, publication. Casadevall et al. (2016) also report that one of the six recommendations to improve the research in biological sciences is to "encourage scientific journals to publish negative data that meet methodologic standards of quality" (p. 1).

The importance of negative results (of all types) for the scientific record is constrained by an emphasis on "innovation"; but innovation for its own sake does not serve science. The philosophical perspective of Positivism encourages *as many lines of evidence as possible* to support findings or theory; in this perspective, the more diverse the methods used to test a theory or result, the stronger the evidence base for that theory or result. "So long as theory withstands detailed and severe tests and is not superseded by another theory in the course of scientific progress, we may say that it has 'proved its mettle'… " (Popper, 1959; p 10). Thus, ideally each result would be contextualized within an array of diverse – rigorous – tests of the plausibility, and strength, of the result. If this cannot occur during the experimental design (e.g., with a series of experiments as is typical in many bench sciences), which is true for most biomedical, epidemiological, and clinical studies, then it

should be apparent in the literature. “What is not controversial is that theories, models, and hypotheses need to be probed to assess their correctness.” (Lewin-Koh et al. 2004; p 4) When “innovation”, rather than confirmation and replication (reproducibility). is a journal’s emphasis, Positivism ceases to characterize the scientific endeavor that journal represents. The lack of negative results in the literature reflects the absence of Positivism, rather than “the mettle” of the results or theories that the literature represents. Therefore, although this article focuses on the importance and features of the informative *review* for negative results, statisticians are also encouraged to advocate for the write up and submission of reports of negative results for peer review and publication. Briefly, when the planned study and analyses yield negative results, it is possible to design sensitivity analyses and simulations that can provide a Positivist, rigorous (and reproducible), demonstration of these negative results. The statistician on a research team can advocate, perhaps better than anyone else on the team, for the write up and submission of such a report. If more and more statistically-literate reviewers will review such work carefully, then Positivism can be infused into the literature.

4.1 Defining negative results

There are two main types of study designs that (following a Positivist approach) are *expected* to yield negative results: those that are specifically designed to determine whether (i.e., test the hypothesis that) two methods/processes cannot be distinguished; and those that are designed to test “the mettle” of a hypothesis or theory when a plausible alternative outcome (i.e., not just “fail to reject the null hypothesis”) is also articulated. When results simply do not support rejecting the null hypothesis (which is the typical sense of “negative results”), the most that can be said is that the results are not informative about that hypothesis. Experiments or clinical trials that find “no significant difference(s)” between groups, when not specifically designed to do so (i.e., unless they are equivalence or non-inferiority trials) cannot be characterized as showing equivalence, although if they are methodologically rigorous, they may be characterized as having “negative results”. These are cases where expected outcomes are not observed (i.e., failures to replicate); consideration of how expected outcomes might not have been replicated should be part of the report (see, e.g., Greenland et al. 2016). Other results that could be considered “negative” are informative pilot studies – i.e., if weaknesses are identified in a pilot study or other preliminary work. These are important for the literature because they correctly represent the challenges of the scientific endeavor. Their absence in the literature may give new authors the mistaken impression that science should never result in the identification of problems in methodology or design (see e.g., Grinnell 2009); further contributing to an unwillingness to write up and submit such results for peer review themselves. Finally, when results critique or conflict with a dominant paradigm or theory, investigators at all stages may not believe they will receive a “fair” review, so they may simply not seek to publish such results. All of these actual negative results are important contributions to the scientific record and should be encouraged.

By contrast, some authors may want to characterize failures to reject a null (e.g., if $p>0.05$ or if a confidence interval includes the null value) as evidence of “equivalence”, and the scientific reviewer who is not statistically literate, or one who may be more statistically literate but who is not experienced with this type of study design, may fail to recognize this mischaracterization. Non-inferiority or equivalence studies are specifically designed and powered to *demonstrate non-inferiority*; therefore, a specific interval of “equivalence” or non-inferiority, and not “statistical significance was set as alpha (or p) <0.05” must be articulated *a priori*. The conclusion of non-inferiority or equivalence from an appropriately designed study is not actually a negative result; but the

mischaracterization of a failure to reject the null ($p>0.05$) from any other type of study should *always* be identified in review. As noted, the statistician/most statistical member of a research team can take the failure to reject the null and pursue post-hoc, even exploratory and simulation-driven, analyses that explore "the mettle" of the new hypothesis – that the negative result is a true failure to replicate whatever effect was expected.

4.2 Ethical obligations to review negative results

Statisticians and data scientists can augment the rigor and reproducibility of these negative results by adding sensitivity analyses, but they can also plan the most rigorous test by ensuring that sample sizes are appropriate, that there is sufficient power to detect an effect if it exists, and that they choose the correct inference test that matches the data and the question. If a statistically-oriented reviewer knows these are options for strengthening *their own* work, then that reviewer should also recognize – and insist – that such strengthening is important in *everyone's* work. This is the implementation of ASA Ethical Guidelines Principles A, B and C. When an author mistakenly describes failures to reject their null hypotheses as "evidence of equivalence" across groups, the statistically literate reviewer should recognize this as a null results, require revision of the narrative to reflect this fact, and ensure that the manuscript is otherwise methodologically sound. As long as it is sound, authors should be encouraged to consider revising the paper so that the negative results are tested and explained – so they can be considered for publication as the potentially important negative results that they are. It is essential for the statistician as well as the authors to recognize that characterizing failures to reject the null as evidence of equivalence is both incorrect and inappropriate. More extensive revision than simply changing that characterization from "evidence of equivalence" to "negative result" is required, but authors should be encouraged to consider doing this. The statistical reviewer strengthens the scientific literature by offering concrete recommendations as to how the authors may best achieve a rigorous and reproducible report of the negative result.

5. Informative reviewing

Whether a submission describes "negative" or "positive" results, statistical and data science practitioners have an ethical and professional obligation to provide an expert evaluation of the statistical methodology. This almost invariably leads to firm conclusions about whether or not any manuscript (or grant proposal) would make a meaningful contribution to the base of scientific knowledge.

5.1 Using reporting requirements to structure a review

Lang & Altman (2013) outline "basic statistical reporting" requirements for biomedical literature; more extensive requirements, for a wider range of statistical methodologies, are given in Hancock & Mueller (2010; see also Lang & Secic 2006). While the guides recommended above are helpful for determining whether (or not) all the required elements of reporting for any given method are included in the write-up, they do not guarantee that a manuscript with all requisite desiderata is a meaningful contribution to the literature. These guidance documents also do not provide guidance for how to begin, and/or structure, the review. It is recommended to use the journal's review criteria to structure the review, and in so doing, consider whether the authors can *or cannot* (or can, but <u>should not</u>) remediate/revise the paper to address the problems you have found. Those journals that do provide any reviewing suggestions or structure indicate there what

features they are obliged to consider (because these are public criteria). If the reporting requirements or guidelines are used, their components can be integrated into the structure requested by the journal. If there is no structure provided by the journal, indicate the reporting guidelines that you use – because the author (and possibly editors) may not be familiar with them.

The informative review focuses on what can, and what cannot, be revised. Some journals instruct reviewers to differentiate between "major" and "minor" revisions; these distinctions are less useful for the ultimate editorial decision about whether the paper should be published. Instead, reviewers can characterize the revisions they recommend or their comments according to whether or not the paper would need to be completely rewritten – or all data re-collected (reject! Do not revise!), as opposed to other issues that, if the paper is revised, could actually be effectively addressed. When deploying the specific skill set the quantitative scientist brings to a review, identifying fundamental and foundational flaws in study design that *cannot be remediated* is one of the most important aspects of having experienced quantitative reviewers' input. These problems do not represent "major" *revisions*, because the data cannot be used as intended – the entire paper has to be re-written (Results, Discussion, Conclusions, and sometimes, Methods). If the data were collected in a sufficiently structured way and simply not correctly analyzed, that *can* be remediated; however, in some cases the correct/appropriate analysis, for example, correcting for multiple comparisons, will eliminate any "significant" results.

An informative review identifies and discusses issues that *cannot* be revised, because addressing them would require all new data, a new experimental design, or to completely re-write the results and conclusions (and abstract, and probably also the title). If there are other issues that *can* actually be effectively revised, providing guidance to the author about how to do these will be helpful to their writing, but when these are presented after *non-revisable* problems are clearly described, it can support stronger design and data collection (as well as writing) in the future. The editor's decision on a paper uses this advice; when comments are characterized instead as "minor" or "major", the input is simply less useful to the editor.

5.2 Borrowing review structure from another journal for a review

If the journal you are reviewing for does not provide review criteria or a structured review sheet, borrow from another journal. As with the use of reporting guidelines to structure a review, it is recommended that reviewers notifying the author – and the editor – that your review/evaluation features criteria that were helpful for structuring *your* thinking about what makes a contribution to the literature worth publishing. It is essential that the criteria from the journal you're borrowing from are a match to the criteria for publication in the journal for which you're writing your review. That is, do not utilize a criterion from Journal B (borrowed) for your review of a manuscript submitted to Journal A that Journal A's editor cannot utilize. For example, if Journal A publishes only positive results and no negative results, but Journal B has a specific policy to publish negative results that are sufficiently rigorous, it does not help the author or the editor for Journal A for you to include comments about how nice it is to see a manuscript with negative results for Journal A. The editor for Journal A cannot use such an assessment, and a comment like that will only confuse the author.

An example of borrowing the review criteria from one journal to review a submission to another journal is to use the criteria for publication from the Public Library of Science

journal PLoS ONE. This journal is somewhat unique because it does not have a single disciplinary focus, and also, because innovation is not one of the publication criteria –so replication studies and papers reporting negative results, for examples, can be published if they meet the criteria for rigor and reproducibility that have been articulated. To be accepted for publication in PLOS ONE, research articles must satisfy the following criteria (PLOS ONE, no date; http://journals.plos.org/plosone/s/criteria-for-publication#loc-1):

- The study presents the results of primary scientific research and represent a contribution to the base of scientific knowledge. **
- Results reported have not been published elsewhere.
- Experiments, statistics, and other analyses are performed to a high technical standard and are described in sufficient detail. **
- Conclusions are presented in an appropriate fashion and are supported by the data. **
- The article is presented in an intelligible fashion and is written in standard English.
- The research meets all applicable standards for the ethics of experimentation and research integrity.
- The article adheres to appropriate reporting guidelines and community standards for data availability.

These criteria capture general features that many, if not most, journals would recognize as important for reviewers to consider (possibly in addition to additional criteria). Papers can, and probably should, always be reviewed at least in terms of the three key attributes indicated with ** above (plus whatever else the journal requests):

- Scientific research representing a meaningful contribution to the base of scientific knowledge;
- Contain experiments, statistics, and other analyses that are performed to a high technical standard and are described in sufficient detail.
- Present conclusions in an appropriate fashion that are supported by the data.

One can *always* integrate these features into a peer review; and one can also integrate these items with the guidelines for reporting mentioned earlier.

5.3 Non-inferiority study

These three "borrowed" review criteria are discussed for a review of a manuscript describing a *non-inferiority* study (see Armitage et al. 2002: 636 -639).

1. Does the paper present the results of primary scientific research and make a meaningful contribution to the base of scientific knowledge?

Non-inferiority and equivalence studies are *designed as such* with *a priori* specified bounds on outcome values representing "equivalence"; when reviewing a manuscript that describes an "equivalence" or non-inferiority study, evidence that this was the intended design should be present. For example, power calculations were based on this outcome; the pre-identified range of results that would be accepted as evidence of equivalence must be given; and the background and introduction must make the case for non-inferiority design. Without these features, it is unlikely that non-inferiority was the original scientific research question, so the manuscript is not a meaningful contribution as presented. This publication criterion is an essential one for determining whether a paper should be revised and resubmitted or not, because if the study design is flawed, no

amount of revision, short of rewriting, will suffice. Finding negative results (failing to reject the null) is not necessarily a sign that the study design was flawed.

2. Does the paper contain experiments, statistics, and other analyses that are performed to a high technical standard and are described in sufficient detail?
Similar to the previous item, the methods section must describe how results do or do not meet the non-inferiority study design features, for example, results fall *above* a 100(1-α)% confidence interval at the *a priori* (i.e., stated in the methods section) lower bound (Armitage et al. 2002, p. 636-7). Moreover, the introduction should have made the case for the design, contextualizing the study design and experiments as negative or non-inferiority by design, and the argument should not use a p-value greater than 0.05 to conclude equivalence. The results should be contextualized with non-inferiority as the stated objective, and this should match both the introduction and conclusions. There should be a sensitivity analysis, and limitations should be mentioned and adequately discussed. These aspects constitute "sufficient detail" for a reviewer to determine if the authors did perform the design and execution of the non-inferiority study to a high technical standard. Statistical analyses can always be redone in a revision, but if a manuscript has incorrect, or inappropriate, (or just flawed) design, revising the analysis will not change the manuscript into a "meaningful contribution to the base of scientific knowledge". Also, if the authors (in the end) wish they had done a non-inferiority study, so that the negative results could be described as such, it is unlikely that simply designing a new analysis plan will change the paper into a meaningful contribution on non-inferiority.

3. Does the paper present conclusions in an appropriate fashion that are supported by the data?
As noted above, in a non-inferiority study, results should be contextualized with non-inferiority as the stated objective (in both introduction and conclusions). The presentation of conclusions in an appropriate fashion includes both using correct and appropriate narrative given that the study was originally designed to be a non-inferiority study. Conclusions that suggest the authors had originally expected to see a difference (i.e., not originally designed as an equivalence or non-inferiority study) should then clearly identify the negative results as failures to reject the null, and the narrative should focus on limitations, sensitivity analyses, and other "sufficient detail" to permit the reader to see that the experiments and analyses were in fact "performed to a high technical standard" and that, because of that fact, failures to reject the null suggest that the alternative hypothesis may need to be reconsidered. This means going beyond stating that "$p>0.05$" or acknowledging that the results are negative; because as Lang & Secic (2006, p. 58) point out, "(g)roups that do not differ statistically cannot necessarily be assumed to be (clinically) equivalent." Sensitivity analysis, and limitations, must be adequately discussed.

5.4 Other Negative Results
Non-inferiority or equivalence studies are more common in clinical and pharmacological research than in other scientific fields. Not all statisticians are familiar with this design and, if they are not experienced with clinical and pharmacologic studies, authors may also not be familiar with the design – specifically, the requirement that the entire study is designed to test the hypothesis of equivalence, meaning that failing to reject the null hypothesis of "no difference" is very different. For this reason, and as noted earlier, reviewers should be alert to the possibility that authors may mistakenly characterize an experiment or trial that found "no significant differences" as an "equivalence" or non-

inferiority study. If there is insufficient evidence to reject the null hypothesis, authors should/should be encouraged to plan and execute additional analyses, as is feasible, to probe the nature of this failure to reject, so that if expected outcomes (e.g., "significant differences across groups") are not observed, or previously published relationships are not replicated, their work will actually be incorporated into the wider literature. If a study was adequately powered (which would be discussed in the methods and discussion, if expected or prior effects are not replicated), and lack of significant effects explored (e.g., using sensitivity analyses; alternative post hoc explanations that are also tested and discussed), then the three critical review features borrowed from PLOS ONE publication criteria may be met and the manuscript would be potentially publishable.

If weaknesses in a study design or paradigm are identified by a pilot study or other work, or if results, being negative or representing failures to reject the null hypothesis, critique a dominant paradigm or theory, it is very important to encourage authors to submit a rigorous and reproducible report. These are essential results – but must have sensitivity analysis or other appropriate and relevant contextualization to ensure that "lessons learned" are conclusions that the data do support. Clearly, the reporting guidelines can be used to ensure the reproducibility of the analyses that are described; the three review features borrowed from PLOS ONE are also important to consider. However, specific requirements from the journal to which the paper was submitted must also be considered (and there will or should be some overlap).

Although Lang & Altman (2013) do not consider negative results specifically, they outline Statistical Analyses and Methods in the Published Literature: The SAMPL Guidelines for reporting Statistical Methods and Results. All peer *reviews* should document that manuscripts adhere to the SAMPL reporting guidelines at a minimum, and both reviewers and authors can also consider resources like Hancock & Mueller (2010)– to ensure that all contributions to the peer-reviewed scientific literature follow consensus guidelines for what is essential and appropriate to be reported in the description of different analytic methods. When authors are drafting manuscripts for peer review (and grant proposals), these criteria and guidelines are helpful to ensure that what is described and reported, irrespective of whether results are "positive" or "negative", is complete and correct.

6. Discussion

"(S)ignificance tests and confidence intervals do not by themselves provide a logically sound basis for concluding an effect is present or absent with certainty or a given probability." (Greenland, et al. 2016; p10) Effective reviewing of negative results can be complicated by the inaccurate perception that studies presenting "positive" results are actually **correct**; since many studies presenting "positive" results may or do turn out to be irreproducible, actually-negative results are already a large part of the research record. Thus, <u>all</u> work should be considered for its potential to "make a meaningful contribution to the base of scientific knowledge" – being both rigorous and reproducible, whether the results are positive or negative. The features of a competent and informative statistical review do not solely focus on the "methods section"; and these should also be utilized by authors to ensure that what the reviewer needs to find in their manuscript can be found.

The ASA Ethical Guidelines for Statistical Practice encourage competent and informative peer review under three key principles: **A**. professional integrity and accountability; **B.** integrity of data and methods; and **C**. responsibilities to science. The ethical statistician

applies these principles whether they are producing the write-up or reviewing it. Following these principles and providing an informative review also meets two other Ethical Guidelines principles: **D.** responsibilities to research subjects; and **F.** responsibilities to other statisticians or statistics practitioners. The practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers drawing on their own unique sets of experiences can arrive at different and potentially diverging judgments about the plausibility of different explanations. Reviewers should be as informative in reviews of negative results as they are for any other results. Negative results may need special attention, but every manuscript should be evaluated for its potential to contribute to the literature. Incomplete or incorrect methods, a failure to align methods with the research question, inappropriate conclusions, and mischaracterization of results are all features that should be identified in manuscripts presenting positive as well as negative results. Structure for informative reviewing can be borrowed, if one finds a journal that has a usefully-structured review form, one can always use that to structure reviews in other (possibly less structured) cases. Reviews can also utilize the guidelines for *reporting* to make comments for the author about what is missing, or how best to revise for the most rigorous and reproducible contribution to the literature.

Every journal requires – whether or not these are featured in the review or publication criteria – that what authors describe in their Methods section is actually what generated the results appearing in the Results section. Further, in *any* submission for peer review, the author must justify that the statistical analysis was appropriate for the data and for the research question; that the results match the problem formulation; that sufficient information is presented to demonstrate that the hypothesis was rigorously tested; and that enough information is provided - clearly enough – so that the work is reproducible. Thus, the statistical reviewer of manuscripts as well as grants *cannot* just read the statistics or methods section. A statistical reviewer should be expected and encouraged to read the entire document; asking for a statistical review that limits or otherwise directs attention to just the methods or just methods and results violates the ASA Ethical Guideline relating to the responsibilities that "employers" of statisticians have, Principle **H**: "Employers, funders, or those who commission statistical analysis have an obligation to rely on the expertise and judgment of qualified statisticians for any data analysis." (see Appendix). Relying on the judgment of a qualified statistical reviewer confers the same obligation to allow the statistical reviewer access to all relevant information (i.e., the entire document) to form that judgment.

To optimize the utility, generalizability, and interpretability of their results, authors should convince the reviewer that they checked to ensure that their data met the assumptional requirements of the analysis method they chose. They must describe exactly what was done – if the analysis did or did not include outlier data points, one must describe the criteria used to identify them. Depending on how "odd" these data points are, the reviewer must also consider the importance for the interpretability of results that the exact same analysis was done with *and without* the outliers included. If those outliers are "odd" but not implausible (i.e., biologically or logically impossible) values, then sensitivity analyses (comparing inferences with and without those points) ensure that their exclusion as outliers does not bias the perception of the relationship of interest. This is especially important if the bias is *away* from the null, causing the false rejection of the null hypothesis when the evidence doesn't warrant that. If the bias is *towards* the null, it might explain how negative results were obtained, if they are the result.

These recommendations for reviewing should be followed whether the article presents positive or negative results. If an author or a research group has gone to the effort to write up what are actually negative results, but have mischaracterized them as positive or "equivalence" results, they should be encouraged to reformulate their argument so that these negative results can become a rigorous and reproducible contribution to the literature.

Negative results are essential contributions to the research record in every scientific discipline, but not if they are badly done, incompletely reported, or otherwise ineffectively reviewed. This article outlined definitions of "negative results" and highlighted key features of the planning, execution, analysis, and write up of studies that can be utilized to compose informative reviews of submitted manuscripts that describe negative (or positive) results. All quantitative scientists are encouraged to review actively and informatively, because they are uniquely qualified to provide specific input and advice to journal editors on the viability of *all* research results with respect to the statistical methods. By carefully reviewing all submissions, including those with "negative" results, statisticians and data scientists can fulfill their ethical and professional obligations while advocating for the integrity of the scientific record.

Acknowledgements

The author is the Chair of the ASA Committee on Professional Ethics (2017-2019) and a Section Editor for the journal PLOS ONE. She also serves on editorial boards for several other journals and reviews grants for the National Institutes of Health, National Science Foundation, and the Department of Defense. The views represented here are her own and do not necessarily reflect the views of any of these journals or funders, the other members of the Committee on Professional Ethics, or the ASA.

APPENDIX: ASA ETHICAL GUIDELINES – REVISED

Ethical Guidelines for Statistical Practice

Prepared by the Committee on Professional Ethics of the American Statistical Association

Approved by ASA Board April 2016

Purpose of the Guidelines

The American Statistical Association's Ethical Guidelines for Statistical Practice are intended to help statistics practitioners make decisions ethically. Additionally, the Ethical Guidelines aim to promote accountability by informing those who rely on statistical analysis of the standards that they should expect. The discipline of statistics links the capacity to observe with the ability to gather evidence and make decisions, providing a foundation for building a more informed society. Because society depends on informed judgments supported by statistical methods, all practitioners of statistics, regardless of

training and occupation or job title, have an obligation to work in a professional, competent, and ethical manner and to discourage any type of professional and scientific misconduct.

Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations. In some situations, Guideline principles may conflict, requiring individuals to prioritize principles according to context. However, in all cases, stakeholders have an obligation to act in good faith, to act in a manner that is consistent with these Guidelines, and to encourage others to do the same. Above all, professionalism in statistical practice presumes the goal of advancing knowledge while avoiding harm; using statistics in pursuit of unethical ends is inherently unethical.

The principles expressed here should guide both those whose primary occupation is statistics and those in all other disciplines who use statistical methods in their professional work. Therefore, throughout these Guidelines, the term "statistician" includes all practitioners of statistics and quantitative sciences, regardless of job title or field of degree, comprising statisticians at all levels of the profession and members of other professions who utilize and report statistical analyses and their implications.

A. Professional Integrity and Accountability

The ethical statistician uses methodology and data that are relevant and appropriate, without favoritism or prejudice, and in a manner intended to produce valid, interpretable, and reproducible results. The ethical statistician does not knowingly accept work for which he/she is not sufficiently qualified, is honest with the client about any limitation of expertise, and consults other statisticians when necessary or in doubt.

The ethical statistician:

1. Identifies and mitigates any preferences on the part of the investigators or data providers that might predetermine or influence the analyses/results.
2. Employs selection or sampling methods and analytic approaches appropriate and valid for the specific question to be addressed, so that results extend beyond the sample to a population relevant to the objectives with minimal error under reasonable assumptions.
3. Respects and acknowledges the contributions and intellectual property of others.
4. When establishing authorship order for posters, papers, and other scholarship, strives to make clear the basis for this order, if determined on grounds other than intellectual contribution.
5. Discloses conflicts of interest, financial and otherwise, and manages or resolves them according to established (institutional/regional/local) rules and laws.
6. Accepts full responsibility for his/her professional performance. Provides only expert testimony, written work, and oral presentations that he/she would be willing to have peer reviewed.

B. Integrity of data and methods

The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may impact the integrity or reliability of the statistical analysis. Objective and valid interpretation of the results requires that the underlying analysis recognizes and acknowledges the degree of reliability and integrity of the data.

The ethical statistician:

1. Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms.
2. Reports the limitations of statistical inference and possible sources of error.
3. In publications, reports, or testimony, identifies who is responsible for the statistical work if it would not otherwise be apparent.
4. Reports the sources and assessed adequacy of the data; accounts for all data considered in a study and explains the sample(s) actually used.
5. Clearly and fully reports the steps taken to preserve data integrity and valid results.
6. Where appropriate, addresses potential confounding variables not included in the study.
7. In publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics.
8. In publications or testimony, identifies the ultimate financial sponsor of the study, the stated purpose, and the intended use of the study results.
9. When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting.
10. To aid peer review and replication, shares the data used in the analyses whenever possible/allowable, and exercises due caution to protect proprietary and confidential data, including all data that might inappropriately reveal respondent identities.
11. Strives to promptly correct any errors discovered while producing the final report or after publication. As appropriate, disseminates the correction publicly or to others relying on the results.

C. Responsibilities to Science/Public/Funder/Client

The ethical statistician supports valid inferences, transparency, and good science in general, keeping the interests of the public, funder, client, or customer in mind (as well as professional colleagues, patients, the public, and the scientific community).

The ethical statistician:

1. To the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or precision.

2. Strives to explain any expected adverse consequences of failure to follow through on an agreed-upon sampling or analytic plan.
3. Applies statistical sampling and analysis procedures scientifically, without predetermining the outcome.
4. Strives to make new statistical knowledge widely available to provide benefits to society at large and beyond his/her own scope of applications.
5. Understands and conforms to confidentiality requirements of data collection, release, and dissemination and any restrictions on its use established by the data provider (to the extent legally required), and protects use and disclosure of data accordingly. Guards privileged information of the employer, client, or funder.

D. Responsibilities to Research Subjects

The ethical statistician protects and respects the rights and interests of human and animal subjects at all stages of their involvement in a project. This includes respondents to the census or to surveys, those whose data are contained in administrative records, and subjects of physically or psychologically invasive research.

The ethical statistician:

1. Keeps informed about and adheres to applicable rules, approvals, and guidelines for the protection and welfare of human and animal subjects.
2. Strives to avoid the use of excessive or inadequate numbers of research subjects, and excessive risk to research subjects (in terms of health, welfare, privacy, and ownership of their own data), by making informed recommendations for study size.
3. Protects the privacy and confidentiality of research subjects and data concerning them, whether obtained from the subjects directly, other persons, or existing records. Anticipates and solicits approval for secondary and indirect uses of the data, including linkage to other data sets, when obtaining approvals from research subjects, and obtains approvals appropriate to allow for peer review and independent replication of analyses.
4. Knows the legal limitations on privacy and confidentiality assurances and does not over-promise or assume legal privacy and confidentiality protections where they may not apply.
5. Considers whether appropriate research-subject approvals were obtained before participating in a study involving human beings or organizations, before analyzing data from such a study, and while reviewing manuscripts for publication or internal use. The statistician considers the treatment of research subjects (e.g., confidentiality agreements, expectations of privacy, notification, consent, etc.) when evaluating the appropriateness of the data source(s).
6. In contemplating whether to participate in an analysis of data from a particular source, refuses to do so if participating in the analysis could reasonably be interpreted by individuals who provided information as sanctioning a violation of their rights.
7. Recognizes that any statistical descriptions of groups may carry risks of stereotypes and stigmatization. Statisticians should contemplate, and be sensitive to, the manner

in which information is framed so as to avoid disproportionate harms to vulnerable groups.

E. Responsibilities to Research Team Colleagues

Science and statistical practice are often conducted in teams made up of professionals with different professional standards. The statistician must know how to work ethically in this environment.

The ethical statistician:

1. Recognizes that other professions have standards and obligations, that research practices and standards can differ across disciplines, and that statisticians do not have obligations to standards of other professions that conflict with these Guidelines.

2. Ensures that all discussion and reporting of statistical design and analysis is consistent with these Guidelines.

3. Avoids compromising scientific validity for expediency.

4. Strives to promote transparency in design, execution, and reporting or presenting of all analyses.

F. Responsibilities to Other Statisticians or Statistics Practitioners

The practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers drawing on their own unique sets of experiences can arrive at different and potentially diverging judgments about the plausibility of different explanations. Even in adversarial settings, discourse tends to be most successful when statisticians treat one another with mutual respect and focus on scientific principles, methodology and the substance of data interpretations. Out of respect for fellow statistical practitioners, the ethical statistician:

1. Promotes sharing of data and methods as much as possible and as appropriate without compromising propriety. Makes documentation suitable for replicate analyses, metadata studies, and other research by qualified investigators.
2. Helps strengthen the work of others through appropriate peer review; in peer review, respects differences of opinion and assesses methods, not individuals. Strives to complete review assignments thoroughly, thoughtfully, and promptly.
3. Instills in students and non-statisticians an appreciation for the practical value of the concepts and methods they are learning or using.
4. Uses professional qualifications and contributions as the basis for decisions regarding statistical practitioners' hiring, firing, promotion, work assignments, publications and presentations, candidacy for offices and awards, funding or approval of research, and other professional matters.
5. Does not harass or discriminate.

G. Responsibilities Regarding Allegations of Misconduct

The ethical statistician understands the difference between questionable scientific practices and practices that constitute misconduct, avoids both, but knows how each should be handled.

The ethical statistician:

1. Avoids condoning or appearing to condone incompetent or unethical practices in statistical analysis.

2. Recognizes that differences of opinion and honest error do not constitute misconduct; they warrant discussion, but not accusation.

3. Knows the definitions of, and procedures relating to, misconduct. If involved in a misconduct investigation, follows prescribed procedures.

4. Maintains confidentiality during an investigation, but discloses the investigation results honestly to appropriate parties and stakeholders once they are available.

5. Following an investigation of misconduct, supports the appropriate efforts of all involved, including those reporting the possible scientific error or misconduct, to resume their careers in as normal a manner as possible.

6. Avoids, and acts to discourage, retaliation against or damage to the employability of those who responsibly call attention to possible scientific error or misconduct.

H. Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners

Those employing any person to analyze data are implicitly relying on the profession's reputation for objectivity. However, this creates an obligation on the part of the employer to understand and respect statisticians' obligation of objectivity.

Those employing statisticians are expected to:

1. Recognize that the Ethical Guidelines exist, and were instituted, for the protection and support of the statistician and the consumer alike.

2. Recognize that valid findings result from competent work in a moral environment. Employers, funders, or those who commission statistical analysis have an obligation to rely on the expertise and judgment of qualified statisticians for any data analysis. This obligation may be especially relevant in analyses that are known or anticipated to have tangible physical, financial, or psychological impacts.

3. Recognize that the results of valid statistical studies cannot be guaranteed to conform to the expectations or desires of those commissioning the study or the statistical practitioner(s).

4. Recognize that it is contrary to these Guidelines to report or follow only those results that conform to expectations without explicitly acknowledging competing findings and the basis for choices regarding which results to report, use, and/or cite.

5. Recognize that the inclusion of statistical practitioners as authors, or acknowledgement of their contributions to projects or publications, requires their explicit permission because it implies endorsement of the work.

6. Support sound statistical analysis and expose incompetent or corrupt statistical practice.

7. Strive to protect the professional freedom and responsibility of statistical practitioners who comply with these Guidelines.

References

Aczel B, Palfi B, Szaszi B. (2017). Estimating the evidential value of significant results in psychological science. *PLoS One*. 2017 Aug 18;12(8):e0182651. doi: 10.1371/journal.pone.0182651. eCollection 2017.

Altman DG. (1998). Statistical reviewing for medical journals. *Statistics in Medicine* (Dec 15)17(23):2661-74.

American Statistical Association (ASA). (2016). *ASA Ethical Guidelines for Statistical Practice*. Downloaded from http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx 20 May 2016

Armitage P, Berry G & Matthews JNS. (2002). *Statistical methods in medical research, 4E*. Malden, MA: Blackwell Science.
Baker M. (2016). Is there a reproducibility Crisis? *Nature 533*, 452–454.

Begley CG, Ioannidis JP. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research* 116(1):11626. doi: 10.1161/CIRCRESAHA.114.303819.

Bloom BJ (Ed), Englehart MD, Furst EJ, Hill WH & Krathwohl DR. (1956). *Taxonomy of educational objectives: the classification of educational goals, by a committee of college and university examiners. Handbook I: Cognitive Domain.* New York: David McKay.

Collins FS & Tabak LA. (2014). Policy: NIH Plans to enhance reproducibility. *Nature* 505, 612–613 (30 January 2014) doi:10.1038/505612a

Casadevall A, Ellis LM, Davies EW, McFall-Ngai M, Fang FC. (2016). A Framework for Improving the Quality of Research in the Biological Sciences. *MBio* 7(4). pii: e01256-16. doi: 10.1128/mBio.01256-16.

Cumming G, Fidler F, Vaux DL. (2007). Error bars in experimental biology. *Journal of Cell Biology, 177*: 7–11.

Curran-Everett D. (2016). Explorations in statistics: statistical facets of reproducibility. *Advances in Physiology Education* 40(2):248-52. doi: 10.1152/advan.00042.2016.

Egger M, Smith GD, Schneider M, Minder C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 315 :629. Downloaded from https://doi.org/10.1136/bmj.315.7109.629 on 18 August 2017

Fidler F, Thomason N, Cumming G, Finch S, Leeman J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science 15*:119–126.

Finch S, Cumming G, Thomason N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement 61*:181–210.

Flier JS. (2017). Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy. *Molecular Metabolism* 6(1):2-9. doi: 10.1016/j.molmet.2016.11.006.

Galipeau J, Barbour V, Baskin P, Bell-Syer S, Cobey K, Cumpston M, Deeks J, Garner P, MacLehose H, Shamseer L, Straus S, Tugwell P, Wager E, Winker M, Moher D. (2016). A scoping review of competencies for scientific editors of biomedical journals. *BMC Medicine* 14:16. doi: 10.1186/s12916-016-0561-2.

Greenland S, Senn SJ, Rothman K, Carlin J, Poole C, Goodman S, and Altman D. (2016). Statistical Tests, P-values, Confidence Intervals, and Power: A Guide

to Misinterpretations. *European Journal of Epidemiology* 31(4):337-50. doi: 10.1007/s10654-016-0149-3. Downloaded from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/ on 12 Sept 2017.

Grinnell, F. (2009). *Everyday Practice of Science: Where Intuition and Passion Meet Objectivity and Logic*. New York, NY: Oxford University Press.

Hancock GR, Mueller RO. (Eds.) (2010). *The Reviewer's Guide to Quantitative Methods in the Social Sciences*; Routledge: New York, NY, USA, 2010.

Hayden EC. (2013). Weak statistical standards implicated in scientific irreproducibility. *Nature News* **2013**, doi:10.1038/nature.2013.14131.

Henson RK, Hull DM, William CS. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Education Research 39*, 229–240.

Hirst A, Altman DG. (2012). Are Peer Reviewers Encouraged to Use Reporting Guidelines? A Survey of 116 Health Research Journals. *PLoS ONE* 7(4): e35621.

Ioannidis JPA. (2017). Acknowledging and overcoming nonreproducibility in basic and preclinical research. *Journal of the American Medical Association* 317(10):1019-1020. doi:10.1001/jama.2017.0549

Ioannidis, JPA. (2005). Why most published research findings are false. *PLoS Medicine 2*, e124

Kenett RS, Shmueli G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nature Methods* 12(8):699. doi: 10.1038/nmeth.3489.

Kim YH, Chiu CY, Bregant J. (2015). Unskilled and Don't Want to Be Aware of It: The Effect of Self-Relevance on the Unskilled and Unaware Phenomenon. *PLoS One.* 10(6):e0130309. doi: 10.1371/journal.pone.0130309. eCollection 2015.

Kruger J, Dunning D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6):1121-34.

Lang T. (2010). CONSORTing with a QUOROM of MOOSES: the standards movement in scientific reporting. *Neurourolgy and Urodynamics* 29(1):28-9. doi: 10.1002/nau.20854.

Lang T, Altman DG. (2013). Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In P Smart, H Maisonneuve, A Polderman (Eds). *EASE Science Editors' Handbook, 2E*. European Association of Science Editors.

Lang TA & Secic M. (2006). *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers 2E*. Philadelphia PA: American College of Physicians.

Lewin-Koh N, Taper ML & Lele SR. (2004). A brief tour of statistical concepts. In ML Taper & SR Lele (Eds.), *The nature of scientific evidence: statistical, philosophical, and empirical considerations*. Chicago, IL: The University of Chicago Press. Pp 3-16.

McNutt M. (2014). Editorial: Journals unite for reproducibility. *Science* 7 November 2014: Vol. 346 no. 6210 p. 679. DOI: 10.1126/science.aaa1724

Moher D, Altman DG (2015) Four Proposals to Help Improve the Medical Research Literature. *PLoS Medicine* 12(9): e1001864. Downloaded from https://doi.org/10.1371/journal.pmed.1001864 on 17 August 2017

Moher D, Weeks L, Ocampo M, Seely D, Sampson M, Altman DG, et al. (2011). Describing reporting guidelines for health research: a systematic review. *Journal of Clinical Epidemiology* 64:718-42

Moher D, Galipeau J, Alam S, Barbour V, Bartolomeos K, Baskin P, Bell-Syer S, Cobey KD, Chan L, Clark J, Deeks J, Flanagin A, Garner P, Glenny AM, Groves T, Gurusamy K, Habibzadeh F, Jewell-Thomas S, Kelsall D, Lapeña JF Jr, MacLehose H, Marusic A, McKenzie JE, Shah J, Shamseer L, Straus S, Tugwell P, Wager E, Winker M, Zhaori G. (2017). Core competencies for scientific editors of biomedical journals: consensus statement. *BMC Medicine* 15(1):167. doi: 10.1186/s12916-017-0927-0.

Nature (2016). *Editorial: Reality check on reproducibility*. Downloaded from https://www.nature.com/news/reality-check-on-reproducibility-1.19961 on 10 Sept 2017

Nicholls SG, Langan SM, Benchimol EI, Moher D. (2016). Reporting transparency: making the ethical mandate explicit. *BMC Medicine* 14:44. doi: 10.1186/s12916-016-0587-5.

PLOS ONE. (no date). Criteria for publication. Downloaded from http://journals.plos.org/plosone/s/criteria-for-publication 20 July 2010

Popper K. (1959/2002). *The logic of scientific discovery*. London, UK: Routledge.

Pusztai L, Hatzis C, Andre F. (2013). Reproducibility of research and preclinical validation: problems and solutions. *Nature Reviews Clinical Oncology* 10(12):720-4. doi: 10.1038/nrclinonc.2013.171.

Szucs D, Ioannidis JPA. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience* 11:390. doi: 10.3389/fnhum.2017.00390.

Tractenberg RE. (2017). How the Mastery Rubric for Statistical Literacy Can Generate Actionable Evidence about Statistical and Quantitative Learning Outcomes. *Education Sciences* 7(1), 3; *doi:10.3390/educsci7010003.*

Vaux, DL. (2012). Research methods: Know when your numbers are significant. *Nature 492*, 180–181.

Vaux DL. (2014). Basic statistics in cell biology. *Annual Review of Cellular and Developmental Biology* **2014**, *30*, 23–37.

Weissgerber TL, Garovic VD, Milin-Lazovic JS, Winham SJ, Obradovic Z, Trzeciakowski JP, Millic NM. (2016). Reinventing biostatistics education for basic scientists. *PLoS Biology 4*, e1002430.