Commentary

Daniel B. Wright*¹

Improving Trust in Research: Supporting Claims with Evidence

<https://doi.org/10.1515/edu-2020-0106>

received April 27, 2020; accepted May 11, 2020.

Abstract: Trust in science is important, and *Open Education Studies* aims to publish trusted research. Two issues are addressed here: access to the data on which the research is based and how these data are analyzed. Some guidelines from other entities are discussed. As a new journal our guidelines should be influenced by the opinions of readers and authors, and as such we welcome discussion of how to ensure trust in the research *OES* publishes.

Keywords: replication; open science; trust.

1 Introduction

Miyakawa (2020) is the editor-in-chief of *Molecular Brain* and receives lots of papers. Sometimes when editors, reviewers, and readers look at the results sections of papers, something does not look right. Miyakawa wrote to 41 authors asking them to provide further evidence (usually raw data) to back up some of the claims made in their papers before sending the papers out for peer review. The response was depressing.

- Twenty-one (51%) of these papers were withdrawn.
- Nineteen (46%) did not provide appropriate data to address Miyakawa's concerns. These were desk rejected (i.e., rejected without peer review).
- One (ONE! Just 2%) provided satisfactory data to relieve Miyakawa's concerns. This single paper was sent for peer review and was ultimately published.

¹ Daniel Wright is the Dunn Family Foundation Endowed Chair and Professor of Educational Assessment, the Department of Educational Psychology & Higher Education, UNLV, USA. Email: daniel.wright@unlv.edu or dbrookswr@gmail.com.

*Corresponding author: Daniel B. Wright, University of Nevada at Las Vegas, Las Vegas, Nevada, United States, E-mail: dbrookswr@gmail.com

Using what Spiegelhalter (2019) calls icon arrays to depict low frequencies and proportions, the proportion of papers where adequate data were provided is shown in Figure 1.

While there are reasons why it may be impractical for a few authors to send their data to an editor (see below). Having only one author provide appropriate data is horrible.

The value of any research study is based on the data (broadly defined) for that study, and a trust the public has in research as a whole, in the individual researchers, and the journal. This has become more difficult in recent years. Some politicians fuel distrust in science/scientists and some media stars endorse bogus or junk science with unfounded claims. Predatory journals feed into this distrust, which is predictable given situations like the clearly spoof essay about the Star Wars entity *midichlondria* being accepted by several outlets (<https://www.discovermagazine.com/mind/predatory-journals-hit-by-star-wars-sting>) with scientific sounding names: *The American Journal of Medical and Biological Research*; *International Journal of Molecular Biology: Open Access*; *Austin Journal of Pharmacology and Therapeutics*; and *American Research Journal of Biosciences*. These sound like they could be legitimate journals. It is important for non-predatory journals to show that trust is deserved for the papers that they publish.

The goal of *Open Education Studies* (OES) is to publish high-quality and trusted research. The goal of this commentary is to provide some guidelines to help you to submit such research. Miyakawa's no-data-no-science rule, providing that it is interpreted broadly and implemented with some common-sense, seems worth adhering to. If you submit an empirical paper and an editor asks for the evidence on which you base your claims, you should provide this evidence or convince the editor why this is not possible. There are data privacy issues that occur in education research, and we will work with you on these. In the next section guidelines/rules from other entities are presented and ways for *OES* to move forward discussed.

Open should mean all aspects are as open as feasible. The open science movement for journals has meant

Figure 1: An icon plot showing the number of authors (1 of 41) providing appropriate evidence to Miyakawa (2020).

allowing everyone with the internet to access research papers (i.e., open access). But some open science initiatives also mean that everyone has access to the research stimuli and data (see <https://osf.io/>). Open should also mean that when a paper is published this opens a dialog among the authors and the public about aspects of the research. The publication of a paper should be thought of as the first step of establishing the paper's impact.

2 Data Storage and Availability

The OES author guidelines do not specify rules about making data available, so guidelines from a funding agency, a learned society and another journal will be considered. First, the Economic and Social Research Council (ESRC) is a government entity that funds education research in the UK. It has a clear policy on making data available:

All data created or repurposed during the lifetime of an ESRC grant must be made available for re-use or archiving within three months of the end of the grant.
[esrc.ukri.org/funding/guidance-for-grant-holders/research-data-policy/](https://www.esrc.ukri.org/funding/guidance-for-grant-holders/research-data-policy/)

Their argument is based on the value for others conducting secondary analysis and that the people of the UK paid for the research so should have access to the data. Because the ESRC fund the research and may provide further funding to those institutions, there is a large incentive for researchers to adhere to their policy. Their archive is: www.data-archive.ac.uk/ and this has become a valuable resource for secondary data analysis.

The American Education Research Association (AERA) describes its policy for empirical research for data access:

7.5. The data or empirical materials relevant to the conclusions should be maintained in a way that a qualified researcher with a copy of the relevant data and description of the analysis

procedures could reproduce the analysis or trace the trail of evidence leading to the author's conclusions.
www.aera.net/Portals/38/docs/12ERv35n6_Standard4Report%20.pdf

Many journals also list their guidelines. The flagship journal of the Association for Psychological Science (APS), the journal *Psychological Science* states: "Researchers are also asked to make their materials, data, and analysis scripts available to reviewers (in ways that are ethically appropriate and practically feasible)." Once an article is accepted: "authors will be required to inform readers how they may access study-related data and materials, or of restrictions thereon" (www.psychologicalscience.org/publications/psychological-science/ps-submissions#CRIT). The APS journals encourage making data and other materials, like statistical code, available and offer "badges" to authors who do different things that are designed to increase trust in research.

The purpose of all of these is not to catch people making up data. The purpose is two-fold: encourage trust and allow other people—from different research perspectives and with different methodological skills—to explore the data. Collecting good data is difficult. It is inefficient and unfortunate if data are left unreported that could positively impact society.

3 Why authors may not have submitted data

More than half of authors who Miyakawa (2020) wrote to withdrew their papers rather than provide evidence for their claims. All of us can speculate on their motives. Some may have made up their data and were fearful that they would be exposed. Unfortunately this occurs. The retractionwatch.com webpage highlights some cases of

research fraud. If the data were made up, it is good that these people retracted their submissions. But there are non-fraudulent reasons why someone might not *wish* to submit their data. Here are some of the reasons that people put forward.

- The data file is not well organized and only interpretable by the author.

Preparing the data file so that others can interpret the data is important both in case something happens to the author (e.g., dies) and for the transparency of the research process. Ideally researchers have their data files in good shape when they submit their paper, but there may be a delay in preparing the data file if requested. This is why the ESRC have a three month grace period. This long period is because their reports have specific due-dates, so grantees may be rushing to finish these. Journals usually do not have these deadlines so authors should not feel rushed (special issues often do have due dates). If data are requested a lengthy delay in submitting the data will prolong the review process. The *Psychological Science* rule (and this is used with other journals too) of telling readers how they can access the data (or why not) seems reasonable. This could be done on the original manuscript. It would be a relatively minor change for most authors to get into the habit of doing this.

- The data include identifiable information.
Regulations in different countries (e.g., FERPA [Family Educational Rights and Privacy Act] in the US, GDPR [General Data Protection Regulation] in Europe) mean that you cannot publish identifiable information without consent. Data files can be de-identified in several ways. Sometimes this means removing some columns and creating new student identifiers. Other times this involves more work, like adding random error to data. *OES* editors will work with you on this. Further, many *Memoranda of Understanding* (*MoU*) and research ethics committees (RECs, often called IRBs) proposals restrict who can see the data. It is important that *MoUs* allow the research to be published and de-identified data to be made available in accordance with the appropriate funding agency/society/journal guidelines.
- The authors wish to be the only ones with access to the data while they are publishing multiple studies from it.
There are several aspects to this reason. Sometimes a big study is conducted and there are several aspects that warrant research. While it can be argued that it is good for science to allow others to also try to glean

useful findings from the data, there are also arguments that the researchers should gain something from gathering the data. Data requests will be for those data relevant to the specific findings and reviewer guidelines stress the confidentiality of the review process. It is likely if data are requested these can be examined by the editor. After publication archiving the data onto a repository so that they can be cited by others is recommended. This can be done just for the data relevant to the paper's findings. At present, archiving data is not required by *OES* (see the next section).

- Some authors worry that people without their expertise/perspective could interpret the data differently and reach different conclusions.

This is why releasing the data is important. If someone from a different perspective and with different analytic tools discovers something different, this allows discussion about the relative merits of both approaches and therefore the validity of both conclusions.

4 Archiving Data

There are several ways to archive your data. This includes on statistical archives like CRAN (CRAN, the Comprehensive R Archive Network), general file sharing sites like GitHub and ResearchGate, personal web pages, and with the journals. For example, in Wright (2019) readers are told that information is available at: <https://github.com/dbrookswr/tarre>. Using GitHub can be intimidating (its advantage for my purpose was that people could add their own examples for the statistical function used in that paper). Other data archives are more user-friendly for most people. I am pleased to say *OES* accepts data and will publish this as supplementary material of the paper. For example, Molinari & Gasparini (2019) have their data published at: https://www.degruyter.com/view/journals/edu/1/1/article-p24.xml?tab_body=supplementaryMaterials-74965. Others are encouraged to use these data.

OES is at present not requiring data to be archived or submitted, even if requested, but this will be taken into account in the decision process. If you make a claim that is based on evidence, but do not share this evidence, it is difficult to evaluate your claim. Evaluation of claims is of course at the heart of accept/reject decisions, and so the likely outcome is rejection if requested material is not made available.

5 Numeric Errors

Sometimes the data requests editors and reviewers have are because while reading the manuscript some of the numbers do not seem corrected. These are often simple data entry errors, and it is worthwhile discussing how to minimize these. Most people use spell checkers to prevent some typos and help with some harder to spell words. No spell checker is perfect. Its suggestions are often for the wrong word (e.g., the word base will not include discipline specific jargon) and it does not catch every error since sometimes the mis-spelling of one word is another word. Checking numeric typos is harder. Consider if a manuscript reports: $F(1, 79) = 2.82, p = .04$. Is this correct? While it is possible to look up each p -value, Nuijten et al. (2016) have automated this process in their R package **statcheck** (Epskamp & Nuijten, 2018). Once this free package is downloaded, within R (which is also free), the following:

```
statcheck("The result is significant, F(1,79) = 2.82, p = .04.")
```

reports that there is an error somewhere, and that if the degrees of freedom and F value are correct, the p -value should be .097. There is also an app version available at <http://statcheck.io/>. This currently (as of 10 March 2020) is a *beta* version, but it is easier to use for most people than the R package **statcheck**. You enter your pdf, Word, or html file, and the app makes a report for the entire file showing any discrepancies it discovered. The journal *Psychological Science* requires (their 2019 rules) authors of accepted articles to submit an error-less report (https://www.psychologicalscience.org/publications/psychological_science/ps-submissions, 10 March 2020). This seems sensible for *OES* authors to also use, though checking this prior to submission may prevent reviewers finding these errors.

The **statcheck** function will not find all errors. It focuses on comparing test statistics, p -values, and degrees of freedom for different distributions. It will not work if, for example, you are copying a table of means from computer output. It cannot say whether a mean of 3.14 really should be 2.72? There are other options. The R package **knitr** (Xie, 2013, see also yihui.org/knitr) allows you to embed statistics code into your LaTeX word processing and this places the R output into the document. Besides R, it can be used to 40 other software including C, Perl, Python, SAS, SQL, and STATA.

Neither **statcheck** nor **knitr** will prevent all errors. Authors need to be careful, no matter what additional techniques that they use, and reviewers need to be

vigilant. And of course many statistical problems can still occur. Some of these are briefly discussed in the next section.

6 Some Additional Statistical Checks for Reviewers to Consider

Several articles exist for statistical guidelines (e.g., Wilkinson & the Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999; Wright, 2003), including many that have been written with the replication crisis in mind (Munafò et al., 2017; Smaldino & McElreath, 2016). The replication crisis is that many studies, perhaps most (Ioannidis, 2005), do not replicate. This topic could be the subject of a longer paper or even series of papers. Here are some of the topics that methodologists believe are a concern. The purpose of this section is to remind reviewers to consider these.

6.1 Too Few Participants

Power analysis (Cohen, 1992) is one method that can be used to help decide how many participants to have in your study. A popular free package is G*Power (Faul et al., 2007). Power is low in many disciplines (e.g., Button et al., 2013). Low power means that there is a high likelihood that nothing meaningful will be found. Low power also increases the likelihood that published studies will report effect sizes that are too large, called Type M (for magnitude) errors, and significant effects in the wrong direction, called Type S (sign) errors (see Gelman & Carlin, 2014). There are issues with how power is used (e.g., Baguley, 2004). The most difficult aspect of power analysis is deciding the minimum effect size that you want to design your study to detect. If your power analysis suggests a larger sample than you can realistically attain, consider changing your design (e.g., using more reliable instruments, a within-subject design, collaborating with others).

Power analysis is not the only way to guide choices about the sample size. In education it is often what data are available. Often this means a large amount of data, like all test scores in a state. Sometimes it means that only a small amount of data is available. It is important that researchers take this into account prior to deciding whether to conduct the study. If they still decide to conduct a study the design of the study may need to be altered. Small numbers of participants are appropriate for some designs if

much data are collected for each participant. For example, psychophysics studies sometimes have a small number of participants do hundreds of tasks. Similarly, education design can involve fairly small numbers of participants answering hundreds of questions, reading hundreds of lines of text, or the researchers use the log files from lengthy online tasks from a small number of participants. The important assumption for these approaches is that their is little inter-individual variability. If this assumption is not plausible, larger samples are needed.

6.2 *p*-fishing

One of the biggest concerns about the reliability of published results is that many researchers try too hard to discover patterns in their data. This is called *p*-fishing. One goal of statistics is to discover patterns in data, but just as patterns emerge if you stare at a cloud long enough, patterns emerge from data sets if you try enough statistical procedures. This goes beyond adjusting for multiple comparisons, though this is a first step. Gelman & Loken (2013) likened this to Jorge Borges essay “The Garden of Forked Paths” (translation at: https://archive.org/stream/TheGardenOfForkingPathsJorgeLuisBorges1941/TheGarden-of-Forking-Paths-Jorge-Luis-Borges-1941_djvu.txt). When you conduct analysis you are faced with several choices, some of these are dictated by the data (e.g., the assumptions of a *t*-test not meet), and this is appropriate (Tukey, 1960). However, it is important to tell readers which choices are made.

Several approaches have been put forward to lessen the negative effects of *p*-fishing. Here are a few.

- If you can identify all the potential ways that you could have conducted your analyses, you could:
 1. Adjust *p*-values with traditional multiple comparison or false discovery rate procedures (Bretz et al., 2010) or
 2. Conduct all these analyses, weight them if appropriate, and see how many are “significant” (Steenen et al., 2016).
- If you know there are a lot of paths, but you are not sure how many, you could adopt a lower *p*-value to declare that the null hypothesis can be rejected. This approach is advocated by some physicists for discoveries, the so-called 5-sigma rule (Lyons, 2013), but not for other empirical findings. Some social scientists advocate using a lower threshold than 5% for all research (Benjamin et al., 2018).
- If you can state what analyses you will do before conducting them, you can pre-register your study.

This can be done at pre-registration cites, like cos.io/prereg/ (COS is the Center for Open Science), or many journals allow you submit plans about your study and analyses. Your plans go through peer review and the journal can decide tentatively to accept the paper. Providing that you do what you say you will do and adequately justify any changes to your plans, they should publish the final piece. *OES* does not currently do this. However, if you have pre-registered your study on a cite like cos.io/prereg/ you should say this in your manuscript and reviewers will likely look favorably upon this.

- The journal *Basic and Applied Social Psychology* banned the use of *p*-values and some related methods. This did not appear to have the desired effect. Fricker et al. (2019) analyzed the first year of articles after the ban and conclude:

We found multiple instances of results seemingly being overstated beyond what the data would support if *p*-values (or some other form of statistical inference) had been used. Thus, the ban seems to be allowing authors to make less substantiated claims, the net result of which could be misleading to readers and could well result in a body of literature within which there are a greater number of irreproducible effects. (p. 383)

- Save *p*-values for the really important hypotheses in your research. Popper ([1959] 2002) describes how he thought empirical research should progress. Bold hypotheses are put forward and evaluated. If *p*-values are considered one of the ways to evaluate empirical results (and there are others), then these should be reserved for a small number of tests for bold conjectures. Sometimes *p*-values are used for not-bold-conjectures, like the reliability of a test being above zero. A hypothesis test should not be conducted unless a reasonable person could believe that the hypothesis were true. It might be useful to have authors provide the rationale for each *p*-value.
- There is much discussion of how greater use of effect sizes (Lipsey et al., 2012), confidence intervals (Cumming, 2014), and Bayesian approaches (Gelman et al., 2014) are better than solely relying on *p*-values. We encourage alternative and complementary approaches, but none of these, on their own, is a *panacea* for the replication crisis. That said, effect sizes should be always be reported, ideally with some kind of interval to show the uncertainty in the estimate.
- Avoid HARKing, or Hypothesizing After Results are Known (Kerr, 1998). Suppose you collect some data, observe an interesting pattern, and then explain

this pattern. This explanation is HARKing. Saying that the data provide evidence for the explanation is different from saying the explanation provides an account for the data. There are likely several possible explanations for the data. The theory should come first, or if created from the data the readers should be told that it is speculative. HARKing is a way to fool yourself into believing an explanation has been confirmed with data when it has not. Most cases of HARKing do not involve the researcher deliberately trying to fool readers.

The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists.

Feynman (1974, p. 12)

- The most important recommendation is not treat a *p*-value as definitive evidence of anything. If you are comparing two groups, a significant *p*-value only allows you to know the likely direction of the effect, and it depends on several assumptions that are likely not valid (e.g., random sampling). The limited information from a *p*-value should be treated as such.

7 Responsibilities of Reviewers

The peer review process relies on the recommendations of reviewers. Editors should not prescribe details of how to review; reviewers are the experts. However, it seems prudent to provide some guidance related to data and statistics. This is common for medical journals (e.g., Altman, 1998; Greenwood & Freeman, 2015), because often the medical researchers have relatively little training in statistics because biostatisticians are often on research teams. All education researchers should receive much methods/analysis training, so this is less often the case in education research. However, with new statistical techniques being developed, no one is an expert in every technique. This is a partial list, but covers the main things to think about.

1. Is the description of the methods and analyses understandable to non-specialists? Do the authors make clear why these methods and these analyses were used?
2. Are psychometric properties of any tests reported. These statistics should be both from the test developers (include the sample they used to find these) and for the sample reported in the paper. Intervals should

be included for these estimates, providing the test publishers report these.

3. Do the authors describe the sampling procedure and characteristics? Do they describe how they decided upon the sample size? Do they discuss the response rate, if appropriate, and if this creates any bias?
4. Are assumptions described, particularly if missing values are imputed?
5. Are Tables and Figures clear? Are units presented in these?
6. If covariates are used in any analyses, are the reasons for these discussed? Sometimes researchers include covariates in their models just because these variables exist in their data set, and this can create numerous problems for interpreting results (e.g., Meehl, 1970).
7. Have authors changed continuous or scale variables into categorical variables without justification?
8. If the authors describe allocating participants to groups, was this done randomly, and was this process described?
9. Are there *p*-values for null hypotheses that no reasonable person would have thought could have been true? Each null hypothesis does not need to be a bold conjecture (Popper, [1959] 2002), but it should not be trivial.
10. If there were multiple hypothesis tests, was this accounted for?
11. Do the conclusions follow from the data?
12. If the data are clustered, like students within classrooms, was this taken into account?
13. Are effect sizes given, and is the uncertainty of these estimates shown (e.g., with confidence or credibility intervals)?
14. Does it appear that HARKing has occurred (Kerr, 1998)?

Most of these will not result in a manuscript being rejected, but they may lead to the authors needing to provide further information. If when reading the paper, if you feel that you or the editor associated with that paper should examine evidence not presented in the paper, tell the editor this. Also, it is important to tell the editor if you are not a specialist for a technique and say if the editor should seek a specialist to comment on this aspect. In many cases the editor will have chosen a methodology reviewer if a new or advanced technique was used in the manuscript.

It is worth stating that if reviewers are advised to examine papers with this list in mind, authors should consider them also.

8 Summary

Miyakawa's (2020) findings show clear evidence of a problem. Disciplines vary in how severe the trust problem is, but all disciplines have room for improvement. It is important that journals help the public trust the findings they report. As a new journal *OES* is developing procedures to help this. The journal takes the OPEN in the title seriously. Yes, the journal is open to all readers who have access to the internet. But ideally open should also mean the public has the opportunity to understand the evidence upon which any claims are made. Sometimes there are reasons why some of this information cannot be shared (e.g., a proprietary test was administered), but encouraging the use of the supplementary materials option can help this. We would also like to encourage discussion post-publication. This is something that is less common in the traditional academic journals, and ideas on how best to facilitate this—beyond just having a comment tab next to the publications on the website—are welcomed!

Author Statement: The Author is a senior editor of Open Education Studies. No funding was received for the article. This commentary was improved by reviewer comments arranged by the journal, but did not go through formal peer review.

References

- Altman, D. G. (1998). Statistical reviewing for medical journals. *Statistics in Medicine*, 17, 2661–2674. doi: 10.1002/(SICI)1097-0258(19981215)17:23<2661::AID-SIM33>3.0.CO;2-B
- Baguley, T. S. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, 35, 73–80. doi: 10.1016/j.apergo.2004.01.002
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2. doi: 10.1038/s41562-017-0189-z
- Bretz, F., Hothorn, T., & Westfall, P. (2010). *Multiple comparisons using R*. Boca Raton, FL: CRC Press.
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365–376. doi: 10.1038/nrn3475
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. doi: 10.1037/0033-2909.112.1.155
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29. doi: 10.1177/0956797613504966
- Epskamp, S., & Nuijten, M. B. (2018). *statcheck*: Extract statistics from articles and recompute *p* values [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=statcheck> (R package version 1.3.0)
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Feynman, R. P. (1974). Cargo cult science: Some remarks on science, pseudoscience, and learning how to not fool yourself. Caltech's 1974 commencement address. *Engineering and Science*, 37 (7), 10–13.
- Fricker, R. D., Jr, Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in *Basic and Applied Social Psychology* after their *p*-value ban. *The American Statistician*, 73, 374–384. doi: 10.1080/00031305.2018.1537892
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9, 641–651. doi: 10.1177/1745691914551642
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from <http://www.stat.columbia.edu/~gelman/research/unpublished/forking.pdf>
- Greenwood, D. C., & Freeman, J. V. (2015). How to spot a statistical problem: advice for a non-statistical reviewer. *BMC Medicine*, 13 (270). doi: 10.1186/s12916-015-0510-5
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2 (8). Retrieved from e124. doi: 10.1371/journal.pmed.0020124
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2, 196–217. doi: 10.1207/s15327957pspr0203_4
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. (NCSE 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf>
- Lyons, L. (2013). Discovering the significance of 5 sigma. *arXiv*(1310.1284).
- Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: Vol IV. Analysis of theories and methods of physics and psychology* (pp. 373–402). Minneapolis, MN: University of Minnesota Press.
- Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain*, 13. doi: 10.1186/s13041-020-0552-2
- Molinari, A., & Gasparini, A. A. (2019). When students design university: a case study of creative interdisciplinarity between design thinking and humanities. *Open Education Studies*, 1.1. doi: 10.1515/edu-2019-0002
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A.

- (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1 (21). doi: 10.1038/s41562-016-0021
- Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48, 1205–1226. doi: 10.3758/s13428-015-0664-2
- Popper, K. ([1959] 2002). *The logic of scientific discovery*. Milton Park, UK: Routledge.
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society: Open Science*(3160384). doi: 10.1098/rsos.160384
- Spiegelhalter, D. J. (2019). *The art of statistics*. New York, NY: Basic Books.
- Steen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702–712.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 448–485). Stanford, CA: Stanford Univ. Press.
- Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. doi: 10.1037/0003-066X.54.8.594
- Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, 73, 123–136. doi: 10.1348/000709903762869950
- Wright, D. B. (2019). Treating rapid responses as incorrect for non-timed formative tests. *Open Education Studies*, 1, 56–72. doi: <https://doi.org/10.1515/edu-2019-0004>
- Xie, Y. (2013). *Dynamic documents with R and knitr*. Boca Raton, FL: Chapman and Hall/CRC.