Improving how Scientific Results are Interpreted

There is a crisis in the educational, behavioral, and social sciences. Many high-profile studies do not replicate (e.g., Camerer *et al*, 2018). Discussion of this replicability crisis comes at a time when the public are increasing skeptical of science and this has led to adoption of procedures without scientific backing and failure to comply with guidelines that do have scientific backing. This has had grave consequences.

There are several reasons why many studies do not replicate and, related to these, there are reasons why some people gain a false belief in the validity of individual findings. Failure to replicate makes people distrust science as a whole. Reasons include fraud (e.g., Wakefield on vaccines, see Deer, 2020), misunderstanding about how to interpret frequentist probabilities (e.g., Colquhoun, 2017; Oakes, 1986), picking and choosing how to analyze data (Gelman & Lokken, 2014; Steegen *et al.*, 2016), creating hypotheses to fit the results (Kerr, 1998), and inappropriate thresholds for supporting (and rejecting) hypothesis (Benjamin *et al.*, 2018). Here the focus is on a failure of researchers to differentiate results from exploratory hypotheses from confirmatory hypotheses, and if labelling the results from these highlights how each should be interpreted by researchers and policy makers. This would allow the potential for researchers to use scientific findings more appropriately.

The Problem: Too Many Hypothesis Tests Taken Too Seriously

There are two related issues that together have created the current crisis. First, the results from any single study may not replicate. There are practices (e.g., more powerful designs, transparency, placing data on public archives, pre-registration) that can increase the likelihood of replication (Wright, 2020), but integral to the scientific process is that some findings will not replicate and this must be understood by those using scientific results. Ideally science would self-correct these problems through critical, often post publication, review but this does not always occur (Feyerabend, 1978, Ioannidis, 2012). The second issue is that some people believe a "significant" result means something definitive and ground-breaking. The choice of the word "significant" in its statistical sense is poor and it misleads readers. Coupled with social media and press officers' desires to create click-bait out of any results, the public more often hear sensationalized findings and may believe whatever claims are reported. It is important for those reporting research not to over-state their claims. Feynman summed this up nicely when discussing scientific integrity:

> I'm talking about a specific, extra type of integrity that is not lying, but bending over backwards to show how you're maybe wrong, that you ought to do when acting as a scientist. And this is our responsibility as scientists, certainly to other scientists, and I think to laymen. (1974, p. 12)

The typical press release does not "bend over backwards" to show how the reported results may not be accurate.

This project is based on the core problem that there are too many undifferentiated hypotheses being conducted and that without differentiation each is treated as definitive evidence of something. Spiegelhalter, both in his discussion (Spiegelhalter, 2017a) of the American Statistical Association's statement on the problems with *p* values and this crisis (Wasserstein &

Lazar, 2016) and in his presidential address to the Royal Statistical Society (Spiegelhalter, 2017b) briefly suggested a simple change that could be beneficial. It is assumed in this proposal that hypotheses are evaluated with *p*-values, though the same labelling could be used for Bayes Factors and other methods for hypothesis evaluation. He suggested that the results from exploratory hypotheses could be labelled p_{exp} and the results from confirmatory hypotheses labelled p_{con}. While some people argue null hypothesis testing of exploratory hypotheses is inappropriate, Spiegelhalter notes that often the findings of exploratory hypothesis can be quite interesting--the science fiction writer Isaac Asimov said: "the most exciting phrase in science is not 'eureka!' but 'that's funny' " (https://quoteinvestigator.com/2015/03/02/eureka-funny/). Further, Spiegelhalter notes that the use of *p* values is so prevalent that banning them would be difficult. And, when *p* values have been banned this has led to researchers over-stating the evidential value of their findings (Fricker *et al*, 2019).

Scientists viewing the results of exploratory hypotheses should do so with appropriate skepticism. They may find them intriguing and worthy of further investigation but should be cautious. Spiegelhalter argues that while p_{exp} results should not be removed from scientific discourse, they should not be discussed in abstracts or conclusions. They should also not be part of press releases or become the basis of high stakes decision making. p_{con} results, however, can become part of this wider discourse, albeit with many of the caveats that have been discussed with respect to the crisis in science considered. These p_{con} results have more evidence for them prior to the study than p_{exp} results.

Addressing: Identify or test strategies for producing more useful research evidence.
Major research question:
Can a simple way in which results are reported improve how educators interpret the results?
Anticipated finding:
Results labeled p_{con} will be viewed as more credible than those labeled p_{exp}.

The Proposal

There are two phases for the proposed research. The first will determine how common p_{exp} and p_{con} are in psychology (or perhaps social science more broadly) research. We will focus on empirical research where there are result sections that address the veracity of specific claims. Any individual study may produce dozens of such hypotheses. Researchers will identify these and label each with p_{exp}, p_{con}, or unable to tell. They will record where they are reported and how (e.g., what test, the degrees of freedom, if the *p* value was adjusted for multiple comparisons). For p_{con} results they also differentiate whether the hypothesis being tested is the predicted by theory or whether the theory predicts the model being rejected. The former is often considered the better scientific approach (e.g., Popper, 2002), but the latter is common in the educational, social, and psychological sciences (Meehl, 1967). Coders will also be allowed to use an "unable to tell" code for this classification. The coders will be urged not to use these final options for

either classification unless necessary. The reliability of both the identification of hypotheses and the classifications will be estimated.

The identifications will include traditional *p* values, but also may be for results based on Bayes Factors, descriptive statistics, effect sizes, and qualitative statements. This will be a time-consuming task and take most of the first year of the project. Several empirical journals will be used to allow us to compare how often exploratory versus confirmatory results are reported. The "unable to tell" hypotheses will also be discussed. If the way an article is presented does not allow the type of hypothesis to be determined, this is poor for communicating results. We will track how often statements about each of these are reported in the abstract.

The second phase of the project is to see if people are able to differentiate the evidential value of results labelled p_{exp} and p_{con}. Materials will be put together for several short research summaries. How hypotheses are labelled will be counter balanced. Participants will be asked about each hypothesis, including the probability of replication, how likely they would alter their behavior based on the result, and, when the result relates to a particular intervention/product, how likely they would use the intervention/product.

Two studies will be conducted in this phase. The first is an in-person "think aloud" study (Ericsson & Simon, 1980). The difference between p_{exp} and p_{con} will be introduced to each participant. They will read the summaries and describe what they are thinking reading through the summaries. They will be asked the questions about the results and told to free respond their thoughts. This will be audio recorded, transcribed, and the content analyzed for differences in responses based on which type of hypothesis was presented (Benoit *et al*., 2018). This study will be conducted in the PI's laboratory at UNLV (the Assessment and Quantitative Analysis [AQUA] lab). Approximately 25 participants will be recruited.

The second study will be a larger quantitative study. The summaries and questions will be displayed via Qualtrics (UNLV has a site license) and this will be distributed to educators volunteering for $10 for the approximate 30-minute study. They will be introduced to the two types of *p* values, read the summaries, answer close-ended questions about the hypotheses, and then be given the opportunity to respond in free format about their views on these two types. The aim is to collect data from 265 participants. This is based on a power analysis with $d_z = .2$ (Cohen's [1992] small effect), power of .90, and $\alpha = .05$. As there are multiple summaries and multiple people, a cross-classified multilevel model will be used for each question (Wright & London, 2009). This is an efficient means to assess whether changing the way *p*-values are reported affects how educators interpret results.

Potential Impact

Efficient use of research requires users to differentiate research more likely to replicate than research less likely to replicate. Numerous journal articles and society statements describe how the over-reliance on *p*-values is problematic. *p*-values are often misused and often used in the wrong situations. The suggestion by Spiegelhalter (2017a,b) to differentiate the results by whether they were of exploratory or confirmatory hypotheses is potentially valuable, but it is necessary to examine how often the different types would be used (phase 1) and what effects this manipulation would have on readers (phase 2). The proposed research will begin by examining education journals to see how often each of these types is used. Next, experimental studies will

examine if using Spiegelhalter's notation for *p*-values does allow researchers to differentiate the evidentiary status of the different results.

References

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* . 10.1038/s41562-017-0189-z

Benoit, K., Watanabe, K., Wang, H, Nulty, P, Obeng, A., Müller, S., Matsuo, A. (2018). **quanteda**: An R package for the quantitative analysis of textual data. *Journal of Open Source Software 3*(30), 774.

Camerer, C. F. *et al. (2018).* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2 (9): 637-644. 10.1038/s41562-018-0399-z*

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. 10.1037/0033-2909.112.1.155

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society: Open Science, 4* (12): 171085. 10.1098/rsos.171085

Deer, B. (2020). *The doctor who fooled the world*. Baltimore, MD. The Johns Hopkins Press.

Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review*. *87*, 215–251. 10.1037/0033-295X.87.3.215

Feyerabend, P. (1978). *Science in a free society.* London: NLP.

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, Jr, R. D., Burke, K, Han X., and 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. 10.1080/00031305.2018.1537892

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460-465.

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

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645-654. 10.1177/1745691612464056

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, *2* , 196–217. 10.1207/s15327957pspr0203_4

Meehl, P. E. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science, 34*, 103-115.

Oakes, M. W. (1986). *Statistical inference: a commentary for the social and behavioral sciences*. Hobokon, NJ: John Wiley & Sons.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349(6251)*. 10.1126/science.aac4716

Popper, K. (2002, original 1959). *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. (2017a). Too familiar to ditch. *Significance, 14*, 41.

Spiegelhalter, D. J. (2017b). Trust in numbers. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*, 948-965. 10.1111/rssa.12302

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11* , 702–712.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: context, process, and purpose. *The American Statistician, 70*, 129-133. 10.1080/00031305.2016.1154108.

Wright, D. B. (2020). Improving trust in research: Supporting claims with evidence. *Open Education Studies*, *2*(1), 1-8. 10.1515/edu-2020-0106

Wright, D. B. & London, K. (2009). Multilevel modelling: beyond the basic applications. *British Journal of Mathematical and Statistical Psychology,* 62, 439–456.