Replication Crisis and Critiques of Significance Testing: Some Suggested Readings

Below are a few sources that are worth reading and giving some thought to. Some of these sources are very constructive. Others I think less so. I do not agree with all of the points made in some of these articles (and, at least in my estimation, some of the assertions are incorrect or unfairly represented points), but these readings present a variety of perspectives and are valuable to consider. Readings are organized around two of the several current streams of discussion in psychology about methods and analysis. I use the term "current", but as one looks into the history of statistics, these very same topics have been debated for 100 years.

"Replication Crisis"

One set of discussions has been prompted by reports a few years ago of failures to replicate findings in psychology. Replication failures are not unique to psychology at all. See the *Nature* poll about replication failures across a number of scientific disciplines (https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970), for instance. Within psychology, much controversy has been triggered by the publication of findings from the Reproducibility Project in the journal *Science* (Open Science Collaboration, 2015), indicating that only 39 out of the 97 results reported in several psychology journals in 2008 could be replicated. See a synopsis and update here https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/.

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

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior research methods, 43*(3), 666-678.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one, 9*, e109019.

Lazzeroni, L. C., Lu, Y., & Belitskaya-Lévy, I. (2016). Solutions for quantifying P-value uncertainty and replication power. *Nature methods, 13*(2), 107-108.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices 1. *Educational Psychology, 28*, 151-160.

Rejection of NHST

The other set of articles are part of the perpetual debate which dates back to Fisher and Neyman and Pearson about significance testing. Articles in this arena are never lacking passionate opinions! Be wary in a few cases of what I suspect is cherry-picked evidence to support the author's argument.

Kline, R. (2012). Beyond significance testing. *Washington, DC: American Psychological Association*.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559-569.

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

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

Reports from the APA Task Force on Statistical Inference. http://www.apa.org/science/leadership/bsa/statistical/index.aspx.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods, 12*(3), 179-185.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.