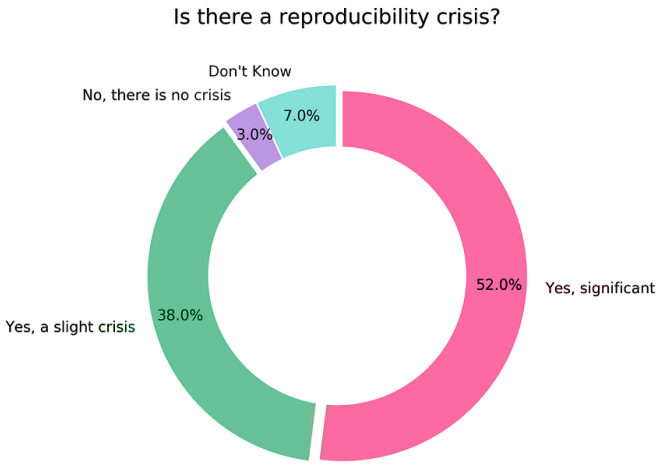
From Scientific Reproducibility To Epistemic Humility

Joseph Wayne Smith and Saxon J. Smith



(Baker, 2016)

1. Introduction: The Reproducibility Crisis

The reproducibility crisis (aka "the replication crisis," aka "the replicability crisis") is a disturbing contemporary theoretical problem whereby, in a large and increasing number of scientific fields, ranging from social sciences such as psychology, sociology, and economics, to natural sciences such as genetics, ecology, and medicine, it has been found that numerous studies have not been able to be reproduced by other researchers and sometimes even by the original research teams that retested their own research (Baker, 2016). The philosophical problem raised here is that the so-called "scientific method," among other things, requires for the empirical sciences that their research be able to be reproduced, hence a failure to be able to do this on a large-scale calls into question the scientificity of the disciplines suffering from this failure of reproducibility.

In this essay we focus on two important recent papers which, we believe, significantly intensify the reproducibility crisis. It has been shown in two distinct studies that scientists can analyze the same sets of data but get significantly divergent results. The two areas where this research was conducted are first, ecology, and second, a social science hypothesis about immigration levels. Both papers show that the conscious and unconscious analytic choices made by researchers saliently shape the outcomes of scientific studies, but even beyond that, there are still outstanding problems in explaining the variation in research results. We'll argue that these results require scientists and philosophers of science to adopt an attitude of *epistemic humility* towards much or even all empirical science.

2. The Ecology Paper

In a recent reproducibility trial in ecology, no less than *246 biologists* got different results from the same data set (Oza, 2023). In this trial, which was conducted by Gould et al., the scientists were given two data sets and two questions. Either: "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" Or "How does grass cover influence *Eucalyptus* spp. seedling recruitment?" With regard to the blue tit study, there was general agreement that sibling competition had a negative impact upon nestling growth. However, there was significant disagreement about the size of the effect. The grass cover research had even greater disagreement. The research is concisely summarized by the authors as follows:

> We used two unpublished datasets, one from evolutionary ecology (blue tit, *Cyanistes caeruleus*, to compare sibling number and nestling growth) and one from conservation ecology (*Eucalyptus*, to compare grass cover and tree seedling recruitment), and the project leaders recruited 174 analyst teams, comprising 246 analysts, to investigate the answers to prespecified research questions. Analyses conducted by these teams yielded 141 usable effects for the blue tit dataset, and 85 usable effects for the *Eucalyptus* dataset. We found substantial heterogeneity among results for both datasets, although the patterns of variation differed between them. For the blue tit analyses, the average effect was convincingly negative, with less growth for nestlings living with more siblings, but there was near continuous variation in effect size from large negative effects to effects near zero, and even effects crossing the traditional threshold of statistical significance in the opposite direction. In contrast, the average relationship between grass cover and *Eucalyptus* seedling number was only slightly negative and not convincingly different from zero, and most effects ranged from weakly negative to weakly positive, with about a third of effects crossing the traditional threshold of significance in one direction or the other. However, there were also several striking outliers in the *Eucalyptus* dataset, with effects far from zero. For both datasets, we found substantial variation in the variable selection and random effects structures among analyses, as well as in the ratings of the analytical methods by peer reviewers, but we found no strong relationship between any of these and deviation from the meta-analytic mean. In other words, analyses with results that were far from the mean were no more or less likely to have dissimilar variable sets, use random effects in their models, or receive poor peer reviews than those analyses that found results that were close to the mean. The existence of substantial variability among analysis outcomes raises important questions about how ecologists and evolutionary biologists should interpret published results, and how they should conduct analyses in the future. (Gould et al., 2023)

The conclusion of the study was that the substantially divergent sets of answers produced by the analysis by the ecologists and evolutionary biologists for the same data sets, indicate that different individuals' data analysis choices, and decisions made in statistical analyses can generate radically different outcomes. This result is

consistent with other published results indicating the same analytic variability in the quantitative social sciences (Fanelli et al., 2017; Deressa et al., 2023).

Analytic variability was also evident in the second paper we'll now examine, but with an even more radical upshot.

3. The Immigration Paper

In a paper recently published in *Proceedings of the National Academy of Sciences*, no less than *161 researchers in 73 research teams* used the same data to test the hypothesis that "greater immigration reduces support for social policies among the public" (Breznau et al., 2023). The teams analysed the data, and then submitted not only results but also descriptions of the models they used. The principal meta-research team then examined these models and found that *107 different analytic decisions* were made in the research, such as decisions about the inclusion and exclusion of data and the variables that were included in the studies. The aim was to isolate the "idiosyncrasy of conscious and unconscious decisions that researchers make during data analysis."

It was found that among the different research teams there was "both widely diverging numerical findings and substantive conclusions despite identical start conditions." The diagram displayed directly below this paragraph provides the meta-researchers' representation of their results. The *x*-axis (horizontal) has a ranking of the estimates ranging from the smallest to largest, and the *y*-axis (vertical) gives the size and sign value (i.e., whether they are positive or negative). The colored dashes are the estimates that the teams made. As can be seen, there was substantial disagreement about whether greater immigration reduces support for social policies, with some research teams supporting the hypothesis, others rejecting it, and some with mixed results.

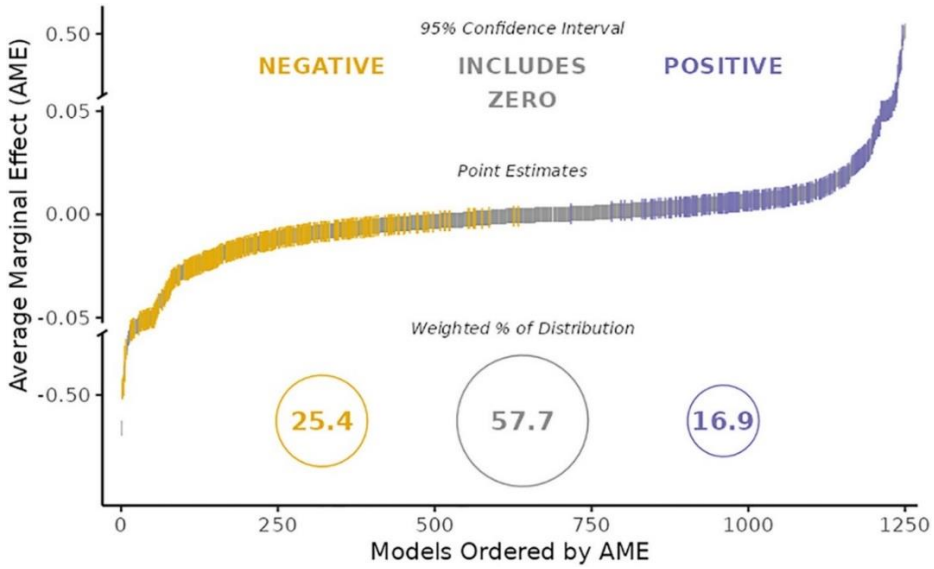


Figure 1 (Breznau et al., 2023)

Correspondingly, the diagram displayed directly below this paragraph represents the attempt by the meta-researchers to examine whether the variation observed was due to analytic decisions made by the different research teams. The factors explaining the variance in results are represented in colors, and the analytic decisions in green. Variation in the results was not explained by the researchers' expertise, expectations, or beliefs; and the choices made in designing statistical tests did not adequately explain the variance. In fact, they concluded: "More than 95 % of the total variance in numerical results remains unexplained even after qualitative coding of all identifiable decisions in each team's workflow. This reveals a universe of uncertainty that remains hidden when considering a single study in isolation."

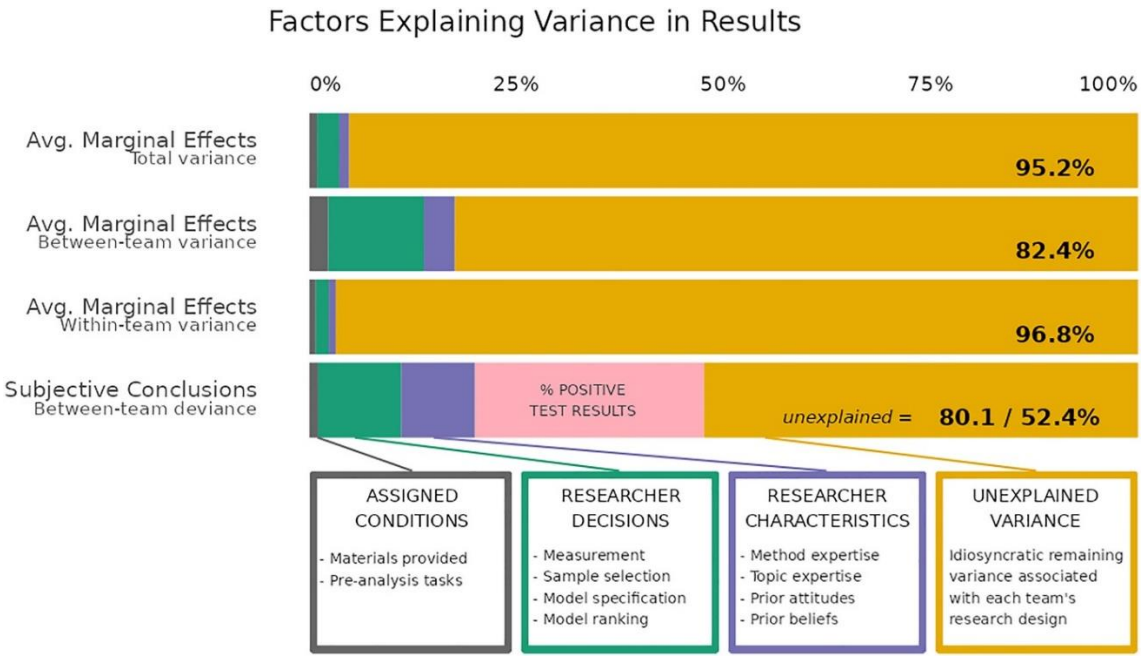


Figure 2 (Breznau et al., 2023)

4. Conclusion: What's Going On Here?

Gould et al., studying the ecology studies' variance, i.e., that there is "substantial heterogeneity due to analytical decisions," concluded that regardless of what produces the variability,

> the existence of such dramatic heterogeneous results when ecologists and evolutionary biologists seek to answer the same questions with the same data should trigger conversations about how ecologists and evolutionary biologists analyze data and interpret the results of their own analyses and those of others in the literature. (Gould et al., 2023)

They proposed that there should be a methodological skepticism about taking single analyses to provide a complete answer to a research question.

But even if one should be in principle skeptical of single analyses, and make definite conclusions only after multiple studies have been done, the question was asked in (Oza, 2023): "how do you know, what is the true result?" Gould et al.'s response is that "part of the solution could be asking a paper's authors to lay out the analytical decisions that they made, and the potential caveats of those choices" (Oza, 2023).

Nevertheless, as has been shown by the immigration studies, there was variability in the research outcomes even while cleaving to the "scientific method" and "state-of-the-art approaches to maximizing reproducibility" (Breznau et al., 2023). As regards the question of detailing explicitly the analytical decisions, as Gould et al. propose should be done, while this can be applied to those decisions that are self-consciously made, and hence known, there are also subtle nonself-conscious decisions that researchers are not aware of, that go "unnoticed as nondeliberate actions following standard operating procedures" (Breznau et al., 2023). These decisions can add up and then collectively operate so as to produce the divergent results.

Breznau et al.'s conclusion is that

> idiosyncratic uncertainty is a fundamental feature of the scientific process that is not easily explained by typically observed researcher characteristics or analytical decisions. (Breznau et al., 2023)

Elaborating on this, they make the cogent observation that a much higher level of uncertainty about research findings exists than was previously thought, and that there is a general need for "epistemic humility." Their conclusion was drawn from a range of empirical studies that used statistical methods and assumptions such as the relevance of the categories of "significant' versus "not significant" data, with a further focus upon variation in significance, which has also been observed to be problematic in debates about the replication crisis and the foundations of statistical methodology (Mathur et al., 2023).

As we're construing it, the attitude of epistemic humility towards empirical science is not an all-out or *destructive* skepticism about empirical science, but instead a measured or *constructive* skepticism that yields a critical awareness of the proper limits and scope of empirical science. Certainly, the reproducibility crisis calls for epistemic humility towards empirical science. Moreover, we also contend that broader issues about *the scientific objectivity* of much or even all empirical science are even more problematic than the issues arising specifically from the reproducibility crisis. We will discuss these broader issues in other essays.

REFERENCES

(Baker, 2016). Baker, M. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533, 7604: 452-454. Available online at URL = <https://www.nature.com/articles/533452a>.

(Breznau et al., 2023). Breznau, N. et al. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." *Proceedings of the National Academy of Sciences* 119, 4: e2203150119. Available online at URL = <https://doi.org/10.1073/pnas.2203150119>.

(Deressa et al., 2023). Deressa, T. et al. "More than Half of Statistically Significant Research Findings in the Environmental Sciences are Actually Not." *EcoEvoRxiv*. Available online at URL = <https://doi.org/10.32942?X24G6Z>.

(Fanelli et al., 2017). Fanelli, D. et al. "Meta-Assessment of Bias in Science," *Proceedings of the National Academy of Sciences* 114: 3714-3719.

(Gould et al., 2023). Gould, E. et al., "Same Data, Different Analysts: Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology." *EcoEvoRxiv*. Available online at URL = <https://doi.org/10.32942/X2GG62>.

(Mathur et al., 2023). Mathur, M. et al., "Variation Across Analysts in Statistical Significance, Yet Consistently Small Effect Sizes." *Proceedings of the National Academy of Sciences* 120, 93: e2218957120. Available online at URL = <https://doi.org/10.1073/pnas.2218957120>.

(Oza, 2023). Oza, A. "Reproducibility Trial: 246 Biologists Get Different Results from the Same Data Sets." *Nature*. 12 October. Available online at URL = <https://www.nature.com/articles/d41586-023-03177-1>.