


# Variation across analysts in statistical significance, yet consistently small effect sizes

Maya B. Mathur<sup>a,1</sup> , Christian Covington<sup>b</sup>, and Tyler J. VanderWeele<sup>c</sup>

Breznau et al. assessed variation in results when 73 research teams analyzed the same dataset to investigate whether greater immigration reduces public support for social policies (1). Breznau et al. reported “massive variation in reported results,” including “widely diverging numerical findings.” These conclusions were based on variation in statistical significance across estimates (e.g., 25% of estimates were significant and negative, 58% were nonsignificant, and 17% were positive and significant). Research teams also reported different subjective conclusions, which closely aligned with whether their estimates were significant.

However, considering variation in effect sizes rather than only variation in statistical significance suggests different conclusions. The point estimates themselves varied minimally across teams, and nearly all estimates were quite close to the null. For example, 90% of standardized estimates were within the range  $[-0.037, 0.037]$ ; that is, 90% of estimates suggested that a one-unit increase in immigration was associated with an increase or decrease in public support of <4% of a SD. While standardized effect sizes can be difficult to interpret and are inherently contextual (2), these effect sizes are more than fivefold smaller than classic benchmarks defining “small” effect sizes (3) and do not seem to support unqualified conclusions of “massive variation.” We would encourage Breznau et al. to grapple with whether, in substantive context, effect sizes falling within this narrow range around the null truly exhibit massive variation.

A well-known problem with exclusive focus on statistical significance is that, for very large datasets, even a very small point estimate can be significant. Breznau et al.’s teams analyzed a dataset containing tens of thousands of observations, which explains the “substantial variation” in estimates’ significance even though the estimates themselves were consistently quite small and close to the null. Given the minimal variation in the estimates, it is also not entirely surprising that little of this variation could be systematically explained by identifiable analytic decisions. If there is little heterogeneity in the true population effect sizes underlying the point estimates, then much of the remaining variation in point estimates may simply be statistical error that would be accurately captured in the estimates’ individual confidence

intervals (CIs). Variation in estimates due to statistical error is not part of a “hidden universe” of decisions but rather is fully captured by appropriate statistical inference, including CIs. One can also use meta-analytic methods to estimate variation in population effect sizes, excluding statistical error (4, 5). This analysis suggests that 90% of standardized population effects would be in an even narrower range around the null of  $[-0.014, 0.014]$  (i.e., effects of < 2% of a SD).

The replication crisis has helped revive long-standing, compelling injunctions to stop degrading evidence into categories of “significant” versus “not significant” (6). Researchers should assess evidence using measures such as effect sizes, CIs, and *P* values treated as continuous values (or Bayesian analogs). Unfortunately, meta-researchers assessing variation in results across analysts or replication studies have themselves sometimes focused entirely on variation in significance; as we have seen here, this practice can lead to potentially misleading conclusions.

## Reproducibility

All data and R code required to reproduce these results are publicly available ([https://github.com/ctcovington/MCV\\_BRWN-22\\_comment](https://github.com/ctcovington/MCV_BRWN-22_comment)).

**ACKNOWLEDGMENTS.** This research was supported by NIH grants R01 LM013866 and R01 CA222147. The funders had no role in the conduct of this research.

Author affiliations: <sup>a</sup>Quantitative Sciences Unit and Department of Pediatrics, Stanford University, Palo Alto, CA 94304; <sup>b</sup>Department of Biostatistics, Harvard University, Boston, MA 02115; and <sup>c</sup>Department of Epidemiology, Harvard University, Boston, MA 02115

Author contributions: M.B.M. designed research; C.C. and M.B.M. performed research; C.C. and M.B.M. analyzed data; C.C. and T.J.V. edited the paper; and M.B.M. wrote the paper.

Competing interest statement: The authors declare competing interests. The authors have organizational affiliations to disclose. M.B.M. is Associate Director of the Stanford University Center for Open and Reproducible Science and is a member of the research advisory boards of Sentience Institute and Greener By Default. T.J.V. has received personal fees from Flerish and from Flourishing Metrics.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution License 4.0 (CC BY).

<sup>1</sup>To whom correspondence may be addressed. Email: [mmathur@stanford.edu](mailto:mmathur@stanford.edu).

Published January 9, 2023.

1. N. Breznau et al., Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2203150119 (2022).
2. P. Cummings, Arguments for and against standardized mean differences (effect sizes). *Arch. Pediatr. Adolesc. Med.* **165**, 592–596 (2011).
3. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 2013).
4. M. Mathur, T. VanderWeele, Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology* **31**, 356–358 (2020).
5. C. Wang, W.-C. Lee, A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Res. Synth. Methods* **10**, 255–266 (2019).
6. B. McShane et al., Abandon statistical significance. *Am. Stat.* **73**, 235–245 (2019).