# A chain as strong as its strongest link? Understanding the causes and consequences of biases arising from selective analysis and reporting of research results[1]

Andrew Gelman[2]

1 May 2023

In policy analysis it is common for there to be selection bias in reporting and publishing. We discuss some potential policy implications of systematic overreporting of positive and statistically significant results which leads to a feedback loop of bad studies, repeatedly spinning off unrealistic optimism followed by disappointment.

Sims et al. (2022) and Simpson (2022) explain how published estimates of effect sizes tend to be overestimates, often by factors of two or even much more. In this commentary, I will discuss how this bias can cause an unfortunate feedback loop of noisy studies and inflated policy prescriptions.

For the reasons considered by the two articles under discussion, we should not be surprised when real-world policy impacts are much smaller than would be expected from a straight reading of the research literature.

This is a big problem in applied statistics in general and education policy in particular. For example, the respected economist James Heckman and his colleagues have claimed (Garcia et al., 2016) that, for a certain early childhood intervention, "The overall rate of return is 13.7% per annum, and the benefit/cost ratio is 7.3." That quoted sentence has two serious problems related to uncertainty and selection bias. To start with, it is an elementary but important error to have written that "the overall rate of return is . . . the benefit/cost ratio is . . ."; each instance of the word "is" in that sentence should immediately have been followed by "estimated at." The next step is to recognize that selection on statistical significance induces biases in these estimates—and, given the small samples, high variabilities, and researcher degrees of freedom in the studies where the estimates came from, the biases could be huge (Gelman, 2017a). The report with its bold claims neither acknowledged this bias nor made any attempt to assess its magnitude.

How can this happen, that respected researchers fail to recognize biases in their estimates on such an important topic? We can attribute some of this to a problem with incentives—if you can get away with presenting biased estimates as the truth, this will make your favorite programs look more effective, presumably leading to a greater chance of adoption—and also to a misunderstanding of statistics, what we might call an ideology, in which the unbiasedness of raw estimates leads researchers to not think about the biases engendered by selection. This attitude has been likened to a chain of reasoning that is believed to be as strong as its strongest [sic] link. In the case of a randomized controlled trial, the strongest link is causal identification,

[1] Discussion of Adrian Simpson, "A recipe for disappointment: Policy, effect size, and the winner's curse," and Sam Sims, Jake Anders, Matthew Inglis, and Hugues Lortie-Forgues, "Quantifying 'promising trials bias' in randomized controlled trials in education," for *Journal of Research in Educational Effectiveness.* We thank the U.S. Office of Naval Research and Institute of Education Sciences for partial support of this work.
[2] Department of Statistics and Department of Political Science, Columbia University, New York.

leading to unbiased estimation—but only if all results are reported. Estimates are highly biased if we consider all the weak links in the chain, including the selection of how to analyze the data and what summaries to report; see, for example, Button et al. (2012) and Gelman (2018).

Here are some potential policy implications of ignoring the biases arising from forking paths and selection in data analysis:

1. Effects of interventions are overestimated, leading to implementations of interventions whose benefit-cost ratios are lower than anticipated, including interventions whose net benefits are negative.
2. Interventions regularly work less well than advertised, leading to disillusionment with the entire process.
3. In a sort of Gresham's Law situation, researchers who *don't* use biased estimates are at a disadvantage, as it's difficult for them to compete with research entrepreneurs who use statistical methods that routinely overestimate effect sizes. To put it another way, the "winner's curse" discussed by Simpson (2022) and Sims et al. (2022) is a plague not on the researchers who regularly promote exaggerated claims but rather on more careful researchers, along with future policymakers and students.
4. Previously published overestimates lead researchers to expect unrealistically large effect sizes, so that they design studies that they believe have 80% power even if the actual power is more in the range of 6%; see discussion in Gelman (2017b).
5. A literature full of overestimates can be summarized to yield a ridiculously overoptimistic estimate of the effect of future treatments, as Szászi et al. (2022) discuss in the context of a flawed meta-analysis of nudge interventions.

The result is a feedback loop of bad studies, repeatedly spinning off unrealistic optimism followed by disappointment, with the Gresham factor making it difficult for realism to break into the cycle. And, as the authors of the two papers under discussion emphasize, all this arises even in the best-case scenario of clean randomized studies of well-defined treatments.

I would also like to add one caution. The two articles under discussion explain the winner's curse in randomized controlled trials. But the same biases occur in observational studies when results are selected based on statistical significance or other criteria that favor large estimates. The problem is not with the randomization, which is perhaps the strongest link in the chain of reasoning; it is with the summaries of results that discard information. Causal identification is a good thing—as long as it does not become an excuse for researchers to ignore issues of measurement, noise, and selection.

## References

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2012). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 365-376.

Garcia, J. L., Heckman, J. J., Leaf, D. E., and Prados, M. J. (2016). The life-cycle benefits of an influential early childhood program. HCEO Working Paper 2016-135, Human Capital and

Economic Opportunity Global Working Group, University of Chicago. https://heckmanequation.org/assets/2017/01/Garcia_Heckman_Leaf_etal_2016_life-cycle-benefits-ecp_r1-p.pdf

Gelman, A. (2017a). How does a Nobel-prize-winning economist become a victim of bog-standard selection bias? Statistical Modeling, Causal Inference, and Social Science blog, 20 July. https://statmodeling.stat.columbia.edu/2017/07/20/nobel-prize-winning-economist-become-victim-bog-standard-selection-bias/

Gelman, A. (2017b). The 80% power lie. Statistical Modeling, Causal Inference, and Social Science blog, 4 Dec. https://statmodeling.stat.columbia.edu/2017/12/04/80-power-lie/

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* 44, 16-23.

Simpson, A. (2022). A recipe for disappointment: Policy, effect size, and the winner's curse. *Journal of Research on Educational Effectiveness*.
DOI: 10.1080/19345747.2022.2066588

Sims, S., Anders, J., Inglis, M., and Lortie-Forgues , H. (2022), Quantifying "promising trials bias" in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*. DOI: 10.1080/19345747.2022.2090470

Szászi, B., Higney, A. C., Charlton, A. B., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., and Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences* 119, e2200732119.