A recipe for disappointment: policy, effect size and the winner's curse.

Adrian Simpson

Journal of Research on Educational Effectiveness

Effect size and policy

Standardized effect size estimates are commonly used by the 'evidence-based education' community as a key metric for judging relative importance, effectiveness, or practical significance of interventions across a set of studies: larger effect sizes indicate more effective interventions. However, this argument applies rarely; only when linearly equatable outcomes, identical comparison treatments and equally representative samples are used in every study.

Even when these assumptions hold, this approach to comparing interventions may not provide the best information to policymakers. Conditional on choosing interventions with larger than average effect sizes, the effect estimates are likely exaggerated, with noisier study estimates potentially inflated above more precise ones. Originally derived from the analysis of sealed bid auctions, this phenomenon is *the winner's curse*.

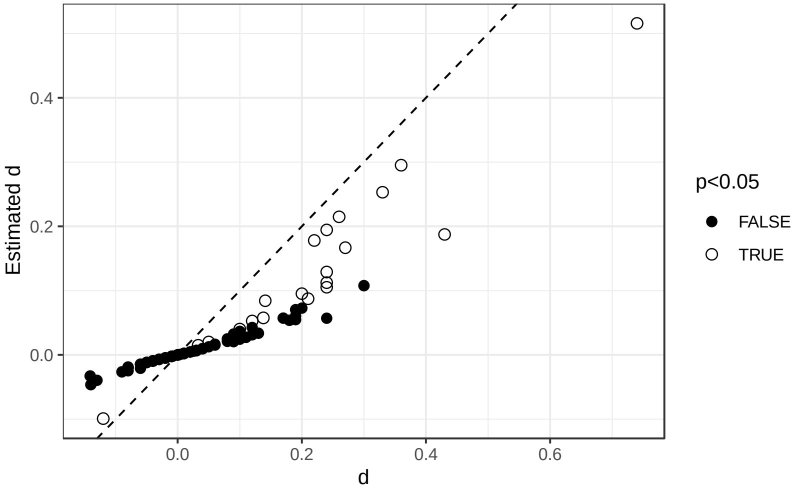
Adjusting for the winner's curse

Recently developed techniques allow adjustment for the winner's curse for a set of studies provided certain properties hold, such as a lack of publication bias or p-hacking. The paper explains the winner's curse and how to adjust for it: modelling the distribution of effect sizes as a mixture of normal distributions then calculating the deconvolution of those with the unit normal distribution (to adjust for the impact of noise). The technique is illustrated on the UK's Education Endowment Foundation (EEF) projects: randomized controlled trials of education interventions conducted with a high degree of transparency.

The results suggest barely significant results in the EEF set might be adjusted for the winner's curse by shrinking estimates by a factor of around 2.5. In addition, barely significant positive effect sizes have around a 16% chance of resulting from negative latent effects – far greater than the 5% (or less) chance one would expect when looking at a given effect estimate alone. Moreover, while the EEF studies commonly aim for 80% power to detect the researcher's intended effect size, the latent effects appear much smaller than those used in the power analyses, suggesting less than 6% of studies achieve 80% power for those latent effects.

The figure below illustrates the impact of adjusting for the winner's curse on this set of studies. Not only are positive effect sizes likely inflated, there are also instances of order reversals: studies that rank higher than others based on the original study effect size, rank lower when adjusted to account for estimation error.

In general, unless adjustment is made, when policy makers select an intervention based on higher effect size estimates from sets of potential interventions, they will likely suffer the winner's curse and be disappointed by subsequent policy outcomes.



Notes: The published effect size (d) and estimated latent effect size (δ) for Education Endowment Foundation (EEF) intervention studies, with results originally flagged as statistically significant indicated.

Full Article Citation:
Simpson (2022). A recipe for disappointment: policy, effect size and the winner's curse. *Journal of Research on Educational Effectiveness*. DOI: https://doi.org/10.1080/19345747.2022.2066588.