Lurking inferential monsters?
Quantifying selection bias in non-experimental evaluations of school programs

Ben Weidmann & Luke Miratrix

1. Background/Context

School systems are awash with observational data, but what can we do with these data? As Edward Leamer noted, the trouble with non-experimental data is that "there is no formal way to know what inferential monsters lurk beyond our immediate field of vision" (Leamer, 1983, 39). The threat of such monsters – the possibility that students who are involved in programs differ in systematic, unobserved ways from other students – has resulted in widespread distrust of causal claims based on observational data (Gargani, 2010).

We aim to evaluate this sense of distrust. While there is no direct test for the Conditional Independence Assumption (CIA), in some instances it is possible to examine whether observational methods can replicate randomized controlled trial (RCT) results by performing a within-study comparison where we compare an experimental estimate (the 'causal benchmark') to an observational analysis that uses the same treatment group (see, e.g., LaLonde, 1986 and later Shadish, 2008 for an earlier example. Also Pohl et al., 2009 or Chaplin et al. 2018). What varies across the two estimates is how the control group is selected: the RCT uses control units that are chosen at random, while the observational study uses a set of non-randomly-selected comparison units (St. Clair, Cook, & Hallberg, 2014). The difference between estimates from the RCT and the observational study can be interpreted as a measure of selection bias (Wong, Valentine, & Miller-Bains, 2017), which is the possible confounder that can destroy the integrity of an observational study.

Purpose/Objective/Research Question:

In this work we do two things. First, we present 14 new within-study comparisons to substantially expand the empirical evidence on selection bias in school evaluations. These new interventions – representing a wide range of school programs from 1-to-1 maths tutoring, to teaching children how to play chess – diversify the pool and extend this area of scholarship beyond the United States.

Second, we present an approach of collectively analysing multiple within-study bias estimates using a meta-analysis framework to investigate the *distribution* of bias in our context. Meta-analysing multiple bias estimates provides a powerful test of whether unobservable factors systematically bias observational evaluations and is central to addressing our core research question: **what is the typical magnitude of selection bias in school evaluations that rely on the CIA?**

Setting and Data:

Our analysis relies on a unique set of linked databases in England. The key data source is an archive of RCTs maintained by the Education Endowment Foundation. Many of the RCTs in the archive can be linked to the National Pupil Database, a census of publicly-funded schools and their pupils.[1]

Research Design:

We perform within-study comparisons on 14 programs. The outcomes we examined were standardised academic achievement at the end of grade 6. For each intervention and outcome, we generate two estimates of bias. The first is a naïve estimate, calculated as a simple difference in means between the experimental control group (CT) and the entire population of potential matches (CO). The magnitude of naïve bias $|\hat{\beta}^{Naive}|$ provides an initial indication of how big the issue of selection bias might be (Wong, Steiner, & Anglin, 2018). Second, we estimate bias *after* conditioning on detailed set of covariates, including a pre-treatment measure of academic performance, giving $\hat{\beta}^{Match}$. The estimation approach we test is 1:1 matching at the school level (see, e.g, Rosenbaum (2000) for an overview). Our simple method was chosen because we wanted an approach that reflected the current state of practice for typical non-experimental evaluations of school programs.

Method for Cross-Study Assessment of Bias

We cannot examine the distribution of raw estimates of selection bias as these distributions reflect both bias and sampling variation. Even if there were no bias to correct, estimates of $\boldsymbol{\hat{\beta}^{Match}}$ would still be non-zero due to sampling error. We therefore design and use a meta-analysis framework (e.g., Higgins, 2009) and testing procedure to explicitly account for sampling variation. This addresses two overlapping goals: to present estimated distributions for β^{Match} and β^{Naive} that are not over-dispersed due to sampling error, and to estimate the typical value of underlying selection bias for our setting.

Observed bias estimates $\hat{\beta}_{kw}$ are assumed to be made up of several components:

$$\hat{\beta}_{kw}|\beta_{kw} \sim N(\beta_{kw}, \sigma^2_{kw})$$
$$\beta_{kw} \sim N(\nu, \tau^2)$$

Where:

- $\nu =$ the mean bias across all interventions and outcomes
- $\beta_{kw} =$ the underlying selection bias for outcome k in intervention w. This has a variance of τ^2, capturing that β_{kw} might change due to context, nature of the program, and so on.
- Observed bias $\hat{\beta}_{kw}$ deviates from underlying bias β_{kw} with a variance of σ^2_{kw}.

Results

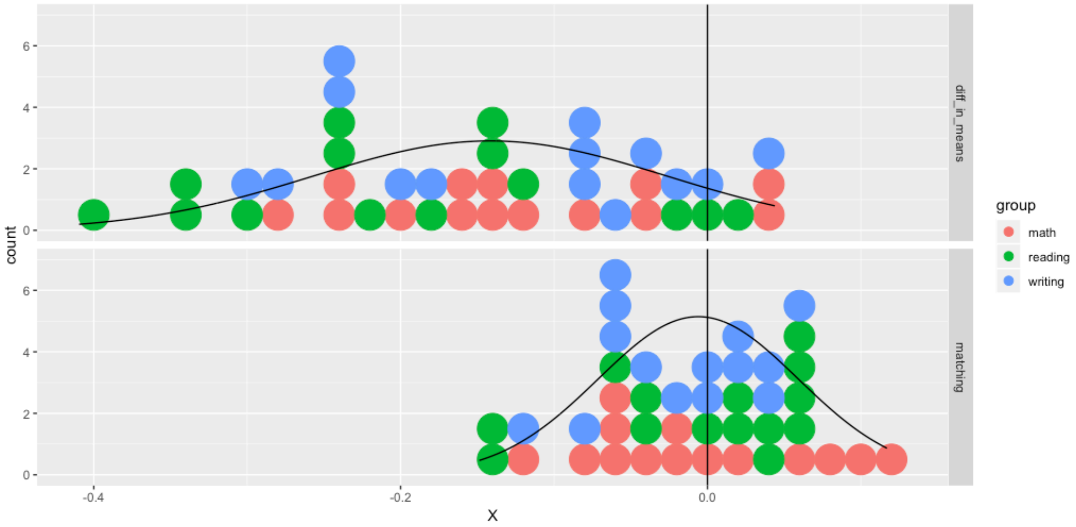
For each intervention-outcome pair we calculate a constrained empirical Bayes' estimate of selection bias, $\tilde{\beta}_{kw}$ (Weiss et al., 2017).[2] Results are on Figure 2, representing the core output of this study: this is our best guess at the distribution of underlying selection bias due to unobserved

[1] This represents over 90 percent of English school children (Department for Education (UK), 2015).

[2] The empirical Bayes estimates are rescaled so that the empirical distribution of $\tilde{\beta}^{Match}$ ultimately has the variance $\hat{\tau}^2$. Failure to constrain the estimates in this way would lead to a distribution that was too narrow (Weiss et al., 2017).

characteristics. The mean of the estimated $\tilde{\beta}^{Match}$ distribution is $\hat{\nu}^{Match} = -0.007\sigma$, indicating that across studies and outcomes the average bias tends to be very close to zero. More importantly, the mean absolute value of $\tilde{\beta}^{Match}$ is 0.03σ, suggesting that the typical magnitude of bias is relatively small. Last, we note that all the estimates of $\tilde{\beta}^{Match}$ are less than 0.11σ.

Figure 2 - estimated distribution of $\tilde{\beta}^{Naive}$ (top) and $\tilde{\beta}^{Match}$ (bottom)

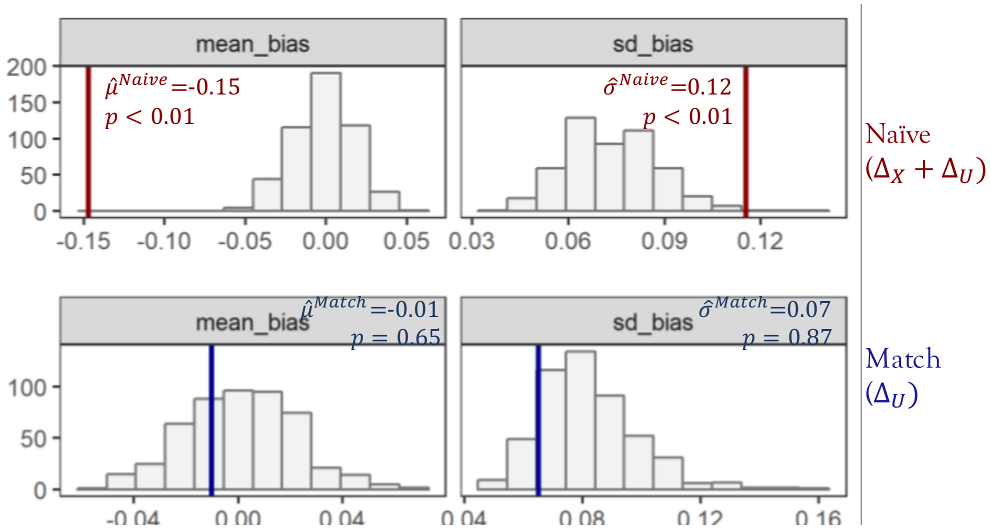


To test the null hypothesis that selection bias is zero across all the studies and outcomes we examine, we used a non-parametric inference procedure.[3] The core idea of this test is to replicate our entire analysis many times in a simulated world where, by design, the only differences between RCT control schools (CT) and comparison schools (CO) is due to sampling variation. Figure 3 presents the reference distributions from this process for four parameters under the null: $\hat{\mu}^{Naive}$ and $\hat{\sigma}^{Naive}$ (top row); $\hat{\mu}^{Match}$ and $\hat{\sigma}^{Match}$ (bottom row).

We find evidence to reject the null that $\hat{\beta}^{Naive}$ estimates are all zero. The observed mean ($\hat{\mu}^{Naive} = -0.15\sigma$) and the observed spread ($\hat{\sigma}^{Naive} = 0.12$) are very unlikely under the null, with p<0.01. In contrast, we find no evidence that that the observed values of $\hat{\mu}^{Match}$ and $\hat{\sigma}^{Match}$ are inconsistent with their reference distributions under the null.

[3] Bias estimates within each intervention are correlated in a complex way. This motivated our use of a simulation-based approach to inference.

Figure 3 – Null hypothesis testing for mean and sd of $\hat{\beta}^{Naive}$ (top) and $\hat{\beta}^{Match}$ (bottom)



6. Conclusion

Across 14 within-study comparisons in UK schools we did not find evidence of substantial selection bias. A meta-analysis of the estimates suggests that, net of sampling variation, the mean absolute value of underlying bias is 0.03σ, with a mean of -0.007σ. None of the estimates of are larger than 0.11σ. While tempting to conclude that observational and experimental evaluations in schools will tend to produce substantively similar results, this conclusion is subject to several caveats. For example, we note that our results may not generalise to other times or contexts.

Our findings cannot disprove the basic truth that in any observational study there may be "inferential monsters" that invalidate causal conclusions. However, in 14 cases when we were able to search for the influence of these monsters, we found little trace of them. So, while RCTs will continue to be invaluable in building a reliable evidence base for school education, we argue that well-designed observational evaluations can, and should, make more of a contribution.

References

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, R. E. (2018). The Internal And External Validity Of The Regression Discontinuity Design: A Meta- Analysis Of 15 Within- Study Comparisons. Journal of Policy Analysis and Management, 37(2), 403-429.

Department for Education (UK). (2015). Statistical first release: Schools, pupils and their characteristics: January 2015 (Ref: SFR 16/2015). *National Statistics*. Retrieved from https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2015. Accessed October 1, 2019.

Gargani, J. (2010). A welcome change from debate to dialogue about causality. *American Journal of Evaluation, 31*(1), 131-132. doi:10.1177/1098214009357612

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re- evaluation of random- effects meta- analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(1), 137-159.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 604-620.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review, 73*(1), 31-43.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. Educational evaluation and policy analysis, 31(4), 463-479.

Rosenbaum, P. R. (2010). Design of observational studies. New York.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. Journal of the American statistical Association, 103(484), 1334-1344.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, *35*(3), 311-327. doi:10.1177/1098214014527337

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843-876. doi:10.1080/19345747.2017.1300719

Wong, V., Steiner, P., & Anglin, K. (2018). What can be learned from empirical evaluations of nonexperimental methods? *Evaluation review, 42*(2), 147-175. doi:10.1177/0193841X18776870

Wong, V. C., Valentine, J. C., & Miller-Bains, K. (2017). Empirical performance of covariates in education observational studies. *Journal of Research on Educational Effectiveness, 10*(1), 207-236. doi:10.1080/19345747.2016.1164781