Abstract Title Page

Title: Asymdystopia: The threat of small biases in evaluations of education interventions that need to be powered to detect small impacts

Authors and Affiliations:
John Deke, Mathematica Policy Research
Thomas Wei, Institute of Education Sciences
Tim Kautz, Mathematica Policy Research

This paper was supported under the U.S. Department of Education, Institute of Education Sciences' Independent Review and Evaluation for Regional Educational Laboratories project (contract ED-IES-12-C-0083).

Abstract Body

Research Context:

Evaluators of education interventions increasingly need to design studies to detect impacts much smaller than the 0.20 standard deviations that Cohen (1988) characterized as "small." For example, an evaluation of Response to Intervention from the Institute of Education Sciences (IES) detected impacts ranging from 0.13 to 0.17 standard deviations (Balu et al. 2015), and IES' evaluation of the Teacher Incentive Fund detected impacts of just 0.03 standard deviations (Chiang et al. 2015).

The drive to detect smaller impacts is in response to strong arguments that in many contexts, impacts once deemed "small" can still be meaningful (Kane 2015). Hill et al. (2008) and Lipsey et al. (2012) suggest multiple substantive benchmarks for assessing what a "meaningful" impact would be for a given intervention and context. These benchmarks often suggest that impacts less than 0.20 standard deviations are meaningful. For example, under the cost-effectiveness benchmark, smaller impacts may be deemed meaningful when evaluating less-expensive interventions.

Though based on a compelling rationale, the drive to detect smaller impacts may create a new challenge for researchers: the need to guard against relatively smaller biases. When studies were designed to detect impacts of 0.20 standard deviations or larger, it may have been reasonable for researchers to regard small biases as ignorable. For example, a bias of 0.03 standard deviations might have been ignorable in a study that could only detect an impact of 0.20 standard deviations. But in a study designed to detect much smaller impacts, such as Chiang et al. (2015) in which the impact estimate was 0.03 standard deviations, a bias of 0.03 standard deviations is no longer small—it is enormous.

Theoretical Background:

As study sample sizes increase to allow evaluators to detect smaller and smaller impacts, it is tempting to believe that *ceteris paribus*, the additional data should (1) always lead us closer to the correct answer and (2) always reduce the probability that we draw false inferences. In other words, we should get closer and closer to asymptopia, a place where "data are unlimited and estimates are consistent" (Leamer 2010).

We define asym*dys*topia as a context in which a larger sample size is not necessarily better and could even be worse from the perspective of controlling the Type 1 error rate. If, as a study becomes larger, the standard error of the impact estimate shrinks while bias stays the same (or shrinks less than the standard error), then Type 1 errors could become more common. This is because the denominator of the t-statistic (the standard error) is shrinking faster than the numerator (the biased point estimate). For example, if the true impact is 0, bias is 0.05, and the standard error is 0.20, then the t-statistic is 0.05/0.20 = 0.25 (not statistically significant). If bias shrinks to 0.025 while the standard error shrinks to 0.01, then the t-statistic becomes 2.5 (statistically significant at the 5 percent level).

Purpose and Research Questions:

We examine the potential for asymdystopia as studies are powered to detect smaller impacts, where even small biases may lead to false inferences about the existence or magnitude of an impact. We focus on the potential for bias from attrition in the case of randomized controlled trials (RCTs) and bias from regression misspecification in the case of regression discontinuity designs (RDDs). While the methodological details are distinct, in both cases we are unpacking a source of bias that may become increasingly problematic when studies are designed to detect smaller impacts.

Our two main research questions are:

1. How problematic is attrition bias in RCTs as studies are powered to detect smaller impacts?
2. How problematic is functional form misspecification bias in RDDs as studies are powered to detect smaller impacts?

Research Design and Methods:

We examine the first research question using an attrition model for RCTs used in several federal evidence reviews, including the What Works Clearinghouse (WWC 2013; 2014). This model assumes that attrition bias is ignorable so long as it accounts for less than 20 percent of whatever size impact is deemed substantively important. Using this model and data from the WWC on attrition from more than 800 prior studies, we examine:

a. How attrition may become less acceptable, leading to higher rates of false inferences, as studies are powered to detect smaller effects;
b. Contexts in which more favorable assumptions about the relationship among attrition, outcomes, and treatment status may allow for greater tolerance of attrition; and
c. The feasibility of achieving lower attrition rates in future studies that are powered to detect small impacts, based on an analysis of attrition in past RCTs that were reviewed by the WWC.

We examine the second research question using Monte Carlo simulations to assess what happens as the sample size of the RDD increases under varying assumptions regarding the true functional form. The data generating processes used for these simulations are based on data from several prior large-scale RCTs in education (James-Burdumy et al. 2010; Constantine et al. 2009; Campuzano et al. 2009). Specifically, we examine the effect of a larger sample size on statistical power, functional form misspecification bias, and the accuracy of estimated *p*-values. We also assess whether a method proposed by Calonico et al. (2014) can be used to calculate accurate *p*-values, thereby limiting false inferences even in this context.

Findings and Conclusions:

Overall, our findings suggest that biases that might have once been reasonably ignorable can pose a real threat in evaluations that are powered to detect small impacts. Our paper identifies and quantifies some of these biases, and shows that they are important to consider when designing evaluations and when analyzing and interpreting evaluation findings. We also discuss potential strategies to address these biases. Our findings should *not* be interpreted as suggesting that researchers should avoid powering evaluations to detect small impacts. The problem of small biases is real but surmountable—so long as it is not ignored.

References

Balu, R., P. Zhu, F. Doolittle, E. Schiller, J. Jenkins, and R. Gersten. "Evaluation of Response to Intervention Practices for Elementary School Reading." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2015.

Calonico, S., M. Cattaneo, and R. Titiunik. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*, vol. 82, no. 6, 2014, pp. 2295–2326.

Campuzano, L., M. Dynarski, R. Agodini, and K. Rall. "Effectiveness of Reading and Mathematics Software Products: Findings from Two Student Cohorts." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Chiang, H., A. Wellington, K. Hallgren, C. Speroni, M. Herrmann, S. Glazerman, and J. Constantine. "Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Two Years." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2015.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences,* 2nd edition. Mahwah, NJ: Lawrence Erlbaum Associates, 1988.

Constantine, J., D. Player, T. Silva, K. Hallgren, M. Grider, and J. Deke. "An Evaluation of Teachers Trained Through Different Routes to Certification, Final Report." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Hill, C. J., H. S. Bloom, A. R. Black, and M. W. Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives*, vol. 2, no. 3, 2008, pp. 172–177.

James-Burdumy, S., J. Deke, J. Lugo-Gil, N. Carey, A. Hershey, R. Gersten, R. Newman-Gonchar, J. Dimino, K. Haymond, and B. Faddis. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Findings from Two Student Cohorts." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2010.

Kane, T. J. "Frustrated with the Pace of Progress in Education? Invest in Better Evidence." Washington, DC: The Brookings Institution, 2015.

Leamer, E. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives*, vol. 24, no. 2, 2010, pp. 31–46.

Lipsey, M. W., K. Puzio, C. Yun, M. A. Hebert, K. Steinka-Fry, M. W. Cole, M. Roberts, K. S. Anthony, and M. D. Busick. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." Washington, DC: National

Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education, 2012.

What Works Clearinghouse (WWC). "Assessing Attrition Bias (Version 2.1)." Washington, DC: Institute of Education Sciences, U.S. Department of Education, 2013.

What Works Clearinghouse (WWC). "Assessing Attrition Bias—Addendum (Version 3.0)." Washington, DC: Institute of Education Sciences, U.S. Department of Education, 2014.