Multiple-Cohort Experiments to Balance Rigor, Replication, and Continuous Improvement in Education Research: An Illustrative Case from a Teacher Coaching Program

David Blazar
University of Maryland, College Park
Matthew A. Kraft
Brown University

Background and Context

Education and social science research have, in recent years, been subject to rigorous debates around appropriate policies and procedures that balance several goals: (a) methodological rigor and research designs that can support causal inferences (ESSA, 2015; Kane, 2016; Murnane & Willett, 2011); (b) studies that directly inform practice and policy, often through "researcher-practitioner partnerships" (Coburn & Penuel, 2016; Snow, 2015); and (c) an ability to replicate findings (Camerer et al., 2018; Miguel et al., 2015).

We argue that these goals can and should be viewed as complementary, even though designing studies that meet all three can be challenging. For example, the very nature of researcher-practitioner partnerships often means that studies are conducted in and meant to inform local policies; thus, the results of any given study very well may not replicate when adapted to other settings. While education researchers (Kane, 2016; Murnane & Willett, 2011) and policymakers (ESSA, 2015) have pushed for methodological rigor in designing studies capable of supporting causal inferences, practitioners often require information on *mechanisms* and *factors* driving effective or ineffective programs in order to inform continuous improvement efforts (Wagner, 1997). Unpacking mechanisms generally is a challenge in causal research.

We propose and illustrate with an example from a teacher coaching program a multiple-cohort, longitudinal experimental design that we believe achieves all three goals: methodological rigor, replication, and an ability to unpack possible mechanisms. The randomized design allows us to draw causal inferences, while the multiple-cohort design allows us to examine whether results replicate (at least in the same setting though amongst a separate set of teachers) and how changes in implementation across cohorts may explain differences in effectiveness. Because changes in implementation were determined as part of the program design prior to the start of a new cohort, between-cohort differences reflect exogenous variation in program characteristics.

Intervention and Experimental Design

The specific intervention we examine is MATCH Teacher Coaching (MTC), a teacher coaching program developed by the MATCH Public Charter School in Boston and implemented in schools across the Recovery School District in New Orleans over the course of three school years (2011-12 through 2013-14). Coaches trained under the MTC program worked with participating teachers during a four-day training workshop over the summer and then one-on-one for either three or four intensive, week-long observation and feedback cycles throughout the school year.

In each of the three cohorts, we randomly assigned half of the teachers who agreed to participate in the study to receive an offer of coaching using a blocked randomized design. In total, 217 participated in the study, including 59 teachers in cohort 1, 94 teachers in cohort 2, and 68 teachers in cohort 3.

From its inception, the developers and funders of MTC have been particularly attuned to assessing the effectiveness of the program and the extent to which there were specific components of the program that could be improved. Therefore, over the course of the three-year study, several key features were adapted: First, several of the coaches turned over across cohorts, driven by an evolving perspective from MTC leaders about the qualities of coaches needed to drive changes in teacher practice. Second, due to the growing scale of the program, MTC reduced the average amount of coaching it provided to teachers throughout the school year from four weeks to three weeks. Third, programmatic changes induced an increased focus on behavior management over other classroom practices. These changes reflect features specific to MTC, but also reflect broader three implementation features – personnel, duration, and content – that are critical components of many educational programs.

Data and Empirical Methods

We utilize three primary sources of data to triangulate the effect of MTC on teachers' practices: a classroom observation protocol developed by MTC and aligned to the coaching program, a principal evaluation derived from previous studies, and the TRIPOD student survey. We focus specifically on process measures and subjective ratings rather than on student achievement given both substantive and practical concerns about using test-score outcomes. Not only do these process measures align with our primary focus of changing teacher practices across grades and subjects, but process measures also align with the needs of continuous improvement efforts by examining effects on outcomes most proximal to the intervention.

The experimental design allows for a straightforward method for calculating treatment effects. Using Ordinary Least Squares (OLS), we predict each outcome as a function of the random offer to participate in treatment and fixed effects for randomization block. We both pool and disaggregate results by cohort, allowing us to examine whether results replicate across cohorts. To examine whether pre-determined implementation features drive differences in outcomes across cohorts, we predict outcomes as a function of these features. In these analyses, we leverage variation *across* cohorts and so must exclude fixed effects for randomization blocks.

Findings

Findings indicate that, on average across all three cohorts, MTC did not improve teachers' instructional practice as measured by classroom observations, principal survey, or student surveys. However, these average treatment effects mask important variation across cohorts. We find large positive effects on several measures of teachers' practices in cohort 1, upwards of 0.5 SD. Comparatively, we find no effects in cohort 2 or cohort 3, which are statistically significant from treatment effect in cohort 1 for several outcome measures. A set of exploratory analyses suggest that the failure to replicate may be attributable to key implementation factors, including differences in coach effectiveness and the focus of coaching across cohorts.

Conclusion

Education agencies and practitioners, including MTC, benefit from information not only about *whether* a given program works to improve desired outcomes but also *why* that program is or is not effective. Multiple-cohort experimental designs should be considered a powerful tool for future researcher-practitioner partnerships in order to estimate causal effects and also to examine how differences in implementation across cohorts relate to differences in effectiveness.

Our study also highlights several additional tensions and tradeoffs, including issues related to statistical power, and an ability to anticipate ex-ante and pre-register analyses regarding specific mechanisms to explore (Gehlbach & Robinson, 2018). Ultimately, we conclude that small studies like ours provide unique opportunities to examine replicability and possible mechanisms that drive program effectiveness, which should be paired with additional and larger-scale studies to confirm the success of a type of intervention and specific implementation features.

References

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2,* 637-644.

Coburn, C. E., & Penuel, W. R. (2016). Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48-54.

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness*, *11*(2), 296-315.

Kane, T. J. (2016). Connecting to practice. *Education Next*, *16*(2).

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... & Laitin, D. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30-31.

Murnane, R., & Willett, J. (2011). *Methods matter: Improving causal inference in education and social science research.* Oxford University Press.

Snow, C. E. (2015). 2014 Wallace Foundation Distinguished Lecture: Rigor and realism: Doing educational science in the real world. *Educational Researcher*, *44*(9), 460-466.

Wagner, J. (1997). The unavoidable intervention of educational research: A framework for reconsidering researcher-practitioner cooperation. *Educational researcher*, *26*(7), 13-22.