AERA Open
July-September 2019, Vol. 5, No. 3, pp. 1–18
DOI: 10.1177/2332858419876252
Article reuse guidelines: sagepub.com/journals-permissions

Balancing Rigor, Replication, and Relevance: A Case for Multiple-Cohort, Longitudinal Experiments

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

Over the past 15 years, the education research community has advocated for rigorous research designs that support causal inferences, for research that provides more generalizable results across settings, and for the value of research-practice partnerships that inform the design of local programs and policies. We propose the multi-cohort, longitudinal experiment (MCLE) as one approach to balancing these three, sometimes competing goals in a single study. We describe our application of an MCLE design to evaluate a teacher coaching program, where we find that changes in program features related to personnel, content, and duration coincided with substantial differences in effectiveness across three cohorts of the experiment. Our analyses and corresponding recommendations can help researchers weigh the benefits and trade-offs of the MCLE design.

Keywords: *research-practice partnership, randomized control trials, replication, continuous improvement, professional development, teacher coaching*

Introduction

Research evidence can be a critical source of information for informing social policy. Because these decisions can have large and lasting consequences, it is important that studies are designed not only to eliminate plausible alternative explanations for hypotheses (i.e., *rigorous*) but also to have external validity across contexts and conditions (i.e., *replicable*) and to inform the practice of partner organizations and local decision makers (i.e., *relevant*).[1] However, achieving these three goals in unison can be challenging, in part, because the goals themselves often compete with each other. For example, conducting *rigorous* randomized control trials (RCTs) may limit the *replicability* of a given study because inferences are limited to those participants who proactively volunteer to participate. Waiting to evaluate a program until after it has been fully developed provides the clearest assessment of its *replicability* but is too late to be *relevant* for the program design process.

Despite these tensions and challenges, we argue that scholars can adopt research designs that strike a better balance at maximizing rigor, replicability, and relevance. As the adage goes, researchers cannot fix by analysis what they bungle by design. We propose the multiple-cohort, longitudinal experiment (MCLE) as a research design that attempts to achieve these three goals and illustrate possible advantages and challenges with one example from an evaluation of a teacher coaching program across three cohorts. MCLEs randomize individuals to treatment versus control groups over multiple cohorts and vary program features across these cohorts. This setup allows researchers to test whether results replicate and, if not, to examine the role of program features in explaining cross-cohort differences in effectiveness.

The key benefit is that this design treats programs as fluid rather than static and allows for evolution and reevaluation over time. We argue that by combining random assignment to treatment with the analysis of implementation features, MCLEs are a promising approach to enable greater use of research in decision making (Tseng, 2012) and to address problems of practice that schools and districts face (Donovan, 2013; Fishman, Penuel, Allen, & Cheng, 2013). In the case study we present, differences in the effectiveness of the coaching program across cohorts coincided with changes in program design features—namely, turnover of coaches, greater emphasis on behavior management relative to other areas of teaching practice, reduction in the dosage of coaching, and increases in teacher-to-coach ratios—that informed ongoing development of the coaching program.

Of course, the MCLE design is not without challenges. It requires a sustained research-practice partnership where the program under study works with multiple cohorts that can be randomized to treatment. We illustrate one approach

to partnership-based research guided by a continuous improvement framework (Bryk, Gomez, & Grunow, 2011), though we recognize that other partnership models also are possible (e.g., those focused on increasing organizations' capacity for research; Roderick, Easton, & Sebring, 2009). As researchers, we worked collaboratively with the program designers and core staff of the MATCH Teacher Coaching (MTC) program to study its effects in the context of its first implementation in a new setting. MTC was developed at the MATCH Public Charter School in Boston and brought to charter schools across the Recovery School District in New Orleans with support and funding from New Schools for New Orleans. (For prior studies describing our work with the MTC program, see Blazar & Kraft, 2015; Kraft & Blazar, 2017.) Although we did not originally set out to conduct an MCLE, the evolution of the program and partnership helped us see the value of such an approach and the benefit of formalizing the design in our own and others' future work.

MCLEs must also examine and address a greater number of assumptions than a traditional randomized experiment. In particular, it is possible that cross-cohort spillover or differences in characteristics of participants across cohorts may drive differential treatment effects, rather than the program characteristics under study. Our case study highlights these and several other trade-offs that researchers and practitioners must consider when deciding whether to adopt an MCLE design. We organize our discussion around eight recommendations for how other researchers may address these challenges, including collecting detailed information on program participants at multiple points in time, collecting detailed data on program implementation, and, to the extent possible, anticipating ex-ante specific mechanisms to explore.

Beyond the methodological contributions of this article, our substantive findings provide further justification for researchers and practitioners to work together and innovate on research designs aimed at addressing problems of practice. Failure to replicate provided an opportunity for MTC developers, funders, program staff, and ourselves as the research partners to identify characteristics of effective teacher coaching. These findings also make clear that experimental designs and a push for causal inference on its own may not lead to better interventions and outcomes. A straightforward experimental design can identify *whether* or not a program works as implemented. But, to know *why* a program does or does not work, evaluation designs must accommodate the dynamic nature of education interventions. It is the combination of design features—random assignment to treatment versus control, and rigorous study of program features that coincide with differences in causal estimates—that define MCLEs and make them most useful both to local programs and to the research community more broadly.

Trade-offs and Challenges in Education Research

Education and social science research have, in recent years, been subject to rigorous debates around the relative importance of several important goals: (1) using methodologically *rigorous* research designs that can support causal inferences (Angrist & Pischke, 2009; Every Student Succeeds Act, 2015; Kane, 2016; Murnane & Willett, 2011); (2) producing findings that are *replicable* (Camerer et al., 2018; Miguel et al., 2014; Schneider, 2017); and (3) designing research to be *relevant* to local communities and program developers, often through research-practice partnerships (Coburn & Penuel, 2016; Snow, 2015; Tseng, 2012).

Over the past 15 years, the education research community has made substantial progress in addressing the first of these three goals.[2] For example, a recent meta-analysis of education and human capital interventions across developed countries identified 196 experiments (Fryer, 2017), representing a substantial increase in rigorously designed analyses from just a few years earlier. Fryer's calculations show that in 2000 only 14% of studies reviewed by the What Works Clearinghouse—a repository for education research—met their standards for supporting causal conclusions "without reservations" (i.e., experiments and regression discontinuity designs); by 2010, that percentage had tripled to 46%. We have seen similar trends in the context of teacher professional development (PD), which is the focus of our case study in this article. In 2007, a comprehensive review of the entire canon of literature on the effects of teacher PD (n = 1,300 studies) found only nine studies that met the What Works Clearinghouse's highest evidence standards (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). In 2018, our own meta-analysis of the causal evidence on teacher coaching—just a subset of teacher PD programs—identified 60 studies with research designs that could support causal inferences, all but one of which were published after the 2007 review (Kraft, Blazar, & Hogan, 2018).

The push for RCTs and other research designs that support causal inferences is important, but not enough. Many RCTs are designed as *efficacy* trials, which examine small programs under conditions that are intended to maximize effects. Over half of the included studies in our meta-analysis of teacher coaching programs had sample sizes of fewer than 100 teachers. Many of the programs under study were designed and implemented by researchers who were highly invested in their success. While efficacy trials provide information that is directly relevant to program developers, by design these studies do not provide information on the extent to which a given program may succeed in other settings or be scaled with fidelity. Researchers have proposed methods for examining the generalizability of findings from one experiment to a broader population (Tipton, 2014), yet none of the studies in our meta-analysis used these methods. The limited statistical power of small efficacy trials also constrains researchers' ability to identify the mechanisms underlying effective programming.

Large-scale *effectiveness* trials implemented across a range of settings can provide greater external validity and generalizability for informing state and federal policy (Wayne, Yoon, Zhu, Cronen, & Garet, 2008). But, by growing in scale, effectiveness trials generally cannot respond directly to the needs and questions generated by local communities and program developers. A recent review of all large-scale RCTs commissioned by the National Center for Educational Evaluation and Regional Assistance in the United States and by the United Kingdom–based Education Endowment Foundation found the interventions under study produced very modest effects (0.06 standard deviations, on average) and were generally underpowered to detect main effects, let alone explore mediators and moderators (Lortie-Forgues & Inglis, 2019). This lack of precision can render the results of these trials largely uninformative for both social policy and local practice.

A concern both for efficacy and effectiveness trials is that they generally are static, one-shot snapshots of program effects, whereas real-world programs and interventions evolve over time and in response to a myriad of factors including available resources, school and district conditions, and the needs of teachers and students participating in a given intervention. Of the 60 studies included in our teacher coaching meta-analysis, only 12 examined the evolving nature of the program under study or among a second or third cohort of participating teachers. It is possible that longer-term evidence may become available in the future as more time passes, or that coaching interventions that were found to be successful in a pilot study may be subject to a second round of evaluation in the future when longer-term data are collected.[3] Yet, at least in the current literature base, limited evidence of program effects over time and across cohorts makes it difficult to produce information that is directly relevant for continuous improvement efforts.

These features of many RCTs mean that simultaneously achieving goals two and three—an ability to replicate findings across multiple settings and to be relevant enough to inform the program or policy under study—can be a considerable challenge. A recent review of the top 100 education journals as measured by impact factor found that 0.13% of all published studies over a 5-year period were replications (Makel & Plucker, 2014). The very nature of research-practice partnerships also means that studies often are conducted in and meant to inform local policies (Bryk et al., 2011; Gutiérrez & Penuel, 2014). Thus, the results of any given study very well may not replicate when adapted to other settings.

Further, while practitioners often require information on factors driving effective or ineffective programs to inform continuous improvement efforts (Wagner, 1997), unpacking mechanisms and identifying mediating pathways generally is a challenge for single-cohort studies. In this setup, it is impossible to disentangle program features that are rolled out simultaneously and often selected by program staff to fit the needs of participants. Disentangling the contribution of specific program factors requires research designs that randomize participants to different variations of the program design. Doing so in a single study with sufficient statistical power can be challenging and prohibitively costly. MCLEs provide a feasible way to rigorously test the importance of different program features by systematically changing them over time and across cohorts.

A Proposal

We propose an alternative approach to the standard design of RCTs, which is considered the gold standard for evaluating education programs and interventions but often fails to simultaneously achieve replicability and relevance. The typical design of RCTs used for program evaluation involves a one-time assessment of program effects on outcomes in the same year in which the program was implemented. Implementation costs and capacity constraints frequently lead to small-scale designs with limited statistical power (Lortie-Forgues & Inglis, 2019). Participants are recruited to be volunteers, sometimes resulting in a highly nonrepresentative sample (Tipton, 2014). Researchers typically have limited interactions with program staff after gathering background information and collecting data. They retreat to conduct their analyses and return to present the finding months or even years later, often well after program staff has had to make programmatic decisions for the following year. When researchers are able to examine mechanisms, they usually do so in an ad hoc exploratory way.

Multiple-Cohort, Longitudinal Experiments

We argue that MCLEs are better equipped to address the tension between rigor, replication, and relevance. The core features of MCLEs are

1. Partnering with an organization for the purposes of evaluating a program for both continuous improvement and to contribute to the broader knowledge base.
2. Randomizing participants to evaluate program effects.
3. Conducting RCTs across multiple cohorts over time.
4. Studying changes in program features over time.
5. Tracking effects beyond the program implementation period.

Research-practice partnerships are at the core of MCLEs because they allow scholars to engage in research that is informed by and relevant to specific programs (Coburn & Penuel, 2016; Tseng, 2012). Randomized designs allow researchers to draw causal inferences, while multiple-cohort

designs provide a mechanism to examine whether results replicate and, if not, to inform ongoing program redesign and improvement efforts by studying the implementation features that drive cross-cohort differences. The longitudinal analyses allow researchers to examine whether results persist or fade out over time.

Several other alternative research designs provide elements and features of MCLEs, but few combine them all into a single design. RCTs with multiple treatment arms have the advantage of testing the effects of different program designs in a way that is not confounded with changes over time and across cohorts (for an example relevant to teacher PD, see Garet et al., 2008). However, multiarm RCTs require researchers and their partners to identify program modifications at the onset of the study. They also require substantially larger sample sizes and program capacity, which often are not feasible for local programs. Similarly, large-scale effectiveness trials with greater external validity can only be conducted for programs that have achieved scale and secured major funding for these resource-intensive studies. More exploratory, observational studies of program mechanisms among the full population of participants offers an alternative design to examine the importance of a range of program features. However, such studies can produce misleading results due to self-selection into treatment and other omitted variables.

Researchers interested in conducting MCLEs will need to work closely with their partners to align the continuous improvement goals of the program with the MCLE design outcomes of interest. For example, some organizations may focus on increasing their impact; others may seek only to sustain impacts while reducing costs or expanding their scale. Responsive researcher partners also will need to be prepared to conduct quick, short-cycle analyses to inform time-sensitive decisions by their partner organization. Fortunately, the main results of well-designed and well-implemented RCTs can be analyzed quickly relative to quasi-experimental research methods (Angrist & Pischke, 2009; Murnane & Willett, 2011).

Challenges

Along with the advantages, MCLEs also present several design and implementation challenges with which researchers will have to contend. Identifying possible solutions to these challenges is the key goal of our illustrative case study in this article.

First and foremost, using the multiple-cohort design to identify the causal effect of changes to the program requires more assumptions than a traditional RCT. Program features under study cannot be designed or selected to fit the unique needs of each cohort. One approach would be to plan for and prespecify the changes that the study will test over time (Gehlbach & Robinson, 2018). This approach ensures that changes do not reflect endogenous selection or context-specific changes across cohorts. However, ideas for program modifications often arise as programs are implemented, as was the case in the partnership we describe below. Restricting changes to those that can be identified ex-ante could unnecessarily limit the dynamic process of ongoing program improvement.

Changes in program features across cohorts may also coincide with changes in the sample of participants. As programs evolve over time, so too might the characteristics of those who volunteer to participate or who are recruited by the program. For example, teacher development programs may originally be interested in working with early career teachers but may find they need to relax this restriction to recruit sufficient numbers in later cohorts. Multiple-cohort designs also create the risk of spillover in exposure to treatment across cohorts. For example, participants randomized to the treatment group in the first cohort may interact with potential participants in future cohorts that could be randomized to the control condition. Individuals in earlier cohorts may also encourage (or discourage) colleagues from participating in a future cohort. Cross-cohort spillover could undermine the internal validity of results, while changes in the composition of cohorts would affect the external validity of the results from each cohort.

Another challenge with the MCLE design is the potentially limited statistical power from analyses using a single cohort or comparing across cohorts. Randomizing at the lowest unit possible (e.g., teachers or classrooms rather than schools) and pooling results across several cohorts can help achieve sufficient statistical power when financial or capacity constraints limit recruitment efforts and sample size with any individual cohort. At the same time, analyses of cross-cohort differences that are a primary focus of MCLE studies require relatively precise estimates to establish that differences in treatment effects across cohorts are statistically significantly different from each other.

Studies that require tracking students or teachers over multiple years may also suffer nontrivial sample attrition in contexts with high student mobility and high teacher turnover. Attrition can be particularly problematic when primary outcomes are measured using original data collected by researchers rather than by administrative data captured for all students and teachers in a district. Attrition from the sample reduces statistical power and can compromise the internal validity of an experiment if it differs across treatment and control groups.

Below we discuss an illustrative case of the MCLE design to highlight the many decisions that researchers will have to make in collaboration with their practice-based partners. We orient our case study around specific challenges noted above and recommendations for ways that researchers might contend with these challenges going forward. We organize our recommendations into two broad categories aligned to the

core work of research-practice partnerships: (1) the research design phase, done in close collaboration with program partners and (2) the analysis phase, meant to produce evidence that can inform the work of collaborating partners.

An Illustrative Case: MATCH Teacher Coaching

Research Design Phase

Recommendation 1: Develop relationships with practitioners, and codesign research in advance. We examine 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 3 school years (2011–12 through 2013–14).

Consistent with the long-standing theory of action underlying coaching programs (Joyce & Showers, 1982; Showers, 1984, 1985), MTC's primary goal was to improve teachers' classroom practice through intensive and sustained observation and feedback cycles. Coaches trained under the MTC program worked with participating teachers during a 4-day training workshop over the summer and then one-on-one for either 3 or 4 intensive, week-long observation and feedback cycles throughout the school year. During each cycle, coaches observed teachers' instruction and then met with teachers to debrief about the observations at the end of the school day. Coaches worked with teachers to set rigorous expectations for growth and evaluated teachers' progress through formative assessments on a classroom observation rubric developed by the coaching program. Between coaching sessions, teachers communicated with coaches about their progress every 1 to 2 weeks via e-mail or phone.

As described in our earlier work (Blazar & Kraft, 2015; Kraft & Blazar, 2017), from its inception the developers and funders of MTC were attuned to assessing the effectiveness of the program. In particular, they were interested in the extent to which MTC changed the experiences of teachers and students, and whether there were specific components of the program that could be improved. They also sought to identify ways to scale the coaching program within resource and financial constraints. As such, the programmatic and evaluation designs were developed in tandem on an ongoing basis throughout the experiment, rather than ex-ante. As researchers, we worked with program staff to identify the research questions, discussed plausible research designs to answer relevant questions, designed or selected measurement tools to capture implementation of the program and teacher/student outcomes, and interpreted results.

A key feature of our collaborative planning was identifying a randomized design that was agreeable to and aligned with logistical constraints facing schools, principals, and site-based staff. 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 most cases, these blocks were the schools in which teachers worked in the spring prior to the study year, though a handful of blocks consisted of teachers from multiple school sites. The blocked randomized design assured principals that approximately half of the teachers they recommended to take part in the experiment (prior to random assignment) received an offer of coaching and established an on-site partner on whom we could rely during data collection efforts. In total, 217 teachers participated in the study, including 59 teachers in Cohort 1, 94 teachers in Cohort 2, and 68 teachers in Cohort 3.

In Table 1, we confirm the success of the randomization process in terms of creating balance between the treatment and control groups on observable characteristics. We do so both pooling and disaggregating by cohort, finding no statistically significant differences on any individual measures or on joint tests across measures ($p = .596$ for a joint test of all observable characteristics across treatment and control groups, pooling across cohorts).

Recommendation 2: As part of coplanning, identify ex-ante likely mechanisms of effective programming and a process for amending these plans as the research evolves. One goal of our partnership with MTC and its core program staff was to provide input into the development of their model, which was being rolled out to an external setting (New Orleans) and at a greater scale for the first time. Therefore, we adapted several key features of the study over the course of the evaluation.

First, several of the coaches turned over across cohorts, driven both by natural career mobility and 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 and an attempt to make the coaching program more affordable for schools, MTC reduced the average amount of coaching it provided to teachers throughout the school year from 4 weeks to 3 weeks between Cohorts 1 and 2, and increased coach-to-teacher ratios between Cohorts 2 and 3. Third, programmatic changes resulted in an increased focus on behavior management over other classroom practices (i.e., instructional deliver, student engagement). Shifts in content were in response to perceived needs of teachers in Cohort 1 and aligned with MATCH Public Charter School's "no-excuses" model that also was adopted by many of the New Orleans charter schools in which the coaching program was implemented. Together, these three changes reflect features specific to MTC as well as broader categories of implementation—personnel, dosage, and content—that are critical components of many educational programs.

Notably, MTC varied program features over the course of the study rather than at the outset. While we publicly posted a pre-analysis plan on a personal website (available on request), we did not pre-register a plan for changing program

TABLE 1
Teacher Characteristics and Balance Between Treatment Groups and Cohorts

	Pooled 3 Cohorts		Cohort 1		Cohort 2		Cohort 3		*p* on Joint Difference Between All Cohorts
	Treatment Mean	Control Mean	Treatment Mean	Control Mean	Treatment Mean	Control Mean	Treatment Mean	Control Mean	
Teacher background characteristics									
Interest in coaching (1–10 scale)	9.16	9.13	9.23	8.98	9.09	9.10	9.19	9.28	.718
Female (%)	0.73	0.74	0.70	0.79	0.69	0.69	0.81	0.74	.504
African American (%)	0.19	0.20	0.20	0.14	0.14	0.22	0.24	0.23	.641
White (%)	0.72	0.74	0.77	0.76	0.75	0.76	0.65	0.69	.452
Age (years)	25.29	26.39	26.13	26.07	24.86	25.73	25.19	27.39	.277
Teaching experience (years)	2.83	3.13	3.93	4.00	2.57	2.71	2.27	2.97	.010
Alternatively certified (%)	0.76	0.72	0.80	0.72	0.82	0.80	0.65	0.62	.059
Master's degree (%)	0.22	0.24	0.20	0.24	0.25	0.24	0.19	0.23	.891
College institution ranked very competitive or higher (%)	0.80	0.72	0.73	0.79	0.84	0.73	0.81	0.64	.686
Teaching and school characteristics									
Teach all subjects (%)	0.36	0.35	0.43	0.41	0.33	0.33	0.35	0.33	.770
Teach humanities (%)	0.41	0.38	0.37	0.34	0.47	0.40	0.35	0.38	.569
Teach STEM (%)	0.28	0.35	0.20	0.24	0.31	0.40	0.30	0.38	.328
Teach elementary school (%)	0.66	0.57	0.73	0.59	0.69	0.64	0.57	0.49	.433
Teach middle school (%)	0.24	0.26	0.23	0.31~	0.25	0.27	0.24	0.23	.942
Teach high school (%)	0.21	0.21	0.17	0.21	0.18	0.13	0.27	0.31	.547
Baseline measures of outcomes									
Observation rubric: Achievement of lesson aim (1–10 scale)	4.98	5.11	4.86	4.89	4.73	5.02	5.41	5.37	.156
Observation rubric: Behavioral Climate (1–10 scale)	5.01	5.02	4.65	4.49	5.20	5.26	5.05	5.16	.268
Principal survey: Overall effectiveness composite (1–9 scale)	5.83	6.01~	6.37	6.62	5.63	5.82	5.65	5.78	.001
p value on joint test		.596		0.768		0.615		0.401	
n (teachers)	116	113	30	29	49	45	37	39	

Note. p Values estimated from regression models that control for randomization block fixed effects, with robust standard errors clustered by school-year. STEM = science, technology, engineering, and mathematics.
~$p < .10$. (On difference between treatment and control means.)

features. Without pre-registration of the cross-cohort changes in programming, we view our implementation analyses for MTC as exploratory.

Recommendation 3: Develop or identify preexisting protocols to closely monitor implementation of these program changes and to examine how they play out in practice. In partnership with MTC staff, we developed a set of protocols to capture implementation features of the coaching program. Each week coaches kept track of the practice areas (e.g., behavior management, instructional delivery, student engagement) that they worked on with a given teacher, and the tools (e.g., direct feedback, collaborative lesson planning, practice teaching) they used in their debriefing sessions. We aimed to reduce the burden on coaches by simplifying the data reporting process using online spreadsheets. Coaches simply checked off which content area they focused on and which tools they used after each coaching session.

In Figure 1, we use these data to show changes in the content of coaching across the school year by cohort. In Cohort 1, coaches worked with teachers on a range of classroom practices, including behavior management, instructional delivery, and student engagement. Coaches started out the year with a strong focus on behavior management but decreased this focus as the school year progressed. Consistent with decisions made by programming staff prior to the start of Cohort 2, we observe a much stronger emphasis on behavior management in Cohorts 2 and 3, both at

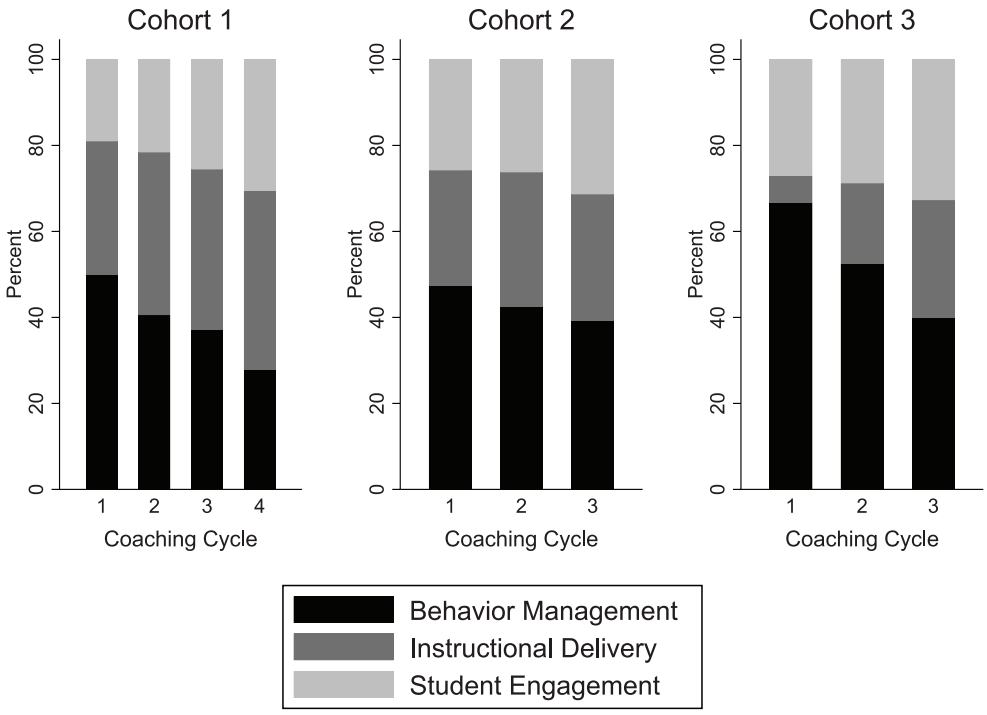

FIGURE 1. *Changes in the content of coaching across coaching sessions, by cohort.*

the start of and throughout the school year. Figure 1 also illustrates variation in coaching dosage across cohorts. On average, teachers in Cohort 1 received 4 cycles/weeks of coaching, while those in Cohorts 2 and 3 received 3 cycles/weeks of coaching.

Implementation data also showed variation in the tools that coaches used when providing feedback to teachers (Figure 2). We observe, for example, that coaches in Cohort 3 provided teachers with few opportunities for direct practice of new skills, while this was one of several tools used by coaches in Cohorts 1 and 2. For those coaches that worked across multiple cohorts, some but not all tools were used consistently.

Recommendation 4: Collect multiple outcome measures, some which are proximal to the intervention and others that are more distal and of policy relevance. When designing our study, we purposefully focused on process measures captured both at the teacher and student levels. Use of process measures aligned with MTC's continuous improvement needs, as we could estimate program effects on outcomes quite proximal to the intervention (i.e., teachers' instructional practice). Collecting outcomes that are more proximal to the treatment also is likely to result in larger effects because they are directly aligned with the treatment. Larger effects require a smaller sample size to achieve sufficient statistical power relative to studies hypothesized to detect smaller effects. Combining these estimates with effects on outcomes that were more distal (i.e., students' experiences in the classroom and their self-reports of the extent to which they learned a lot everyday) also provides an opportunity to consider how effects on teachers' practice translated into student outcomes.

We used three primary sources of data to triangulate the effect of MTC on teachers' practices. (See Appendix A for additional information and details of these data, including reliability statistics.) The first is an observation protocol developed by MTC and aligned to the coaching program, which includes two dimensions of classroom practice: *Achievement of Lesson Aim* and *Behavioral Climate*. The second is a principal survey derived from previous studies (Harris & Sass, 2009; Jacob & Lefgren, 2008), capturing a range of classroom- and school-based behaviors. We created a composite measure of all items, which we call *Overall Effectiveness*. Third is the Tripod student survey, which asks students to reflect on teachers' instructional practice and students' own experiences in the classroom (Ferguson, 2008). In the design phase of the study, we chose to focus on two of the seven domains, *Challenge* and *Control*, because of their close alignment to the aims of the coaching program. We also examined the proportion of students who agreed with a single item from the Tripod instrument: "In this class, we learn a lot every day." In an effort to guard against false positives and facilitate a parsimonious discussion of our results, we also created a *Summary Index* that is a weighted average of all measures.

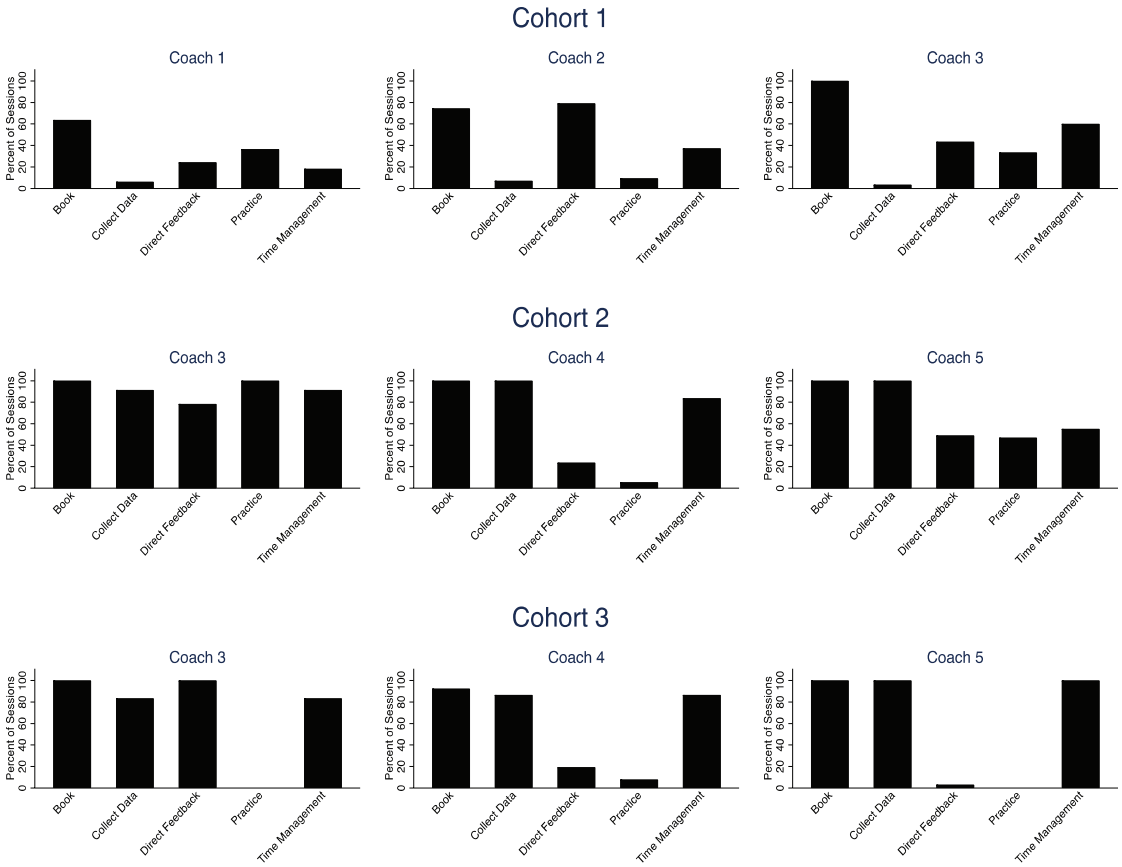

FIGURE 2. *Techniques used in debriefing sessions with teachers, by cohort and coach.*

Analyzing student achievement outcomes was not feasible given that MTC needed to recruit teachers across all grades and subjects to make the program financially viable. The majority of our teachers did not work in a tested grade and subject. This was a necessary compromise, but also limited the scope of our analyses and the degree to which we could address a common question from policymakers about how programs affect student achievement.

Recommendation 5: Track outcomes at multiple time points and, where possible, rely on preexisting administrative records to reduce the burden of primary data collection efforts, increase statistical power, and guard against sample attrition. For teacher-level outcomes, we captured data at baseline, at the end of the intervention year, and at the end of the follow-up year. The baseline data are not necessary to include in our analyses given the randomized design. But, controlling for these measures increased the statistical power of our analyses by explaining residual variation in our outcomes. This approach is particularly useful in smaller scale efficacy trials. Baseline data also allowed us to examine possible threats to internal validity due to differences in samples across cohorts (see additional discussion below). The follow-up year data allowed us to track the persistence or fade out of program effects. For the student survey, we captured data at the latter two time points, but not at baseline due to logistical and cost constraints.

Despite having designed our study to allow for rich primary data collection efforts (e.g., randomizing teachers within schools to gain the support of principals and their help collecting data), a substantive portion of teachers attrited from the study. The high turnover rates among teachers in New Orleans charter schools, many of whom are not from the city, fueled much of this attrition. Some teachers also withdrew their participation from the study. Attrition was particularly pronounced in the follow-up data. As shown in Appendix B: Table B1, roughly 70% of treatment teachers in Cohort 1 worked with us to collect classroom observation and student survey data a year after coaching ended, compared with roughly 40% of control-group teachers. The reverse was true in Cohort 3, where treatment teachers were much less likely than control teachers to participate in follow-up data collection (27% compared with 64%). We did not find any difference in the rate at which treatment or control group teachers contributed primary data at the end of the coaching year.

We also examined whether attriters from the treatment group differed from attriters in the control group on observable characteristics (see Appendix B: Table B2). To do so, we regressed each observable characteristic on indicators for attrition and treatment status, and their interaction. Pooling across cohorts, we find no evidence of differential attrition either at the end of the coaching year or at the end of the follow-up year. We do find some evidence of differential attrition when disaggregating by cohort. However, the characteristics that are related to attrition differ across cohorts and time points (i.e., end of coaching year vs. end of follow-up year).

One way to limit threats to internal validity due to nonrandom attrition and increase power may be to rely on preexisting data collected by partner school districts (or state agencies). Historically, student test scores—which we could not use in our analyses—have been the primary focus of administrative data. However, in the years following our study, additional student and teacher data have become more widely integrated into state-based data systems (Steinberg & Donaldson, 2016). Student test scores, attendance and disciplinary records, and on-time grade advancement, as well as teacher observation scores collected by districts or states may be a rich source of data for future MCLEs that can relieve financial and logistical constraints and minimize attrition from the data. Administrative data can also maximize power by providing multiple data points for outcomes over time (McKenzie, 2012).

Analysis Phase

Recommendation 6: Share results both pooled and disaggregated by cohorts, including nonsignificant point estimates. In order to provide timely and informative data back to our partners, we analyzed data and shared preliminary results midway through the summer after the end of each cohort and before a new cohort began. The randomized design allowed for a quick and straightforward approach to estimating the causal effect of MTC on teacher and student outcomes. (See Appendix C for additional information and details of our analytic approach and methods, including the specific models and estimation techniques.) We used ordinary least squares (OLS) regression to estimate differences in means between treatment and control groups, controlling for a baseline measure of the outcome (where available) and fixed effects for randomization blocks that match our blocked randomized design. To account for the clustered nature of the data (i.e., teachers within schools, students within classrooms), we clustered standard errors at the school-year level for our teacher-level analyses and at the classroom level for our student-level analyses. We both pooled and disaggregated results by cohort to examine whether results replicate across cohorts.

Findings indicate that, when pooling across all three cohorts, MTC did not improve teachers' instructional practice as measured by classroom observations, principal surveys, or student surveys. However, these average treatment effects mask important variation across cohorts. As shown in Table 2, we find large positive effects on several measures of teachers' instructional practices in Cohort 1 at the end of the coaching year. Treatment teachers scored 0.55 standard deviations higher than control group teachers on the *Summary Index* of effective teaching practices. In Cohort 1, access to the coaching program increased the probability that students reported that they "learned a lot everyday" by 8.5 percentage points. Comparatively, we generally find no effects of MTC in Cohort 2 or Cohort 3. In Cohort 3, we observe statistically significant negative effects of the random offer of coaching on the *Control* and *Challenge* constructs from the Tripod survey. Differences in effectiveness between Cohorts 1 and 2, and between Cohorts 1 and 3 often are statistically significant. We do not observe any differences in effectiveness between Cohorts 2 and 3 at the end of the coaching year.

We also observe cross-cohort differences in the effect of the MTC program in the spring of the follow-up year, a year after teachers stopped receiving coaching services. In Cohort 1, we continue to observe positive point estimates that are similar in magnitude to those at the end of the coaching year; however, for most outcome measures, these effects are not statistically significantly different from zero (two exceptions are for *Achievement of the Lesson Aim* and percentage of students who reported that they "learned a lot everyday"). For Cohorts 2 and 3, follow-up effects often are negative in magnitude, and many of these point estimates are statistically significantly different from the follow-up effect for Cohort 1. The follow-up effect for Cohort 3 on the *Summary Index* is especially large and negative in magnitude. However, we are cautious in placing too much emphasis on this estimate—and follow-up effects for Cohort 3 generally—given concerns that very high attrition rates of treatment teachers relative to control group teachers leads us to substantially understate these follow-up effects (see Appendix B).

Although several of these estimates of the follow-up effects of MTC were less precise than necessary to achieve traditional levels of statistical significance, they still provided important information to core staff about the potential sustained effects of the program. As other scholars have argued, achieving traditional levels of statistical significance may be less relevant for informing ongoing programmatic improvement when organizations are working with limited information and inflexible timelines (Conaway & Goldhaber, 2018). Imprecise point estimates provide additional data that can be combined with qualitative experiences on the ground and participant feedback to make real-time decisions.

TABLE 2
Parameter Estimates of the Effect of MATCH Teacher Coaching

		Observation Rubric		Principal Survey	TRIPOD Student Survey		
	Summary Index	Achievement of Lesson Aim	Behavioral Climate	Overall Effectiveness Composite	Control	Challenge	Learn a Lot
PANEL A: Spring of intervention year							
				Regression 1: Pooled results			
Treat	0.044 (0.144)	0.152 (0.164)	0.234 (0.141)	−0.074 (0.144)	−0.056 (0.053)	−0.004 (0.045)	−0.007 (0.018)
				Regression 2: Results by cohort			
Treat * Cohort 1	0.551* (0.243)	0.577 (0.324)	0.671* (0.318)	0.219 (0.156)	0.130 (0.101)	0.327*** (0.073)	0.085** (0.029)
Treat * Cohort 2	−0.190 (0.240)	−0.190 (0.245)	0.027 (0.194)	−0.261 (0.280)	−0.069 (0.099)	−0.132 (0.082)	−0.040 (0.031)
Treat * Cohort 3	−0.069 (0.213)	0.232 (0.283)	0.142 (0.229)	−0.084 (0.187)	−0.195* (0.083)	−0.145* (0.071)	−0.048 (0.032)
p value on test between cohort coefficients							
Cohort 1 versus Cohort 2	.035	.060	.088	.124	.161	.000	.004
Cohort 1 versus Cohort 3	.059	.426	.182	.231	.014	.000	.003
Cohort 2 versus Cohort 3	.711	.262	.702	.595	.364	.910	.871
n (teachers)	199	196	197	192	173	173	173
n (students)	—	—	—	—	5,249	5,261	5,147
PANEL B: Spring of follow-up year							
				Regression 1: Pooled results			
Treat	−0.078 (0.318)	0.280 (0.270)	−0.068 (0.341)	−0.071 (0.336)	−0.115 (0.078)	−0.012 (0.080)	0.029 (0.037)
				Regression 2: Results by cohort			
Treat * Cohort 1	0.538 (0.480)	0.977* (0.386)	0.688 (0.669)	0.247 (0.469)	−0.066 (0.141)	0.161 (0.158)	0.114~ (0.069)
Treat * Cohort 2	0.061 (0.454)	0.139 (0.336)	−0.248 (0.459)	0.143 (0.588)	−0.225~ (0.127)	−0.035 (0.124)	0.006 (0.057)
Treat * Cohort 3	−1.093*** (0.303)	−0.290 (0.526)	−0.623 (0.503)	−0.797* (0.366)	0.169 (0.146)	−0.240** (0.087)	−0.044 (0.044)
p value on test between cohort coefficients							
Cohort 1 versus Cohort 2	.472	.111	.257	.891	.435	.361	.241
Cohort 1 versus Cohort 3	.006	.058	.125	.082	.288	.038	.058
Cohort 2 versus Cohort 3	.037	.493	.581	.172	.038	.178	.502
n (teachers)	107	103	103	103	88	88	88
n (students)	—	—	—	—	2,773	2,781	2,709

Note. All regression models include fixed effects for randomization blocks and a baseline measure of the outcome where available. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains. Robust standard errors clustered at the school-year level (for teacher-level outcomes) or at the class level (for student-level outcomes) in parentheses.
~$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Recommendation 7: Examine threats to internal validity of cross-cohort differences related to the research design (e.g., characteristics of participants, spillover effects). Differences in treatment effects across cohorts coincided with changes in implementation that we describe above and that we model formally below. However, it is premature to conclude that the program features *caused* changes in treatment effects. It is possible that design features may have led to omitted variables that drove these differences. An important step in studies that leverage the MCLE design is to rule out alternative explanations.

As described in our prior work (Blazar & Kraft, 2015), one possible explanation for differential treatment effects related to the research design may be that the treatment remained constant across cohorts but that the counterfactual experiences of control-group teachers changed across years. This might be true if there were general improvements in programming provided to teachers across cohorts, or if spillover effects meant that control-group teachers in later cohorts had access to strategies used in the MTC program that those in the first cohort did not. Both would result in the same reduced treatment-control contrast.

Like others (Angrist, Pathak, & Walters, 2013), we examine this form of bias by comparing control groups from different cohorts on baseline measures of teacher practice (see Table 1). We find no difference between cohorts on the two dimensions from the observation rubric (p = .156 and .268 on joint tests that cohort indicators predict baseline scores, for *Achievement of Lesson Aim* and *Behavioral Climate*, respectively), but we do find a difference on principal reports of teacher effectiveness at baseline (p = .001). Average baseline scores were higher in Cohort 1 relative to the other two cohorts. We view these patterns as unlikely to explain larger effects of MTC in Cohort 1 relative to the other two cohorts given that higher baseline scores generally leave less (not more) room for improvement.

It also is possible that the composition of teachers changed across cohorts. Notably, teachers' baseline interest in participating in the coaching program did not differ between Cohort 1 and the other two cohorts (see Table 1). However, we identify cross-cohort differences in years of prior teaching experience (p = .010) and certification pathways (p = .059). Teachers in Cohort 1 were more experienced, on average, than those in the other cohorts, while teachers in Cohort 3 were less likely to have gone through an alternative certification program. It is possible that differences in these teacher characteristics could explain cross-cohort differences in treatment effects if more experienced teachers benefited more from teacher coaching. However, in prior published work, we disaggregated treatment effects by teacher experience levels and found that treatment effects were similar across cohort within experience levels (Blazar & Kraft, 2015).

Spillover, either within or across cohorts, can create another threat to internal validity for the MCLE design. To examine this possibility, we administered an end-of-year survey that, among other information, asked control-group teachers whether they (1) learned about strategies taught by MTC during the coaching year or (2) used these strategies in their instruction. Our data suggest that spillover occurred to some degree in all three cohorts, but that control-group teachers in Cohort 1 were *more* likely than those in other cohorts to use MTC-based strategies in their instruction. Spillover may have been higher in our blocked randomized trial rather than in an alternative design where teachers were randomized across schools, or where schools were randomly assigned to treatment. At the same time, prior research suggests that exposure to treatment via peers is unlikely to be a first-order concern for most education interventions (Rhoads, 2016). Randomizing at the teacher rather than at the school level helped maximize our sample size and facilitate greater support for school-based collection efforts by assuring all principals that some of their teachers would benefit from coaching. We see the benefits to statistical power from this approach outweighing the threat of spillover.

Because selection into and out of treatment may vary based on the specific program under study, we encourage researchers to work collaboratively with program designers to outline other possible threats to internal validity and design data collection instruments to assess the presence of these threats. We also encourage researchers interested in using the MCLE design to survey participants at baseline on a range of characteristics, including motivation and interest in the program, which facilitates comparisons of cohort characteristics, and also to survey participants at the end of treatment to document possible spillover and understand movement into and out of the program.

Recommendation 8: Model the relationship between changes in program features and changes in outcomes across cohorts. If MTC had varied just one program feature across cohorts, then—after ruling out alternative explanations related to the research design—any differences in observed effects across cohorts should reflect the causal effect of that specific feature. Because MTC varied several features at once, we cannot reasonably identify the unique contribution of each simply by examining cross-cohort differences in program effects. This may be true for other MCLE studies, where multiple program features changed in tandem.

Instead, to examine whether predetermined implementation features were related to differences in outcomes across cohorts, we conducted more exploratory descriptive analyses by modeling changes in outcomes as a function of program characteristics. We view this approach as an important supplement to qualitative analysis of program implementation. Instead of a treatment indicator, our main predictors were sets of variables describing variation in program implementation, including dummy variables for individual coaches, a count of the number of coaching sessions each

TABLE 3
Parameter Estimates of the Effect of Match Teacher Coaching on Teachers' Practices Dissaggregated by Coach

		MATCH Rubric		Principal Survey	TRIPOD Student Survey		
	Summary Index	Achievement of Lesson Aim	Behavioral Climate	Overall Effectiveness Composite	Control	Challenge	Learn a Lot
Coach 1	0.345 (0.235)	0.427 (0.256)	0.408 (0.248)	0.255 (0.325)	−0.091 (0.177)	−0.021 (0.135)	−0.004 (0.060)
Coach 2	0.544* (0.239)	0.988** (0.320)	1.095*** (0.267)	−0.339 (0.298)	0.454** (0.167)	0.389*** (0.100)	0.095** (0.031)
Coach 3	0.379 (0.270)	0.404 (0.293)	0.519˜ (0.275)	0.256 (0.258)	0.006 (0.107)	0.073 (0.084)	0.050˜ (0.030)
Coach 4	−0.340* (0.164)	−0.227 (0.204)	−0.169 (0.147)	−0.236 (0.174)	−0.201* (0.081)	−0.096 (0.065)	−0.030 (0.027)
Coach 5	0.334˜ (0.197)	0.315 (0.210)	0.321˜ (0.167)	0.319 (0.222)	−0.010 (0.112)	−0.053 (0.085)	−0.069* (0.032)
p Value on joint test of coach coefficients	.012	.024	.001	.189	.012	.003	.006
n (teachers)	199	196	197	192	173	173	173
n (students)	—	—	—	—	5,249	5,261	5,147

Note. Estimates in each column are from separate regression models. Coach indicator variables weighted by the amount of time a teacher spent with one coach versus another; these always are coded as 0 for control group teachers. All regressions include fixed effects for cohort and a baseline measure of the outcome where available. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains. Robust standard errors clustered at the school-year level (for teacher-level outcomes) or at the class level (for student-level outcomes) in parentheses.
˜$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

teacher received, and a vector of variables indicating the number of these sessions that a teacher worked on each of three instructional focus area (i.e., behavior management, instructional delivery, student engagement). We removed fixed effects for randomization block given the observational nature of these analyses. We added a cohort indicator to hold constant any differences in outcomes across years due to, for example, differences in classroom raters across years. (See Appendix C for additional discussion and estimation equations.)

These exploratory analyses suggest that the failure to replicate may be attributable to key implementation factors, including differences in coach effectiveness and the instructional focus areas across cohorts. In Table 3, we disaggregate effects by coach, controlling for cohort. Because several coaches worked across cohorts, we are able to separate coach effects from cohort effects. (Of the five coaches, Coaches 1 and 2 worked only in Cohort 1, Coach 3 worked in all three cohorts, and Coaches 4 and 5 worked in Cohorts 2 and 3.) We find consistently large positive effects for one of the coaches from Cohort 1. *P* values on tests of the null hypothesis that coach indicators are jointly equal to zero are less than the 0.05 threshold for most outcome measures. We conclude from these tests that differences in coach effectiveness likely are a key driver of differences in the effectiveness of the coaching program between cohorts. Although our study was not designed to identify the characteristics of effective versus less effective coaches, analyses of implementation data suggest that coaches differed in the tools (e.g., direct feedback, collaborative lesson planning, practice teaching) they used in their debriefing sessions with teachers (see Figure 2).

A second key change between Cohort 1 versus Cohorts 2 and 3 that may explain differences in effectiveness is the instructional focus areas of coaching sessions. Consistent with shifts in the content of coaching across cohorts (see Figure 1), we find that coaching time spent on some areas of instructional practice may be more beneficial than time spent on others. In Table 4, we present estimates from a model of the relationship between the number of sessions focused on each classroom practice area (i.e., behavior management, instructional delivery, student engagement). We included cohort fixed effects as well as baseline observation scores and the total number of weeks of coaching received, as we recognized that teachers who required more support overall or in a given area likely received more coaching aligned to that area. We observe that more time spent on behavior management is associated with decreased effectiveness, while more time spent on student engagement is associated with increased effectiveness. While these analyses are descriptive in nature, they align with the cross-cohort differences described earlier. That is, effects are largest in Cohort 1 where observation and feedback focused on all three areas of practice, relative to effects in Cohorts 2 and 3 that focused much more on behavior management.

Two additional features of MTC that varied across cohorts were dosage and teacher-to-coach ratios. As shown in Figure 1, teachers in Cohort 1 received 4 cycles/weeks of coaching, on average, while those in Cohorts 2 and 3 received 3 cycles/weeks of coaching, on average. Changes in dosage and differences in the sample size of treatment teachers across cohorts also resulted in differences in teacher-to-coach ratios. However, the design of our study does not allow us to tease out the effect of coach workload

TABLE 4

Parameter Estimates of the Effect of Match Teacher Coaching on Teachers' Practices Disaggregated by Focus of Coaching

		MATCH Rubric		Principal Survey	TRIPOD Student Survey		
	Summary Index	Achievement of Lesson Aim	Behavioral Climate	Overall Effectiveness Composite	Control	Challenge	Learn a Lot
Behavior management	−0.139 (0.093)	−0.236* (0.106)	−0.243** (0.085)	0.063 (0.105)	−0.162** (0.052)	−0.069˜ (0.039)	−0.032* (0.013)
Instructional deliver	−0.158˜ (0.084)	−0.160 (0.107)	−0.094 (0.087)	−0.116 (0.071)	−0.036 (0.048)	−0.034 (0.039)	0.009 (0.016)
Student engagement	0.271* (0.116)	0.215˜ (0.122)	0.333** (0.114)	0.163 (0.100)	0.057 (0.049)	0.076˜ (0.041)	0.025 (0.016)
Number of weeks of coaching	0.107 (0.093)	0.237* (0.100)	0.166* (0.081)	−0.053 (0.100)	0.110* (0.052)	0.044 (0.040)	0.008 (0.015)
p values for tests between focus area coefficients							
Behavior management = Instructional deliver	.004	.010	.000	.345	.001	.005	.003
Behavior management = Student engagement	.888	.645	.248	.226	.102	.560	.075
Instructional deliver = Student engagement	.013	.046	.014	.067	.264	.112	.553
n (teachers)	199	196	197	192	173	173	173
n (students)	—	—	—	—	5,249	5,261	5,147

Note. Estimates in each column are from separate regression models. Focus area variables indicate the number of sessions that a teacher worked on a given area; these always are coded as 0 for control group teachers. All regressions include fixed effects for cohort and baseline measures of the outcome where available. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains. Robust standard errors clustered at the school-year level (for teacher-level outcomes) or at the class level (for student-level outcomes) in parentheses.

˜$p < .10$. *$p < .05$. **$p < .01$.

and dosage from differences in coach effectiveness or focus on different classroom practices. This is a limitation of our MCLE design and motivates additional work with a research design that varies these features separately from others.

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. Through our research-practice partnership with MTC, we add to a growing body of literature that examines the efficacy of teacher coaching as a development tool (Kraft et al., 2018) by providing experimental evidence to show that MTC can improve teachers' instructional practice.

However, we also failed to replicate the encouraging findings from Cohort 1 across two subsequent cohorts. Several implementation features appear to explain this pattern, including coach turnover and a greater focus on behavior management relative to other areas of teaching practice. After ruling out alternative explanations for cross-cohort differences in effectiveness, we view our results as suggestive that individual coach effects and coaching content play key roles in determining the overall effectiveness of the MTC program. In our own partnership with MTC and in research-practice partnerships more broadly, this is the sort of information that is necessary to drive continuous improvement efforts.

The MTC case study serves as an example for future research-practice partnerships about how to use MCLE designs to balance methodological rigor, replicability, and relevance. Other research designs may achieve similar goals to MCLEs, and we encourage researchers, evaluators, and program staff to consider the range of options that most closely aligns with their own continuous improvement efforts.

Although we find value in exploiting cross-cohort differences in the effectiveness of MTC, failure to replicate also raises concerns regarding findings from small pilot studies in education research. Ultimately, drawing conclusions about the benefit of any given type of education intervention and investing heavily in these interventions at the state or federal level will require evidence of replicability. MCLE designs build in initial tests of the replicability of program effects. The results of these within-study replication attempts

can solidify our confidence about the efficacy of a program and prevent a premature policy rush to scale up programs with limited evidence from small trials.

Appendix A

Data Sources

We used three data sources to triangulate the effect of MTC on measures of teachers' instructional practice and effectiveness.

MATCH Classroom Observation Rubric. As described in prior work (Blazar & Kraft, 2015; Kraft & Blazar, 2017), the MATCH rubric is composed of two overall codes, *Achievement of Lesson Aim* and *Behavioral Climate*. Each code is scored holistically on a scale of 1 to 10 based on key indicators observed in a lesson. Indicators for *Achievement of Lesson Aim* include clarity and rigor of the aim, alignment of student practice, and assessment and feedback. Indicators for *Behavioral Climate* include time on task, transitions, and student responses to teacher corrections. Coaches observed and rated teachers on the rubric in the spring semester prior to randomization. In the spring at the end of the intervention year and in the spring of the follow-up year, experienced outside observers who were blind to treatment status observed and rated a class taught by each teacher on two separate occasions (one rater at each occasion). After receiving training on how to use the instrument, raters achieved one-off agreement rates with the director of MTC of 80% or higher. We created teacher scores for each code by averaging raw scores across our two raters and then standardizing average scores in each year to be mean zero and standard deviation of one.

Principal Survey

We used a principal survey adapted from surveys developed by Jacob and Lefgren (2008) and Harris and Sass (2009), both of which were found to be moderately correlated with teacher value-added scores in math and reading (0.32 and 0.29, respectively, for the former survey, and 0.28 and 0.22, for the latter). Principals rated teachers on a scale from 1 (*inadequate*) to 9 (*exceptional*) across 10 items: *Overall Effectiveness, Dedication and Work Ethic, Organization, Classroom Management, Time Management in Class, Time on Task in Class, Relationships with Students, Communication with Parents, Collaboration with Colleagues*, and *Relationships with Administrators.* One additional item asked principals to rank teachers in a given quintile of effectiveness compared with all the teachers at their school. Principals completed survey evaluations for each teacher in the spring prior to the coaching year, at the end of the following academic year at the end of the intervention year, and in the spring at the end of the follow-up year. We created a composite score of teachers' overall effectiveness, *Overall Effectiveness*, by standardizing individual items within each year, averaging scores across all 11 items above, and then restandardizing this composite score to be mean zero and standard deviation one. We estimated an internal consistency reliability of 0.91 or greater in all administrations. It was not feasible to keep principals blind to teachers' experimental condition. This could potentially bias principal evaluations scores if principals were inclined to rate teachers who participated in coaching more favorably. However, there was no incentive to do so, as results of the experiment did not affect funding for the program or any school evaluation.

Tripod Student Survey

The Tripod survey (Ferguson, 2008) is composed of items designed to capture students' opinions about their teacher's instructional practices. In the design phase of the study, we chose to focus on two of the seven domains, *Challenge* and *Control*, because of their alignment to the coaching program. These two measures also were found to be most predictive of teachers' value-added scores with correlations of 0.22 and 0.14 in math and reading (Kane & Staiger, 2011). We also examined the proportion of students who agreed with a single item, "In this class, we learn a lot every day." Upper elementary and secondary students rated each item on a 5-point Likert-type scale, while early elementary students had three response choices: no, maybe, and yes. Students completed the survey once at the end of the coaching year, and a separate group of students rated these teachers again at the end of the follow-up year. Following the practices of the Tripod project (Ferguson, 2008), we derived scores for each domain by rescaling items to be consistent across all forms, standardizing Likert-type scale response options for each item, and calculating the mean response across items. We then restandardized average scores for each domain to be mean zero and standard deviation one.

Summary Index

In an effort to guard against false positives and facilitate a parsimonious discussion of our results, we created a summary index of these three measures. We created this *Summary Index* by taking a weighted average of the five scores described above—the two items from the MATCH observation rubric, the principal survey composite, and the two Tripod composites (for similar approaches, see Anderson, 2008; Kling, Liebman, & Katz, 2007). For our primary analyses, all three data sources were given equal weight. We then standardized the index to be mean zero and standard deviation one. We also tested the robustness of our findings to alternative composites that gave more weight to the principal and student surveys, which were less proximal to the coaching program than the MATCH rubric; we found that results were similar. Pooling across all cohorts, internal consistency reliability for the five measures that comprise the *Summary Index* is 0.73.

Appendix B

Attrition

TABLE B1

Proportion of Teachers With Outcome Data on Different Measures

	Pooled 3 Cohorts		Cohort 1		Cohort 2		Cohort 3	
	Treatment	Control	Treatment	Control	Treatment	Control	Treatment	Control
PANEL A: Spring of intervention year								
MATCH Teacher Observation Rubric	0.888	0.841	0.933	0.828	0.918	0.822	0.811	0.872
Principal Survey of Teachers	0.862	0.823	0.933	0.828	0.898	0.800	0.757	0.846
Tripod Student Survey of Teachers	0.759	0.752	0.867	0.828	0.714	0.667	0.730	0.795
PANEL B: Spring of follow-up year								
MATCH Teacher Observation Rubric	0.466	0.442	0.667	0.380*	0.490	0.311	0.270	0.641***
Principal Survey of Teachers	0.448	0.460	0.700	0.414*	0.429	0.333	0.270	0.641***
Tripod Student Survey of Teachers	0.457	0.442	0.700	0.414*	0.449	0.289	0.270	0.641***
n (teachers)	229		59		94		76	

*$p < .05$. ***$p < .001$. (On difference between treatment and control.)

TABLE B2

Parameter Estimates of the Difference in Demographic Characteristics of Attritors Across Treatment and Control Groups

	Pooled 3 Cohorts		Cohort 1		Cohort 2		Cohort 3	
	Interaction Coefficient	*p*	Interaction Coefficient	*p*	Interaction Coefficient	*p*	Interaction Coefficient	*p*
PANEL A: Spring of Intervention Year								
Interest in coaching (1–10 scale)	−0.70	.161	−0.84	.261	−0.63	.388	−0.37	.697
Female (%)	0.22	.149	0.29	.443	−0.02	.942	0.32	.180
African American (%)	0.01	.950	0.01	.973	−0.40	.080	0.23	.300
White (%)	−0.13	.551	−0.54	.180	0.28	.331	−0.31	.227
Age (years)	−3.24	.155	−5.35	.042	0.56	.699	−7.77	.170
Teaching experience (years)	−1.47	.307	−2.00	.051	0.97	.542	−5.21	.100
Alternatively certified (%)	−0.07	.653	0.31	.332	−0.18	.454	−0.04	.895
Master's degree (%)	−0.06	.815	−0.50	.174	0.47	.141	−0.34	.389
College institution ranked very competitive or higher (%)	−0.05	.805	0.13	.697	0.18	.430	−0.29	.442
PANEL B: Spring of follow-up year								
Interest in coaching (1–10 scale)	−0.11	.747	−1.25	.027	0.51	.227	−0.07	0.922
Female (%)	0.17	.332	−0.24	.264	−0.12	.628	0.90	0.004
African American (%)	0.01	.949	0.07	.814	0.02	.905	−0.02	0.932
White (%)	−0.14	.364	−0.14	.679	−0.15	.506	−0.06	0.844
Age (years)	−1.23	.431	0.64	.651	0.71	.630	−4.89	0.253
Teaching experience (years)	0.28	.712	0.14	.913	1.77	.047	−1.74	0.296
Alternatively certified (%)	−0.13	.437	−0.23	.293	−0.07	.801	−0.11	0.757
Master's degree (%)	−0.06	.668	−0.23	.369	0.04	.845	−0.02	0.955
College institution ranked very competitive or higher (%)	0.08	.589	0.14	.601	0.25	.229	−0.29	0.351
n (teachers)	229		59		94		76	

Note. Coefficients come from a regression model that includes a treatment indicator, an indicator for attrition, and the interaction between the two (this is the coefficient presented in the table), as well as fixed effects for randomization blocks. Robust standard errors clustered at the school-year level.

Appendix C

Methods and Analyses

We estimated the effect of MTC on our outcomes of interest using OLS regression. We analyzed our teacher-level measures, including observation scores, principal ratings, and teacher self-evaluations, by fitting the following OLS regressions, where *Y* represents a given outcome of interest for teacher *j* at time *t*:

$$Y_{jt} = Y_{j,t=0} + \beta MTC_j + \alpha_{s,t=0} + \varepsilon_{jt} \quad (1)$$

We specified separate models for outcomes captured at the end of the coaching year (i.e., $t = 1$) and at the end of the follow-up year (i.e., $t = 2$). For each of our teacher-level outcomes, we were able to include a baseline measure, $Y_{j,t=0}$, to increase the precision of our estimates. For the *Summary Index*, we calculated a baseline measure from the MATCH rubric and principal survey, excluding the student survey data, as data collection costs prohibited us from administering this measure at the beginning of the school year. To match our research design, we included fixed effects for our randomization blocks, $\alpha_{s,t=0}$; in most cases, these blocks are the schools where teachers worked in the year prior to coaching. Because randomization blocks are unique across cohorts, treatment teachers are compared with control group teachers in their same block and cohort. We clustered standard errors at the school-year level in the current year to account for the nested structure of the data. We also tested the robustness of our results to model specifications that replaced randomization blocks with school-by-cohort fixed effects and found similar results.

We analyzed our student-level survey outcomes for student *i* at the end of the coaching year and at the end of the follow-up year by fitting an analogous model, but without controls for baseline measures of the outcome:

$$A_{ijt} = \beta MTC_j + \alpha_{s,t=0} + \varepsilon_{ijt} \quad (2)$$

To account for the nesting of students within classrooms, we clustered standard errors at the class level.

In both models, the coefficients β on the indicator for whether a teacher was randomly offered the opportunity to participate in MTC are our parameters of interest. We focus on these Intent to Treat (ITT) estimates, given that few treatment teachers dropped coaching, and most of these teachers were censored from our data because they either left teaching or did not want to participate in data collection. These data constraints mean that we are not able to calculate formally Treatment on the Treated. However, if we assume that attrition is random, which seems plausible given the circumstances described to us by many of the teachers who left the study, as well as analyses exploring differential attrition between treatment and control groups at the end of the year of coaching, then we can calculate Treatment on the Treated estimates by scaling our Intent to Treat estimates by the inverse of the take-up rate. We both pool and disaggregate results by cohort, allowing us to examine whether results replicated across cohorts.

To examine whether predetermined implementation features drove differences in outcomes across cohorts, we predicted outcomes (at the end of the coaching year only) as a function of these features. These exploratory analyses derive from slight modifications to the regression models described above. Specifically, the teacher- and student-level models that describe the relationships between coaching characteristics and each of our outcomes measures are given by Equations (3) and (4), respectively:

$$Y_{jt} = Y_{j,t=0} + \beta COACHING\_CHARACTERISTIC_j + \delta_h + \varepsilon_{jt} \quad (3)$$

$$A_{ij} = \beta COACHING\_CHARACTERISTIC_j + \delta_h + \varepsilon_{ij} \quad (4)$$

Here, $COACHING\_CHARACTERISTIC_j$ represents either a set of indicators for individual coaches or a vector of variables indicating the number of sessions that a teacher worked on each focus area (i.e., behavior management, instructional delivery, student engagement). We removed fixed effects for randomization block given the observational nature of these analyses. That is, coaches were not randomly assigned but were matched with teachers by coaches' expertise in a given school level (i.e., elementary, middle, or high) based on prior teaching experience. In addition, the number of sessions that teachers worked on a given focus area is based on teachers' needs and is an endogenous choice of coaches. We added a cohort indicator, δ_h, for cohort *h* to hold constant any differences in outcomes across school years due to, for example, differences in classroom raters across years.

ORCID iD

David Blazar https://orcid.org/0000-0001-5596-1552

Notes

1. *Relevance* may also refer to a broader research and policy audience. However, in this article, we use the term to refer to *relevance* to local communities, primarily those that created, implemented, or directly participated in the program under study.

2. In particular, the passage of the Education Sciences Reform Act in 2002, which authorized the Institute for Education Research, raised the standards for methodological rigor in educational research and created new funding sources for large-scale program evaluation studies.

3. See, for example, several studies evaluating the My Teaching Partner coaching program, where earlier evaluations examined short-term effects only (Mashburn, Downer, & Hamre, 2010; Pianta, Mashburn, Downer, Hamre, & Justice, 2008) and later evaluations examined effects in the follow-up year (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Pianta et al., 2017). However, the My Teaching

Partner program and the large body of evidence on its effectiveness was an exception in our meta-analysis. Most studies evaluated separate coaching programs.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*, 1034–1037.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*, 1481–1495.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Blazar, D., & Kraft, M. A. (2015). Exploring mechanisms of effective teacher coaching: A tale of two cohorts from a randomized experiment. *Educational Evaluation and Policy Analysis*, *37*, 542–566.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. Hallinan (Ed.), *Frontiers in sociology of education* (pp. 127–162). Dordrecht, Netherlands: Springer Verlag.

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*, 48–54.

Conaway, C., & Goldhaber, D. (2018). *Policy-relevant confidence intervals and the standard of evidence for education policy decision-making* (CEDR Policy Brief No. 04032018-1-2). Seattle: The Center for Education Data and Research, University of Washington Bothell.

Donovan, M. S. (2013). Generating improvement through research and development in educational systems. *Science*, *340*, 317–319.

Every Student Succeeds Act, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Ferguson, R. F. (2008). *The Tripod project framework*. Cambridge, MA: Tripod Project.

Fishman, B. J., Penuel, W. R., Allen, A.-R., & Cheng, B. H. (Eds.). (2013). *Design-based implementation research: Theories, methods, and exemplars*. New York, NY: Teachers College Press.

Fryer, R. G., Jr. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments* (Vol. 2, pp. 95–322). New York, NY: North Holland.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . Sztejnberg, L. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, *43*, 19–23.

Harris, D. N., & Sass, T. R. (2009, September). *What makes for a good teacher and who can tell?* (CALDER Working Paper No. 30). Retrieved from https://www.urban.org/sites/default/files/publication/33276/1001431-What-Makes-for-a-Good-Teacher-and-Who-Can-Tell-.PDF

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *20*, 101–136.

Joyce, B., & Showers, B. (1982). The coaching of teaching. *Educational Leadership*, *40*(1), 4–10.

Kane, T. J. (2016). Connecting to practice. *Education Next*, *16*(2), 80–87.

Kane, T. J., & Staiger, D. O. (2011). *Learning about teaching: Initial findings from the measures of effective teaching project* (Policy and practice brief, MET Project). Seattle, WA: Bill & Melinda Gates Foundation.

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*, 83–119.

Kraft, M. A., & Blazar, D. (2017). Improving teachers' practice across grades and subjects: Experimental evidence on individualized teacher coaching. *Educational Policy*, *31*, 1033–1068.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *8*, 547–588.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*, 158–166.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316.

Mashburn, A. J., Downer, J. T., & Hamre, B. K. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, *14*, 179–196.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, *99*, 210–221.

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

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

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*, 431–451.

Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., LoCasale-Crouch, J., . . . Scott-Little, C. (2017). Early childhood professional development: Coaching and coursework

effects on indicators of children's school readiness. *Early Education and Development*, *28*, 956–975.

Rhoads, C. (2016). The implications of contamination for educational experiments with two levels of nesting. *Journal of Research on Educational Effectiveness*, *9*, 531–555.

Roderick, M., Easton, J. Q., & Sebring, P. B. (2009, February). *The Consortium on Chicago School Research: A new model for the role of research in supporting urban school reform*. Retrieved from https://files.eric.ed.gov/fulltext/ED505883.pdf

Schneider, M. (2017). *A more systematic approach to replicating research: Message from IES director*. Washington, DC: U.S. Department of Education, Institute for Education Sciences.

Showers, B. (1984). *Peer coaching: A strategy for facilitating transfer of training* (A CEPM R&D Report). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Showers, B. (1985). Teachers coaching teachers. *Educational Leadership*, *42*(7), 43–48.

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

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, *11*, 340–359.

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*, 478–501.

Tseng, V. (2012). *Partnerships: Shifting the dynamics between research and practice*. New York, NY: William T. Grant Foundation.

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

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, *37*, 469–479.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Authors

DAVID BLAZAR is an assistant professor of education policy and economics at the University of Maryland College Park. His research and teaching interests include the economics of education, education policy analysis, and applied quantitative methods for causal inference.

MATTHEW A. KRAFT is an associate professor of education. He studies human capital policies in education with a focus on teacher effectiveness and organizational change in K–12 urban public schools.