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

David Blazar
University of Maryland College Park

Matthew A. Kraft
Brown University

Abstract

Over the past 15 years, the education research community has advocated for the application of more 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. However, these goals are often in tension with each other. We propose a research design – the multi-cohort, longitudinal experimental (MCLE) design – as one approach to balancing these competing goals of high-quality research. We illustrate the uses and benefits of MCLEs with an example from a research-practice partnership aimed at evaluating the effect of a teacher coaching program. We find that the coaching program failed to replicate its effectiveness with an initial cohort, likely due to changes in personnel, duration, and content. Our analyses can help researchers weigh the tradeoffs of different design features of MCLEs.

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 inform the practice of partner organizations and policymakers (i.e., *relevant*). However, achieving these three goals can be challenging, in part, because the goals themselves often compete with each other. For example, conducting *rigorous* randomized control trials may limit the *replicability* of a given study because the research is 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 tensions and challenges, we argue that researchers can adopt research designs that attempt to 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. The MCLE design treats programs as fluid rather than static and allows for evolution and re-evaluation over time. This proposal responds directly to increased calls from education researchers to determine research quality based on the combination of design features (e.g., Bryk, Gomez, & Grunow, 2011; Gutiérrez & Penuel, 2014) and, in particular, those that leverage the knowledge of local communities to address problems of practice that schools and districts face (Donovan, 2013; Fishman et al., 2013). We argue that MCLEs are a promising approach to enable greater use of research in decision making (Tseng, 2012).

Of course, the MCLE design is not without limitations. It requires a sustained research-practice partnership, where the program under study works with multiple cohorts that can be randomized to treatment. We illustrate the opportunities and challenges of one MCLEs by describing our application of this design to evaluate a teacher coaching program implemented across three cohorts. As researchers, we worked collaboratively with the program designers and core staff of the MATCH Teacher Coaching program to study its effects in the context of its first implementation in a new setting. MATCH Teacher Coaching 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. Our case study highlights several tensions and tradeoffs that researchers and practitioners must consider when designing future MCLE studies, including issues related to statistical power and an ability to anticipate ex-ante and pre-register hypotheses about specific mechanisms to explore (Gehlbach & Robinson, 2018).

Beyond the methodological contributions of this paper, our substantive findings provide further justification for researchers and practitioners to work together and innovate on research designs aimed at addressing problems of practice. We found that MATCH Teacher Coaching had a positive effect on teachers' instructional practice in the first cohort, but that it failed to replicate these positive results in the latter two cohorts. After ruling out alternative explanations (e.g., differences in characteristics of teachers across cohorts, including interest in receiving coaching), we attribute differences in effectiveness to the turnover of coaches, increased focus on behavior management relative to content and student engagement, and decreased number of coaching sessions. We believe that these findings can inform current efforts by coaching programs to improve their effectiveness while attempting to reduce program costs. These findings also make clear that causal inference on its own is unlikely to lead to better interventions and outcomes, and that evaluation designs must accommodate the dynamic rather than static nature of educational programs.

Tradeoffs 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) methodological *rigor* and research designs that can support causal inferences (Angrist & Pischke, 2008; Every Student Succeeds Act [ESSA], 2015; Kane, 2016; Murnane & Willett, 2011); (2) an ability to *replicate* findings (Camerer et al., 2018; Miguel et al., 2014; Schneider, 2017); and (3) *relevance* for informing specific programs as well as practice and policy more generally, often through research-practice partnerships (Coburn & Penuel, 2016; Snow, 2015; Tseng, 2012).

Over the last 15 years, the education research community has made substantial progress in addressing the first of these three goals.[1] 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. By the author's calculations, in 2000 only 14% of studies reviewed by the What Works Clearinghouse (WWC) – 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 paper. In 2007, a comprehensive review of the entire canon of literature on the effects of teacher PD (*n* =

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

1,300 studies) found only nine studies that met the WWC's highest evidence standards. 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 randomized control trials (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 be as conducive as possible to maximizing effects. In our prior meta-analysis of teacher coaching programs (Kraft et al., 2018), over half of the included studies 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. The limited statistical power of small efficacy trials also constrains researchers' ability to look at mechanisms of effective programming or subgroup effects. As such, decision makers at state and federal levels often are interested in large-scale *effectiveness* trials implemented across a range of settings, which provide greater external validity and generalizability (Wayne et al. 2008). But, by growing in scale, effectiveness trials generally cannot respond directly to the needs and questions generated by local education communities.

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 myriad factors including available resources, school and district conditions, and 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 amongst a second or third cohort of participating teachers. Without tracking and comparing findings over time and across cohorts, it is difficult to produce information that is directly relevant to programming staff 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 estimating results that are relevant to and immediately inform the program or policy under study – can be a considerable challenge. In particular, the very nature of research-practice partnerships means that studies often 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. Practitioners also often require information on mechanisms and factors driving effective or ineffective programs to inform continuous improvement efforts (Wagner, 1997). However, unpacking mechanisms and identifying mediating pathways generally is a challenge in causal research. Mechanisms and implementation factors of interest (e.g., personnel, duration, content) often are self-selected by participants or program staff, and so are endogenous. Program evaluations often are static in nature, while real-world programs are engaged in continuous improvement. Therefore, researchers often evaluate a program before it has time to improve.

A Proposal

RCTs have become the new standard for evaluating educational programs and interventions (Angrist & Pischke, 2009; ESSA, 2015; Kane, 2016; Murnane & Willett, 2011). The standard 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 constraints frequently lead to small-scale designs with limited statistical power. Participants are recruited to be volunteers, sometimes resulting in a highly non-representative sample. Researchers 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 later, often well after program staff has had to make programmatic decisions for the following year. If researchers examine mechanisms in any way, they usually do so in an ad-hoc exploratory way.

Multiple-Cohort, Longitudinal Experiments

We propose an alternative approach to this standard design of RCTs in order to better address the tension between rigor, replication, and relevance: multiple-cohort, longitudinal experimental designs (MCLE). 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 to ongoing program redesign and improvement efforts. Pooling data across multiple cohorts also provides researchers with increased statistical power for detecting smaller effects and tests the replicability of effects across different cohorts. The longitudinal nature of this design allows researchers to examine whether results persist or fade out over time.

Several other alternative research designs provide elements of these design features, but few combine them all into a single design. First, 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, as is the case with MCLEs. However, multi-arm 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. Second, large-scale effectiveness trials with greater external validity can only be conducted for programs that have achieved substantial scale and have secured substantial funding for these resource-intensive studies. Third, more exploratory observational studies of program mechanisms using the full population of participants allow researchers to examine the importance of a range of program features but can be biased due to self-selection and other omitted variables.

Researchers interested in conducting MCLEs will need to work closely with their program partners to establish the continuous improvement goals of the program and what outcomes are most aligned with these goals. 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, a well-designed and well-implemented RCT requires very limited analyses and robustness checks relative to quasi-experimental research designs (Angrist & Pischke, 2009; Murnane & Willett, 2011).

Challenges

Along with the advantages, MCLEs also present several important design and implementation challenges with which researchers will have to contend. Using the multiple-cohort design to identify the effect of changes to the program is challenging. One approach would be to plan for and pre-specify 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. Restricting changes to those that can be identified ex-ante could unnecessarily limit program improvement.

A second challenge is the potentially limited statistical power from analyses using a single cohort and comparing across cohorts. Researchers can address this challenge to some degree, randomizing at the lowest unit possible (e.g. classrooms rather than schools) and collecting outcome measures at baseline to increase the precision of their estimates. The limited statistical power of a single cohort RCT also suggests the importance of collecting outcomes that are proximal and directly aligned with the treatment in addition to primary measures. However, achieving statistical significance may be less relevant for informing ongoing programmatic improvement when organizations are working with limited information and inflexible timelines (Conaway & Goldhaber, 2018).

The multi-cohort feature of MCLEs can create two additional challenges: Some partners may be reluctant to continue randomizing participants to a control condition if the program is shown to be successful in the first year. These type of equity concerns are most easily resolved when there is limited capacity to provide the intervention and randomizing is a fair way to distribute this resource. Multi-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. One approach to minimizing this threat would be to use different research contexts across cohorts. At the same time, prior research suggests the exposure to treatment via peers is unlikely to be a first-order concern for most educational interventions (Rhoads, 2016).

The longitudinal nature of the MCLE design presents a final challenge. Studies that require tracking students or teachers over multiple years can suffer non-trivial sample loss 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.

Given several decisions that researchers will have to make in collaboration with their practice-based partners, below we discuss an illustrative case of the MCLE design.

An Illustrative Case

Intervention and Partnership

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).

Consistent with the longstanding 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 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. During each cycle, coaches observed teachers' instruction and then debriefed at the end of the school day about what they observed. Coaches worked with teachers to set rigorous expectations for growth and then 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 one-to-two weeks via email or phone.

As described in our earlier work with MTC (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, programmatic and evaluation designs were developed in tandem. This work stems from that collaboration and partnership. 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.

Experimental Design

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 at least 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 that we could rely on in data collection efforts. (MCLEs may be designed as blocked RCTs, but not always.) 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 Appendix 1: Table 1, we confirm the success of the randomization process in terms of creating balance between the treatment and control groups. 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* = 0.42 for test that differences in all observable characteristics between treatment and control groups are equal to zero, pooling across cohorts).

Changes in Program Features

One goal of our partnership with MTC and its core program staff was to provide input into the development of their model, which was brand new in the sense that it was being rolled out from the MATCH school in Boston to a new setting (New Orleans) at a greater scale for the first time. Therefore, over the course of the three-year study, several key features were adapted. First, several of the coaches turned over across cohorts, driven both by natural movement in and out of the district 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 four weeks to three 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 support, student engagement). Below, we present data on how these programmatic changes played out in practice. These three changes reflect features specific to MTC but also broader categories of implementation – i.e., personnel, dosage, and content – that are critical components of many educational programs.

One challenge with RCTs generally is that participants are self-selected volunteers, and their characteristics may not match the broader population of interest. For MCLEs that exploit variation in program feature across cohorts, it is possible that the first cohort of participants is more interested in participating relative to those who participate in later cohorts. Differences in this and other characteristics of participants across cohorts is a problem not only for generalizability but also for internal validity. That is, differences in treatment effect estimates across cohorts may be due to characteristics of participants rather than program features.

However, in our study, this does not appear to be a concern. In Appendix 1: Table 1, we show that characteristics of teachers, including the self-reported level of interest in participating in MTC, are similar across cohorts. Two exceptions are years of teaching experience and certification through alternative versus traditional routes; however, in prior work (Blazar & Kraft, 2015), we show that differences in these characteristics do not explain differences in effectiveness across cohorts. Researchers who use the MCLE design in future studies will need to collect characteristics of participants at baseline and compare these characteristics across cohorts.

Data

We used three primary sources of data to triangulate the effect of MTC on teachers' practices. (See Appendix 2 for additional information and details of these data, including reliability statistics.) 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*. 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. It was not possible in our study to evaluate the effect of coaching on student test scores, given that the program served teachers across grades and subject areas. Therefore, many teachers did not prepare students for state tests, and even if they did there was no guarantee that their randomization block partner(s) also had student test scores. Instead, we 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.

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). 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.

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, by controlling for these measures we were able to increase statistical power by explaining residual variation in our outcomes. This approach is particularly useful in smaller-scale efficacy trials. 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.

A second possible threat to internal validity of an RCT – in addition to baseline balance between treatment and control groups – is attrition from the sample. In our case attrition occurred either because teachers dropped from the study (e.g., left teaching, chose not to participate) or were unable to participate in data collection. In tables presented in Appendix 1, we assess the extent to which attrition and missing data might bias our results. First, we examined differential rates of missing data between treatment and control groups (Appendix 1: Table 2), where we found no differences at the end of the coaching year but some differences in availability of data in the follow-up year between treatment and control groups, particularly in cohorts 1 and 3. Second, we examined whether attriters from the treatment group differed from attriters in the control group on observable characteristics. To do so, we regressed each observable characteristic on indicators for attrition and treatment status, and their interaction. We report coefficients and *p*-values on the interaction term in Appendix 1: Table 3. We find some evidence that treatment group teachers who were missing data at the end of the coaching year had lower overall interest in coaching and tended to be younger than control group teachers who were missing data (pooling across all cohorts). In light of these findings, we exclude follow-up estimates in cohort 3 from our main result figures but include estimates in an appendix.

Analyses

The randomized design allowed for a straightforward approach to estimating the causal effect of MTC on teacher and student outcomes. (See Appendix 3 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 block 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 in all analyses and included teacher- and classroom-level random effects in the student-level analyses. We both pooled and disaggregated results by cohort in order to examine whether results replicate across cohorts.

If MTC had varied just one program feature across cohorts, then 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. Instead, to examine whether pre-determined implementation features were related to differences in outcomes across cohorts, we first examined qualitatively changes in program characteristics across cohorts. We supplemented these data with quantitative analyses that modeled changes in outcomes as a function of program characteristics, with slight modifications to the regression models described above. 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 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.

Results

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 Figure 1, 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 0.59 SD than control group teachers on the *Summary Index* of effective teaching practices. (See Appendix 1 for corresponding tables with regression coefficients and associated standard errors). 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 (see Figure 2). 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 (see Appendix 1: Table 4).

We also observe cross-cohort differences in the effect of the MTC program in the spring of the follow-up 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 (one exception is for *Achievement of the Lesson Aim*). For cohort 2, no point estimates in the follow-up year are statistically significantly different from zero, and several of these point estimates can be distinguished from the follow-up effect for cohort 1 (see Appendix 1: Table 4). We exclude from Figure 1 estimates for the follow-up year in cohort 3 given concerns that very high attrition rates of treatment teachers relative to control group teachers leads us to substantially under-state these follow-up effects (see Appendix 1: Table 3). However, we include this point estimate in the appendix, where we observe that all but one point estimate is negative in magnitude, and two (including for the *Summary Index*) are statistically significantly different from zero.

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 instructional focus areas across cohorts. In Appendix 1: Table 5, 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.

We hypothesize that dimensions of effective coaching include the knowledge and skills that coaches bring to their work as well as their ability to have productive interactions and develop strong interpersonal relationships with teachers. While our study was not designed to test these relationships, descriptive analyses of implementation data show some differences in the tools that coaches used with teachers (see Figure 3). 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.

A second key change between cohorts 1 versus cohorts 2 and 3 that may explain differences in effectiveness is the instructional focus areas of coaching sessions. In Figure 4, we 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 of content, and student engagement. Coaches started out the year with a stronger focus on behavior management and decreased focus on this area 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 the start of and throughout the school year.

In Appendix 1: Table 6, we present estimates from a model of the relationship between the number of sessions focused on each classroom practice area. 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 an increased focus 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.

The third feature of MTC that varied across cohorts was dosage. As shown in Figure 5, teachers in cohort 1 received four weeks of coaching, on average, while those in cohorts 2 and 3 received three weeks of coaching, on average. However, the design of our study does not allow us to tease out the effect of dosage from differences in coach effectiveness or focus on different classroom practices.

Discussion

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 on 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, several implementation features related to personnel, duration, and content appear to mediate these effects. We find that MTC became less effective due to turnover of coaches and a greater focus on behavior management relative to other instructional areas (i.e., instructional support, student engagement). While implementation almost always is considered endogenous due to the self-selection of these features, leveraging variation in implementation across three cohorts of the experiment provides a unique opportunity to rigorously test the efficacy of these program features. In our own partnership with MTC and in research-practice partnerships more broadly (Tseng, 2012), this is the sort of information that is necessary to drive continuous improvement efforts.

We use this specific case as an illustration of our proposed research design, MCLEs, that we argue can serve as an example for future research-practice partnerships about how 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 a range of options that most closely align with their own continuous improvement efforts.

Our study also illustrates some of the challenges of executing MCLE designs. The nature of our research-practice partnership with MTC – and of many continuous improvement efforts (Wagner, 1997) – meant that we discovered mechanisms to explore empirically during the course of the study. We agree with others regarding the importance of pre-registering analyses in education research (Gehlbach & Robinson, 2018), particularly for confirmatory and causal work. Pre-registration helps guard against endogenous selection or context-specific changes across cohorts. Without pre-registration of the cross-cohort changes in programming, we view our implementation analyses as exploratory. We encourage future research of this kind to use theory and exploratory analyses to identify ex-ante a likely set of implementation mechanisms to vary across cohorts. For example, our work with MTC has led to several additional analyses related to coaching duration and coach quality, which we see as critical when considering how best to scale-up teacher coaching programs. At the same time, we encourage scholars to adapt these pre-registered plans when new insights arise during the course of study about how best to modify the program for future cohorts. We recognize that results from a first cohort may lead to a new understanding of program effects and interest in additional research questions not anticipated prior to that first cohort.

A second tension we identified relates to sample size and statistical power. 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 require relatively precise estimates to establish that differences in treatment effects across cohorts are statistically significant from each other. In our context sample sizes were limited by the capacity of MTC coaches and our ability to recruit school to participate in the study.

In turn, we make several recommendations for researchers and program staff looking to use MCLE designs to balance methodological rigor, replication, and relevance, which we roughly organize in temporal order:

1. Develop relationships with practitioners, and co-design research in advance.
2. As part of co-planning, identify ex-ante likely mechanisms of effective programming (e.g., personnel, duration, content), and make a plan for changing some of these features across cohorts. Ideally, manipulate program characteristics separately, rather than as a group, in order to be able to tease out the role of one versus another.
3. Identify data sources to closely monitor implementation of these program changes and to examine how they play out in practice.
4. Survey participants at baseline on a range of characteristics including motivation and interest in the study to examine how cohorts differ over time.
5. Be conscious of and examine the degree to which the RCT samples are similar across cohorts, and representative of a broader population of interest.
6. Develop outcome measures that are both proximal to the intervention under study and more distal outcomes of policy relevance (i.e., student test scores, student classroom experiences).
7. Track outcomes at baseline (to increase statistical power), at the end of the intervention year (to detect immediate effects), and in the year(s) after the intervention ends (to examine whether effects persist or fade out over time).

8. Share results both pooled and disaggregated by cohorts, including null findings, so that others can learn from prior work.
9. If new research questions arise after interpreting results from an early cohort, amend pre-registration plans about mechanisms to explore in a future cohort.
10. Leverage administrative data to capture long-term outcomes among participant populations.
11. Communicate results in a timely and accessible way to partner organizations.

Finally, although we find value in exploiting cross-cohort differences in the effectiveness of MTC, failure to replicate raises concerns regarding findings from small pilot studies in education research. Our evaluation was designed under a best-case scenario, with the same evaluators and program. Yet, even here effects differed substantially. We recognize that the design of pilot studies often are purposeful as findings can be very useful for the practitioners and policymakers who designed the program. Small-scale programs, though, also stem from budget constraints and other real-world scenarios that result in causal research in non-representative settings. Ultimately, drawing conclusions about the benefit of any given type of education intervention (e.g., teacher coaching) 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 or prevent a premature policy rush to scale up programs with limited evidence from small efficacy trials.

References

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., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. 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*(4), 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, the Netherlands: 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*(1), 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, WA: 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: The Tripod Project.

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

Fryer Jr, R. G. (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). North-Holland.

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

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, *43*(1), 19-23.

Harris, D.N., & Sass, T.R. (2009). *What makes for a good teacher and who can tell?* CALDER Working Paper No. 30.

Jacob B. A., & Lefgren L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 20*(1), 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).

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. 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). Individualized coaching to improve teacher practice across grades and subjects: New experimental evidence. *Educational Policy, 31*(7), 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*, *88*(4), 547-588.

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.

Rhoads, C. (2016). The implications of contamination for educational experiments with two levels of nesting. *Journal of Research on Educational Effectiveness*, 9(4), 531-555.

Schneider, M. (2017). *A more systematic approach to replicating research: Message from IES director*. 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*(9), 460-466.

Tseng, V. (2012). P*artnerships: 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*(8), 469-479.

Figures

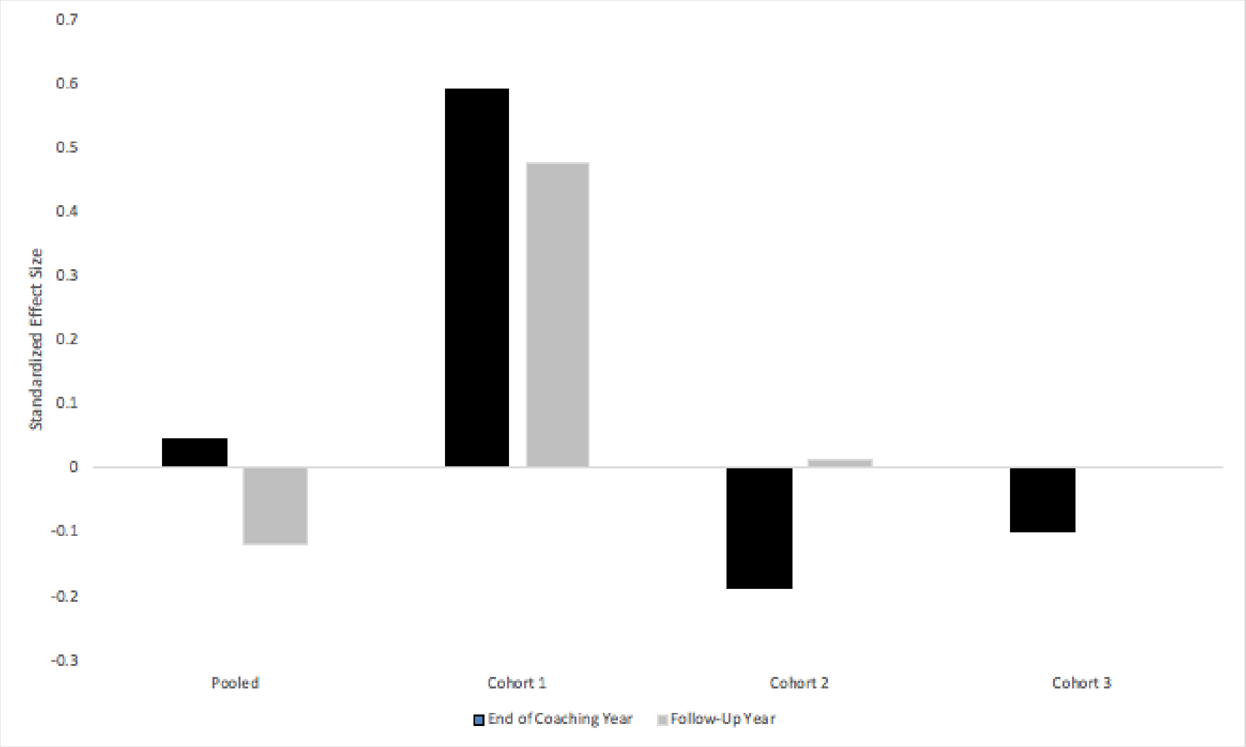

Figure 1. Standardized effect sizes of the effect of MTC on the *Summary Index* of effective teacher practices, by cohort and year.

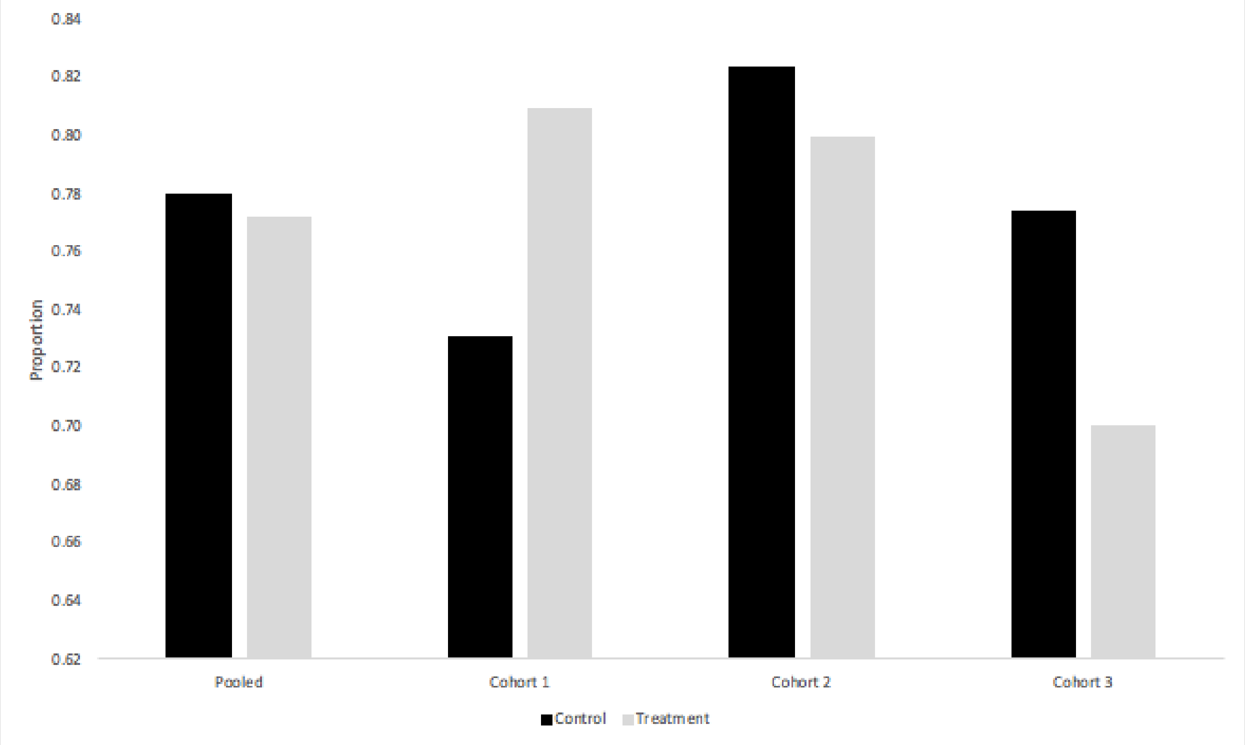

Figure 2. Proportion of students who felt that they "learned a lot in class every day," by treatment versus control and cohort; end of Year 1 only.

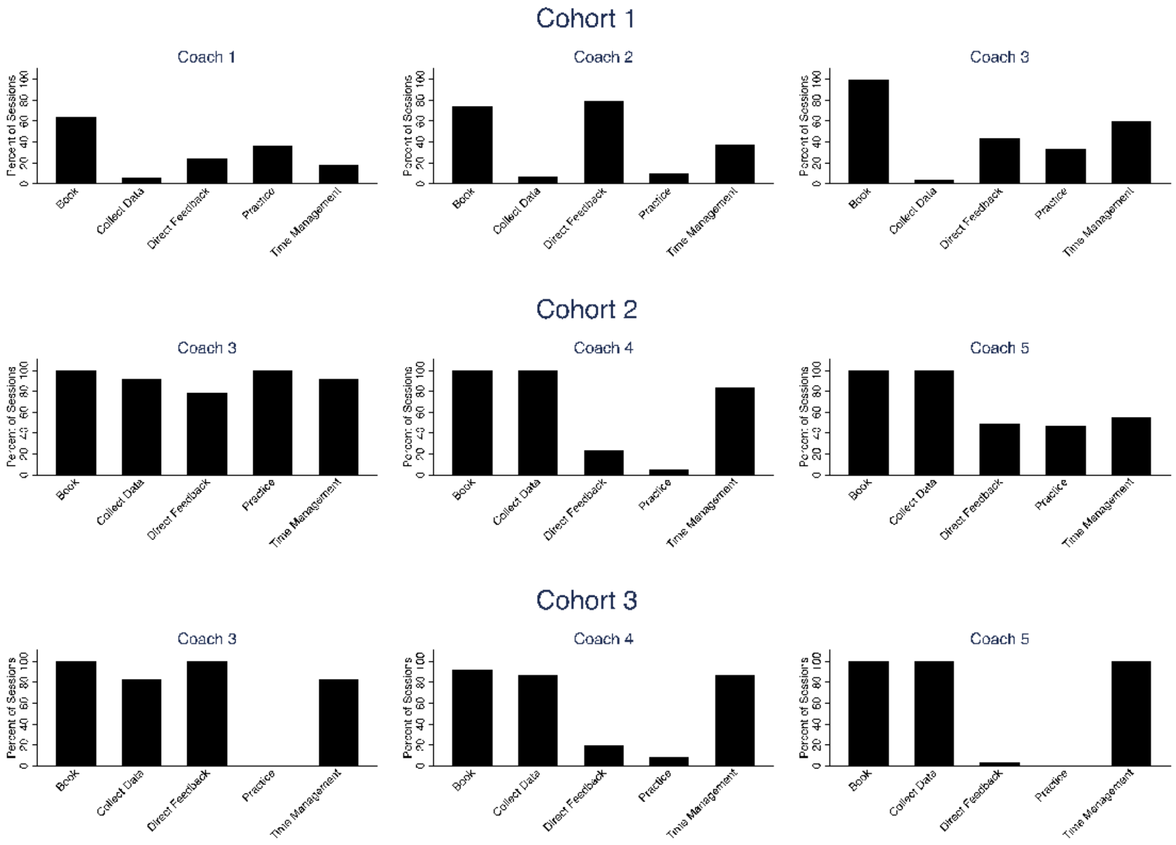

Figure 3. Techniques used in debriefing sessions with teachers, by coach and cohort.

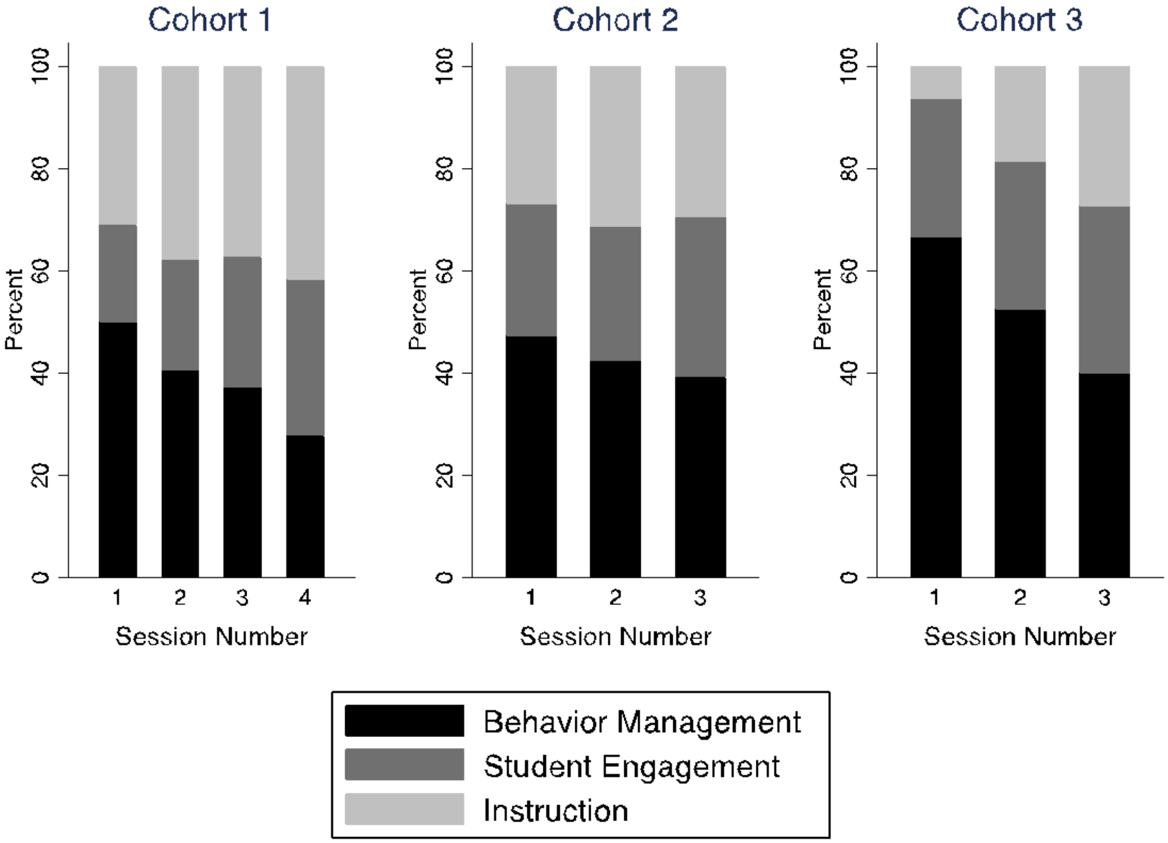

Figure 4. Changes in the content of coaching across coaching sessions, by cohort.

Appendix 1: Tables

Appendix Table 1
Teacher Characteristics and Balance Between Treatment Groups and Cohorts

	Pooled 3 Cohorts		Cohort 1		Cohort 2		Cohort 3		*p*-value on
	Treatment Mean	Difference for Control Group	Treatment Mean	Difference for Control Group	Treatment Mean	Difference for Control Group	Treatment Mean	Difference for Control Group	Joint Difference between All Cohorts
Teacher Background Characteristics									
Interest in Coaching (1-10 scale)	9.16	-0.03	9.23	-0.25	9.09	0.01	9.19	0.09	0.63
Female (%)	0.73	0.00	0.70	0.09	0.69	0.00	0.81	-0.07	0.45
African American (%)	0.19	0.01	0.20	-0.06	0.14	0.08	0.24	-0.01	0.55
White (%)	0.72	0.01	0.77	-0.01	0.75	0.00	0.65	0.04	0.38
Age (years)	25.29	1.10	26.13	-0.06	24.86	0.88	25.19	2.20	0.29
Teaching Experience (years)	2.83	0.30	3.93	0.07	2.57	0.14	2.27	0.70	0.00
Alternatively Certified (%)	0.76	-0.04	0.80	-0.08	0.82	-0.02	0.65	-0.03	0.03
Masters' Degree (%)	0.22	0.02	0.20	0.04	0.25	0.00	0.19	0.04	0.86
College Instution Ranked Very Competitive or Higher (%)	0.80	-0.08	0.73	0.06	0.84	-0.10	0.81	-0.17	0.63
Teaching and School Characteristics									
Teach All Subjects (%)	0.36	-0.01	0.43	-0.02	0.33	0.01	0.35	-0.02	0.47
Teach Humanities (%)	0.41	-0.02	0.37	-0.02	0.47	-0.07	0.35	0.03	0.53
Teach STEM (%)	0.28	0.08	0.20	0.04	0.31	0.09	0.30	0.09	0.20
P-value from joint test		0.42		0.56		0.61		0.25	
n(teachers)	116	113	30	29	49	45	37	39	

Notes: Treatment and control group means are estimated from regression models that control for randomization block fixed effects.

Appendix Table 2
Proportion of Teachers with Outcome Data on Different Measures

	Pooled 3 Cohorts		Cohort 1		Cohort 2		Cohort 3	
	Treatment Mean	Difference for Control Group	Treatment Mean	Difference for Control Group	Treatment Mean	Difference for Control Group	Treatment Mean	Difference for Control Group
				Panel A: Spring of Intervention Year				
MATCH Teacher Observation Rubric	0.888	-0.047	0.933	-0.106	0.918	-0.096	0.811	0.061
Principal Survey of Teachers	0.862	-0.039	0.933	-0.106	0.898	-0.098	0.757	0.089
Tripod Student Survey of Teachers	0.759	-0.006	0.867	-0.039	0.714	-0.048	0.730	0.065
				Panel B: Spring of Follow-Up Year				
MATCH Teacher Observation Rubric	0.466	-0.023	0.667	-0.287*	0.490	-0.179	0.270	0.371***
Principal Survey of Teachers	0.448	0.012	0.700	-0.286*	0.429	-0.095	0.270	0.371***
Tripod Student Survey of Teachers	0.457	-0.014	0.700	-0.286*	0.449	-0.160	0.270	0.371***
n(teachers)	116	113	30	29	49	45	37	39

Appendix Table 3

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* -Value	Interaction Coefficient	*p* -Value	Interaction Coefficient	*p* -Value	Interaction Coefficient	*p* -Value
				Panel A: Spring of Intervention Year				
Interest in Coaching (1-10 scale)	-0.70	0.07	-0.84	0.36	-0.63	0.27	-0.37	0.57
Female (%)	0.22	0.23	0.29	0.52	-0.02	0.96	0.32	0.29
African American (%)	0.01	0.93	0.01	0.98	-0.40	0.11	0.23	0.38
White (%)	-0.13	0.49	-0.54	0.15	0.28	0.35	-0.31	0.32
Age (years)	-3.24	0.08	-5.35	0.10	0.56	0.80	-7.77	0.06
Teaching Experience (years)	-1.47	0.11	-2.00	0.28	0.97	0.51	-5.21	0.00
Alternatively Certified (%)	-0.07	0.69	0.31	0.46	-0.18	0.47	-0.04	0.91
Masters' Degree (%)	-0.06	0.73	-0.50	0.22	0.47	0.10	-0.34	0.21
College Instution Ranked Very Competitive or Higher (%)	-0.05	0.77	0.13	0.76	0.18	0.48	-0.29	0.34
				Panel B: Spring of Follow-Up Year				
Interest in Coaching (1-10 scale)	-0.11	0.72	-1.25	0.04	0.51	0.22	-0.07	0.91
Female (%)	0.17	0.26	-0.24	0.43	-0.12	0.59	0.90	0.00
African American (%)	0.01	0.95	0.07	0.74	0.02	0.92	-0.02	0.94
White (%)	-0.14	0.33	-0.14	0.59	-0.15	0.50	-0.06	0.83
Age (years)	-1.23	0.40	0.64	0.77	0.71	0.65	-4.89	0.20
Teaching Experience (years)	0.28	0.70	0.14	0.91	1.77	0.10	-1.74	0.26
Alternatively Certified (%)	-0.13	0.38	-0.23	0.42	-0.07	0.72	-0.11	0.73
Masters' Degree (%)	-0.06	0.65	-0.23	0.41	0.04	0.85	-0.02	0.94
College Instution Ranked Very Competitive or Higher (%)	0.08	0.58	0.14	0.62	0.25	0.22	-0.29	0.31
n(teachers)	116	113	30	29	49	45	37	39

Notes: 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 block.

Appendix Table 4
Parameter Estimates of the Effect of MATCH Teacher Coaching

		MATCH 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	0.045	0.163	0.236	-0.066	-0.056	-0.004	-0.007
	(0.144)	(0.164)	(0.143)	(0.142)	(0.053)	(0.045)	(0.018)
				Regression 2: Results by Cohort			
Treat*Cohort 1	0.589*	0.579	0.663*	0.293	0.130	0.327***	0.085**
	(0.243)	(0.316)	(0.319)	(0.172)	(0.101)	(0.073)	(0.029)
Treat*Cohort 2	-0.187	-0.170	0.033	-0.272	-0.069	-0.132	-0.040
	(0.237)	(0.246)	(0.198)	(0.290)	(0.099)	(0.082)	(0.031)
Treat*Cohort 3	-0.099	0.234	0.148	-0.100	-0.195*	-0.145*	-0.048
	(0.211)	(0.288)	(0.227)	(0.161)	(0.083)	(0.071)	(0.032)
P-value on test between cohort coefficients:							
Cohort 1 versus Cohort 2	0.025	0.503	0.066	0.127	0.161	0.435	0.000
Cohort 1 versus Cohort 3	0.036	0.008	0.424	0.082	0.014	0.288	0.000
Cohort 2 versus Cohort 3	0.782	0.073	0.291	0.541	0.364	0.038	0.910
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	-0.119	0.271	-0.118	-0.084	-0.115	-0.012	0.029
	(0.332)	(0.284)	(0.339)	(0.363)	(0.078)	(0.080)	(0.037)
				Regression 2: Results by Cohort			
Treat*Cohort 1	0.476	0.955*	0.552	0.240	-0.066	0.161	0.114
	(0.462)	(0.391)	(0.680)	(0.467)	(0.141)	(0.158)	(0.069)
Treat*Cohort 2	0.013	0.132	-0.233	0.082	-0.225	-0.035	0.006
	(0.507)	(0.359)	(0.458)	(0.677)	(0.127)	(0.124)	(0.057)
Treat*Cohort 3	-1.078**	-0.288	-0.656	-0.748	0.169	-0.240**	-0.044
	(0.316)	(0.581)	(0.548)	(0.403)	(0.146)	(0.087)	(0.044)
P-value on test between cohort coefficients:							
Cohort 1 versus Cohort 2	0.098	0.343	0.099	0.848	0.361	0.004	0.241
Cohort 1 versus Cohort 3	0.193	0.172	0.100	0.115	0.038	0.003	0.058
Cohort 2 versus Cohort 3	0.703	0.556	0.607	0.297	0.178	0.871	0.502
n (teachers)	107	103	103	103	88	88	88
n (students)	--	--	--	--	2773	2781	2709

Notes: *p<0.05, **p<0.01, ***p<0.001. All regression models include fixed effects for randomization block. Standard errors clustered by school-year in parentheses. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains.

Appendix Table 5

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	0.427	0.408	0.255	-0.177	-0.068	-0.036
	(0.235)	(0.256)	(0.248)	(0.325)	(0.162)	(0.143)	(0.052)
Coach 2	0.544*	0.988**	1.095***	-0.339	0.319*	0.453***	0.116**
	(0.239)	(0.320)	(0.267)	(0.298)	(0.146)	(0.113)	(0.037)
Coach 3	0.379	0.404	0.519	0.256	-0.133	-0.007	0.028
	(0.270)	(0.293)	(0.275)	(0.258)	(0.117)	(0.100)	(0.039)
Coach 4	-0.340*	-0.227	-0.169	-0.236	-0.052	-0.076	-0.022
	(0.164)	(0.204)	(0.147)	(0.174)	(0.071)	(0.071)	(0.033)
Coach 5	0.334	0.315	0.321	0.319	-0.150	-0.078	-0.051
	(0.197)	(0.210)	(0.167)	(0.222)	(0.114)	(0.082)	(0.031)
P-value on test between coach coefficients	0.012	0.024	0.001	0.189	0.105	0.003	0.010
n (teachers)	199	196	197	192	173	173	173
n (students)	--	--	--	--	5249	5261	5147

Notes: + p<0.1, *p<0.05, **p<0.01, ***p<0.001. 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. Standard errors clustered by school-year in parentheses. All regressions include fixed effects for cohort. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains.

Appendix Table 6

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

		MATCH Rubric		Principal Survey	TRIPOD Student Survey		
	Summary Index	Achievement of Lesson Aim	Behavioral Climate	Overall Effectiveness Composite	Challenge	Control	Learn a Lot
Behavior Management	-0.139	-0.236*	-0.243**	0.063	-0.133**	-0.116*	-0.042**
	(0.093)	(0.106)	(0.085)	(0.105)	(0.046)	(0.046)	(0.014)
Instruction	-0.158	-0.160	-0.094	-0.116	0.006	-0.022	0.002
	(0.084)	(0.107)	(0.087)	(0.071)	(0.042)	(0.034)	(0.013)
Student Engagement	0.271*	0.215	0.333**	0.163	0.074	0.092*	0.036*
	(0.116)	(0.122)	(0.114)	(0.100)	(0.041)	(0.041)	(0.015)
Number of Weeks of Coaching	0.107	0.237*	0.166*	-0.053	0.053	0.066	0.016
	(0.093)	(0.100)	(0.081)	(0.100)	(0.044)	(0.041)	(0.014)
P-values for tests between focus area coefficients							
Behavior Management = Instruction	0.004	0.010	0.000	0.345	0.000	0.001	0.000
Behavior Management = Student Engagement	0.888	0.645	0.248	0.226	0.043	0.131	0.032
Instruction = Student Enagement	0.013	0.046	0.014	0.067	0.329	0.058	0.160
n (teachers)	199	196	197	192	173	173	173
n (students)	--	--	--	--	5,249	5,261	5,147

Notes: + p<0.1, *p<0.05, **p<0.01, ***p<0.001. 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. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains.

Appendix 2: 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 comprised of two overall codes, *Achievement of Lesson Aim* and *Behavioral Climate*. Each code is scored holistically on a scale of 1-10 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 follow-up spring, 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 ten 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 to 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 re-standardizing 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 impact funding for the program or any school evaluation.

Tripod Student Survey

The Tripod survey (Ferguson, 2008) is comprised 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 five-point Likert 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-scale response options for each item, and calculating the mean response across items. We then re-standardized 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.

Appendix 3: Methods and Analysis

We estimated the effect of MTC on our outcomes of interest using Ordinary Least Squares (OLS) and multi-level 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*:

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

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 to control group teachers in their same block and cohort. We omitted random effects for the schools where teachers worked during the coaching year because they were highly collinear with our blocking indicators. However, we clustered standard errors at the school-year level in the current year. 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 by fitting an analogous multilevel model where students, *i*, were nested within classrooms, *c*, and teachers, *j*:

$$A_{it} = \beta MTC_j + \alpha_{s,t=0} + (\nu_j + \varphi_c + \varepsilon_{it}) \qquad (2)$$

As noted above, we did not include a baseline measure, as the student survey was administered only once at the end of the school year. To account for the nested nature of the data, we included random effects for teachers, ν_j, and classrooms, φ_c. We again clustered our standard errors at the school-year level in the current year.

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 (TOT). 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 (see Blazar & Kraft, 2017), as well as analyses exploring differential attrition between treatment and control groups, then we can calculate TOT estimates by scaling our ITT estimates by the inverse of the take-up rate. We both pool and disaggregated results by cohort, allowing us to examine whether results replicated across cohorts.

To examine whether pre-determined implementation features drove differences in outcomes across cohorts, we predicted outcomes 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:

$$Y_j = Y_{j,t=0} + \beta COACHING\_CHARACTERISTIC_j + \delta_h + \varepsilon_j \qquad (3)$$

$$A_i = \beta COACHING\_CHARACTERISTIC_j + \delta_h + (\nu_j + \varphi_c + \varepsilon_i) \qquad (4)$$

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 numbers 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, to hold constant any differences in outcomes across years due to, for example, differences in classroom raters across years.