Educational Evaluation and Policy Analysis
December 2015, Vol. 37, No. 4, pp. 542–566
DOI: 10.3102/0162373715579487
© 2015 AERA. http://eepa.aera.net

Exploring Mechanisms of Effective Teacher Coaching: A Tale of Two Cohorts From a Randomized Experiment

David Blazar
Harvard Graduate School of Education

Matthew A. Kraft
Brown University

Although previous research has shown that teacher coaching can improve teaching practices and student achievement, little is known about specific features of effective coaching programs. We estimate the impact of MATCH Teacher Coaching (MTC) on a range of teacher practices using a blocked randomized trial and explore how changes in the coaching model across two cohorts are related to program effects. Findings indicate large positive effects in Cohort 1 but no effects in Cohort 2. After ruling out explanations related to the research design, a set of exploratory analyses suggest that differential treatment effects may be attributable to differences in coach effectiveness, coaching dosage, and the focus of coaching across cohorts.

Keywords: *teacher coaching, professional development, randomized control trial, causal mechanisms*

Teacher coaching is considered a high-quality professional development opportunity that emphasizes job-embedded practice, intense and sustained durations, and active learning (Desimone, 2009; Russo, 2004). Generally, coaches observe teachers in their classes and then provide targeted feedback aimed at improving these practices. Coaching is closely related to teacher mentoring, which also targets instructional improvement through one-on-one relationships between a novice and more veteran teacher; however, mentoring often focuses on providing general advice rather than responding directly to observed classroom practices (Wildman, Magliaro, Niles, & Niles, 1992). To date, experimental evidence on teacher coaching has been largely positive (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Campbell & Malkus, 2011; Neuman & Cunningham, 2009; Powell, Diamond, Burchinal, & Koehler, 2010; Sailors & Price, 2010). This is particularly noteworthy compared with mixed results on the effectiveness of school workshops and trainings that characterize much of the professional development offerings provided to teachers (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007), as well as more intensive development and mentoring opportunities (Garet et al., 2008; Garet et al., 2011; Glazerman et al., 2008).

At the same time, this research base is new, and little is known about the effectiveness of specific design features or practices of different coaching models. Understanding which program features are critical for success is an important line of inquiry for the continued improvement of teacher professional development. This is especially true given the substantial costs of coaching (Allen et al., 2011) and the fact that coaching is being adopted widely by schools, districts, and teacher preparation programs such as the Long Beach Unified School District, charter management organizations (e.g., Aspire, KIPP, Uncommon Schools, YES Prep), Teach for America, and The New Teacher Project (Lake et al., 2012; Maier, Cellini, & Grogan, 2012; Sawchuk, 2009; Smith, 2013).

In this article, we estimate the impact of MATCH Teacher Coaching (MTC) on a range of teacher practices using a blocked randomized trial and explore how changes in the coaching model across two cohorts are related to program effects. In the 2011–2012 and 2012–2013 school years, coaches worked with treatment teachers in charter schools across the Recovery School District in New Orleans on improving practices common across grades and subjects, including behavior management, instructional delivery, and student engagement. We explore whether this program is effective at improving teachers' practices by drawing on classroom observations, principal evaluations, and student surveys. We chose to focus on these process measures and subjective ratings, instead of test-score outcomes, given that our primary objective is to examine whether the coaching program improves teachers' practices across a wide range of grades and subjects. Use of these measures also allows us to triangulate the effect of coaching on a range of practices.

Importantly, several significant changes in the design and delivery of the coaching model in the second cohort provide a unique opportunity to explore potential mechanisms by which coaching may lead to improved teacher practice. By design, the scale of the MTC program increased between the 2 years, with 49 teachers offered coaching in Cohort 2 compared with 30 teachers in Cohort 1. To accommodate this change, MTC reduced the average amount of coaching it provided to teachers from 4 weeks to 3 weeks throughout the school year and increased teacher-to-coach ratios. In addition, all of the coaches except for the program director changed across years. Finally, programmatic changes resulted in an increased focus on behavior management over other classroom practices. All of these were strategic changes made by the MTC staff rather than a targeted response to the specific set of teachers who participated in the second cohort. Therefore, between-cohort differences reflect plausibly exogenous variation in program characteristics, which we exploit in our analyses.

Results indicate no effect of coaching on any of our outcome measures when data are pooled across all teachers. However, this finding masks substantial variability in the effectiveness of coaching across cohorts. For Cohort 1, we find that coached teachers scored 0.56 *SD*s higher on a summary index of effective teacher practices. In contrast, we find no effect of coaching among Cohort 2 teachers. By ruling out explanations related to the research design (i.e., differences in the counterfactual and potential spillover effects, the sample of teachers included in each cohort, randomization block outliers), we attribute differential treatment effects to changes in the program model, which we describe in detail using a rich set of qualitative data including coaching logs and conversations with coaches. Although we lack the statistical power to conduct a thorough heterogeneity analysis and, therefore, cannot determine with certainty which of the components listed above contribute most to these results, a set of exploratory analyses provide suggestive evidence of differences in treatment effects by coach, coaching dosage, and the focus of coaching. To our knowledge, this is the first article to document variation in program effectiveness across individual coaches and coaching content with empirical evidence. We discuss the implications of these findings for policy and practice.

Background and Context

Although teacher coaching is gaining appeal as a way to develop a range of teachers' practices (e.g., Lake et al., 2012; Maier et al., 2012; Sawchuk, 2009; Smith, 2013), the experimental evaluation literature has focused on coaching's effectiveness in a few key areas. In particular, the bulk of this work has examined early literacy coaching models such as Reading First, the Literacy Collaborative, and Content-Focused Coaching. Sailors and Price (2010) found that classroom-based support around reading and comprehension strategies improved these practices by 0.64 to 0.78 *SD*. Neuman and Cunningham (2009) and Powell et al. (2010) identified similar results for teachers' literacy practices at the preschool level; the former study also found effects of between 0.18 and 0.22 *SD*s on students' early literacy skills.

Although research on other content areas has lagged behind, Campbell and Malkus (2011) found that 2 years of on-site coaching on mathematical content knowledge, pedagogy, and curriculum by trained mathematics coaches increased student achievement between 0.14 and 0.19 *SD*s. Focusing on noncontent-specific teaching practices, Gregory, Allen, Mikami, Hafen, and Pianta (2014) found that Web-based coaching around teacher–student relationships increased teachers' in-class behaviors upward of 0.25 *SD* at the end of the coaching year. Allen and colleagues (2011) also found positive effects of this program on student achievement of 0.22 *SD* in the postintervention year.

Despite growing evidence of the benefits of high-quality coaching, questions remain about the efficacy of different types of coaching programs. For example, studies have not compared the relative benefit for teachers of coaching geared toward content-specific knowledge and pedagogy versus general teaching skills. The high cost of coaching also raises important questions about the optimal design of program features such as coaching dosage and teacher-to-coach ratios that maximize effects relative to costs.

We are aware of just two studies that examine the characteristics of coaching programs that may lead to desired outcomes. In their descriptive, cross-sectional study of a literacy coaching program, Marsh and colleagues (2008) found that teachers' assessments of coach quality were related to teachers' self-assessments of the effect of coaching on their instruction. This finding provides some suggestive evidence on the variability of coach quality, although it is limited by the self-report nature of the data and the lack of an experimental design. In the early childhood setting, Ramey and colleagues (2011) found that teachers randomly assigned to an immersion, high-density coaching program over 5 weeks showed larger gains in classroom quality relative to teachers who received the same number of total hours of coaching but spread out over 20 weeks. Similarly, research indicates that the dosage of standard professional development offerings is related to program effectiveness (Garet, Porter, Desimone, Birman, & Yoon, 2001; Yoon et al., 2007).

Our study builds on this prior work in two key ways. First, we estimate effects of a new coaching program focused on skills common across grades and subjects. Second, we examine whether there exist differences in treatment effects by specific characteristics of the coaching program that vary across cohorts. Results can inform efforts to expand teacher coaching as a core component of professional development efforts.

Research Design

MTC

As described in prior work (see Kraft & Blazar, 2014), MTC is an individualized coaching program focused on improving teachers' practices common across grades and subjects, including classroom management and general pedagogical practices. Three coaches in each cohort (with five coaches total across the two cohorts) worked with participating teachers during a 4-day training workshop over the summer and then one-on-one for either three or four intensive, weeklong 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 1 to 2 weeks via email or phone.

From its inception, the developers and funders of MTC have been very 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/or students, and whether there were specific components of the program that could be improved. As such, programmatic and evaluation designs were developed in tandem. This work stems from that collaboration. Prior to beginning the evaluation, we provided MATCH and made publicly available a report (available upon request) that outlined our research design, including outcome measures designated for confirmatory analyses.

Sample

Participating teachers came from charter schools across the Recovery School District in New Orleans. The Recovery School District is a statewide district in Louisiana formed in 2003 to transform underperforming schools, the vast majority of which are in New Orleans and are operated as charters. In partnership with New Schools for New Orleans, MTC coaches recruited teachers of all grade levels and subject areas but with a focus on early- and midcareer teachers—a population known to require on-site support and assistance (Kaufman, Johnson, Kardos, Liu, & Peske, 2002). Given the capacity constraints with three coaches, MTC staff chose to limit the pool of teachers who would be eligible to receive coaching to those teachers who expressed high levels of interest in the program, completed all required paperwork, and received permission from their principal. In Cohort 1, this restriction resulted in a final sample of 59 teachers from 20 schools. In planning for Cohort 2, program leaders reduced the number of weeks of coaching a teacher received from four to three to provide coaching to a larger group of teachers. Using these sample selection criteria, coaches recruited and selected 94 teachers to participate in Cohort 2. None of these teachers were members of either the treatment or control group in Cohort 1. These teachers worked across 25 schools, 17 of which were the same as those in Cohort 1.

Among the participating teachers in each cohort, we randomly assigned half 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. Three of the 40 total blocks consisted of teachers from multiple school sites. This was true when there was only one teacher at a given school or, in Cohort 2, when we recruited additional teachers after the initial round of randomization.

In Tables 1 and 2, we present descriptive statistics on participating teachers and schools, respectively. Thirty-three percent of all teachers taught humanities, and 23% taught science, technology, engineering, and mathematics (STEM) subjects (i.e., science or math). In all, 71% were female, 76% were White, and 18% were African American. Over three fourths of the teachers entered the profession through alternative certification programs, such as Teach for America or TeachNOLA, and attended an undergraduate institution whose admission process is rated as “Very Competitive” or higher by Barron’s rankings. Twenty-four percent held a master’s degree. Comparing across cohorts, we find that the samples of participating teachers were fairly similar on observable characteristics. The only variable for which we detect a statistically significant difference between cohorts is teacher experience, where 27% of teachers in Cohort 1 were in their first or second year of teaching, compared with 63% of teachers in Cohort 2 ($p < .001$).

Of the 28 schools that participated in at least 1 year of the study, roughly one third were at the elementary level and an additional third span kindergarten through eighth grade. Fourteen percent and 18% of schools were at the middle and high school levels, respectively. All schools served student populations that were more than 90% African American; in all but one, more than 90% of students were eligible for free or reduced-price lunch. School rankings on a state “performance index” ranged from 27.3 to 112.9 with an average of 75.8, slightly higher than the Recovery School District average of 74, but notably lower than the state average of 99. We find no statistically significant differences in school characteristics across cohorts given that 17 schools participated in the program in both years ($F = 0.16$, $p = .988$).

Data and Measures

We utilize three primary sources of data to triangulate the effect of MTC on teachers’ practices: a classroom observation protocol developed by MTC and aligned to the coaching program, a principal evaluation derived from previous studies, and the TRIPOD student survey. We focus specifically on process measures and subjective ratings rather than on student achievement given both substantive and practical concerns about using test-score outcomes. First, these process measures align with our primary focus of changing teachers’ practices across grades and subjects. Given this goal, fewer than half of the teachers in our study taught in tested grades and subjects. Within randomization blocks, there was no guarantee that

TABLE 1

Teacher Characteristics and Balance Between Treatment Groups and Cohorts

		Cohort 1				Cohort 2				
		M				*M*				
	All teachers	All teachers	Treatment teachers	Control teachers	*p* value on difference	All teachers	Treatment teachers	Control teachers	*p* value on difference	*p* value on difference between cohorts
Teacher background characteristics										
Interest in coaching (scale from 1 to 10)	9.1	9.1	9.2	9.0	.319	9.1	9.1	9.1	.829	.929
Female (%)	71.2	74.6	70.0	0.8	.267	69.1	69.4	68.9	.820	.474
African American (%)	17.6	16.9	20.0	0.1	.858	18.1	14.3	22.2	.244	.859
White (%)	75.8	76.3	76.7	0.8	.564	75.5	75.5	75.6	.984	.918
Age (years)	25.6	26.1	26.1	26.1	.957	25.3	24.9	25.7	.308	.175
Teaching experience (years)	3.2	4.0	3.9	4.0	.893	2.6	2.6	2.7	.674	.001
First- or second-year teacher (%)	49.0	27.1	26.7	27.6	.866	62.8	65.3	60.0	.537	<.001
Third- or fourth-year teacher (%)	31.4	42.4	53.3	31.0	.124	24.5	22.4	26.7	.803	.020
Fifth- or higher year teacher (%)	19.6	30.5	20.0	41.4	.107	12.8	12.2	13.3	.513	.007
Alternatively certified (%)	79.1	76.3	80.0	72.4	.623	80.9	81.6	80.0	.661	.501
Master’s degree (%)	23.5	22.0	20.0	24.1	.755	24.5	24.5	24.4	.941	.732
College ranked very competitive or higher (%)	77.8	76.3	73.3	79.3	.561	78.7	83.7	73.3	.353	.725
Teaching and school characteristics										
Teach all subjects (%)	43.8	42.4	43.3	41.4	.826	44.7	42.9	46.7	.794	.781
Teach humanities (%)	33.3	35.6	36.7	34.5	.923	31.9	36.7	26.7	.242	.641
Teach STEM (%)	22.9	22.0	20.0	24.1	.794	23.4	20.4	26.7	.346	.846
F statistic from joint test					0.460				0.620	2.060
p value					.924				.820	.024
n (teachers)	153	59	30	29		94	49	45		

Note. Treatment- and control-group means are estimated from regression models that control for fixed effects for randomization blocks. Joint tests include teachers’ experience coded as a continuous variable and not the three individual dummies. STEM = science, technology, engineering, and mathematics.

TABLE 2

School Characteristics and Balance Between Cohorts

	All schools	Cohort 1	Cohort 2	*p* value on difference
Elementary schools (%)	32.1	30.0	36.0	.680
K–8 schools (%)	32.1	40.0	32.0	.588
Middle schools (%)	14.3	15.0	16.0	.929
High schools (%)	17.9	10.0	16.0	.567
Enrollment	491.7	532.7	473.9	.306
African American (%)	94.0	92.7	96.6	.252
Free or reduced-price lunch eligible (%)	92.5	91.8	94.5	.305
English as a second language (%)	0.8	1.0	0.9	.822
Special education (%)	15.6	16.5	16.2	.941
Student-to-teacher ratio	15.1	15.4	15.1	.676
Louisiana school performance score	75.8	75.2	77.4	.693
F statistic from joint test				.160
p value				.988
n (schools)	28	20	25	

Note. One school in Cohort 1 spans all grades, K–12, and therefore is excluded from individual school-level categories.

both treatment and control teachers had test-score data, further reducing the effective sample with test-score data. Second, examining effects on student achievement would require us to combine measures of student performance across grades and subjects despite the fact that these tests are not equatable and measure vastly different skills. Third, process measures allow us to triangulate the effect of coaching on a range of teacher practices rather than focusing on a narrower measure of teacher effectiveness based solely on test scores. Fourth, observation and survey measures are policy relevant as they are the primary evaluation measures available for the majority of classroom teachers. Finally, inconsistencies in district-level data, driven in large part by the high mobility of teachers and students across classrooms and schools in the New Orleans charter sector, limit the reliability of test-score data link to teachers.

MATCH Classroom Observation Rubric. As described in prior work (see Kraft & Blazar, 2014), 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 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 following 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 create teacher scores for each code by averaging raw scores across our two raters and then standardizing average scores in each year to be mean 0 and *SD* 1. Of our three sources of outcome data, the MATCH rubric is most aligned to treatment; therefore, if program effects exist, we anticipate finding the largest effects on this measure.

Principal Survey. We utilize 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 (.32 and .29, respectively, for the former survey, and .28 and .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 and at the end of the following academic year. For those 17 schools that participated in the program in both cohorts, all but 1 had the same principal across school years. We create a composite score of teachers' overall effectiveness, *Overall Effectiveness Composite*, by standardizing individual items within each year, averaging scores across all 11 items above, and then restandardizing this composite score to be mean 0 and *SD* 1. We estimate an internal consistency reliability of .91 or greater in all administrations. It is important to note that 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 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 .22 and .14 in math and reading (Kane & Staiger, 2012). We also examine the proportion of students who agreed with a single item, "In this class, we learn a lot every day." We present exploratory analyses of the effect of coaching on the other five TRIPOD domains, *Care, Captivate, Clarify, Confer*, and *Consolidate*, in an online appendix (available at http://epa.sagepub.com/supplemental). Upper elementary and secondary students rated each item on a 5-point Likert-type scale, whereas early elementary students had three response choices: no, maybe, and yes. Students completed the survey once at the end of the coaching year. Following the practices of the TRIPOD project, we derive 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 restandardize average score for each domain to be mean 0 and *SD* 1.

Summary Index. In an effort to guard against false positives and facilitate a parsimonious discussion of our results, we create a summary index of these three measures. We create 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, we give all three data sources equal weight. We then standardize the index to be mean 0 and *SD* 1. We also test the robustness of our findings to alternative composites that give more weight to the principal and student surveys, which are less proximal to the coaching program than the MATCH rubric.

Data Analyses

We estimate the effect of MTC on our outcomes of interest using ordinary least squares (OLS) and multilevel regression. We analyze 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 in school s at time t:

$$Y_{jst} = Y_{j,t-1} + \beta \text{MTC}_j + \alpha_{s,t-1} + \varepsilon_{jst}. \qquad (1)$$

For each of our teacher-level outcomes, we are able to include a baseline measure, $Y_{j,t-1}$, to increase the precision of our estimates. For the *Summary Index*, we calculate 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 include fixed effects for our randomization blocks, $\alpha_{s,t-1}$; 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 omit random effects for the schools where teachers worked during the coaching year because they are highly collinear with our blocking indicators. However, we cluster our standard errors at the school level in the current year. We also test the robustness of our results to model specifications that replace randomization blocks with school-by-cohort fixed effects.

We analyze our student-level survey outcomes by fitting an analogous multilevel model, where students, *i*, are nested within teachers, *j*, classrooms, *c*, and schools, *s*:

$$A_{ijcs} = \beta \mathrm{MTC}_j + \alpha_{s,t-1} + \left(\nu_j + \varphi_c + \varepsilon_{ijcs}\right). \qquad (2)$$

As noted above, we do not include a baseline measure, as the student survey was administered only once at the end of the year. We include random effects for teachers, ν_j, and classrooms, φ_c. We again cluster our standard errors at the school level in the current year.

In both models, the coefficients β on the indictor 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 only 10 treatment teachers dropped coaching (2 from Cohort 1 and 8 from Cohort 2). Of these, 5 were censored from our data because they either left teaching or withdrew from 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, as well as analyses presented below 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, or 1.14 (79 divided by 69).

Findings

Pooled Treatment Effects

Prior to presenting treatment effects, we confirm the validity of our randomization process by comparing the demographic characteristics of teachers assigned to treatment and control groups. The results reported in Table 1 provide strong evidence that randomization processes in both cohorts were implemented with fidelity. Differences in mean values of observable teacher characteristics across treatment and control groups within cohorts are small and insignificant for each measure; a joint test of significance fails to reject the null hypothesis that these characteristics are the same between treatment and control groups (pooling across cohorts, $F = 0.77$, $p = .704$, not shown in Table 1; Cohort 1: $F = 0.46$, $p = .924$; Cohort 2: $F = 0.62$, $p = .820$).

In Table 3, we present results pooling data across cohorts. Here, we find no statistically significant effects of coaching on any of our outcome measures, including the *Summary Index* of teacher effectiveness consisting of observation scores, principal evaluations, and student surveys. These findings remain consistent when we recalculate the *Summary Index* such that the principal and student surveys are given more weight than the MATCH rubric (see Table A1 in online appendix).

At the same time, these pooled analyses fail to examine the consequences of the major changes in program design and delivery that MTC underwent from Cohort 1 to Cohort 2. In the following section, we disaggregate our pooled results by cohort to examine variation in treatment effects.

Differential Treatment Effects Across Cohorts

In Table 4, we present treatment-by-cohort estimates for each of our outcome measures. To do so, we replace our main treatment indicator in Equations 1 and 2 with two cohort-specific treatment indicators, $\mathrm{MTC}_j \times COHORT\_1_j$ and $\mathrm{MTC}_j \times COHORT\_2_j$, where $COHORT\_1_j$ and $COHORT\_2_j$ each indicate the cohort that a given teacher participated in the study. Therefore, the interaction between these variables and the treatment indicator, MTC_j, identifies the effect of treatment within each cohort.

For Cohort 1, we find that MTC improved teachers' effectiveness across a range of practices. Coached teachers scored 0.56 *SD* higher than control-group teachers ($p = .023$) on our *Summary Index*. Specifically, trained classroom observers

TABLE 3

Parameter Estimates of the Effect of MATCH Teacher Coaching on Teachers' Practices

		MATCH rubric		Principal survey	TRIPOD student survey		
	Summary index	Achievement of lesson aim	Behavioral climate	Overall effectiveness composite	Challenge	Control	Learn a lot
Treat	0.115	0.096	0.263	−0.050	0.068	−0.002	0.018
	(0.185)	(0.202)	(0.178)	(0.188)	(0.086)	(0.101)	(0.028)
n (teachers)	135	134	134	132	115	115	115
n (students)	—	—	—	—	3,404	3,399	3,334

Note. Estimates in each column are from separate regression models. Standard errors clustered by school year in parentheses. All regressions include fixed effects for randomization blocks. The summary index includes the five main outcome variables: the two observation items, the principal evaluation, and the two student survey domains.
$^{\dagger}p < .1$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

TABLE 4

Parameter Estimates of the Effect of MATCH Teacher Coaching on Teachers' Practices Disaggregated by Cohort

		MATCH rubric		Principal survey	TRIPOD student survey		
	Summary index	Achievement of lesson aim	Behavioral climate	Overall effectiveness composite	Challenge	Control	Learn a lot
Treat × Cohort 1	0.564*	0.577†	0.663*	0.228	0.302**	0.093	0.080*
	(0.239)	(0.322)	(0.318)	(0.157)	(0.113)	(0.166)	(0.034)
Treat × Cohort 2	−0.173	−0.216	0.005	−0.234	−0.130	−0.084	−0.034
	(0.236)	(0.243)	(0.192)	(0.268)	(0.094)	(0.115)	(0.036)
Test between cohort coefficients							
F or χ^2 statistic	4.724	3.985	3.149	2.517	8.627	0.770	5.330
p value	.035	.052	.083	.120	.003	.380	.021
n (Teachers Cohort 1)	52	52	52	52	50	50	50
n (Students Cohort 1)	—	—	—	—	1,451	1,449	1,414
n (Teachers Cohort 2)	83	82	82	80	65	65	65
n (Students Cohort 2)	—	—	—	—	1,953	1,950	1,920

Note. Standard errors clustered by school year in parentheses. See Table 3 for further details.
$^{\dagger}p < .1$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

rated coached teachers 0.58 *SD* (*p* = .080) and 0.66 *SD* (*p* = .043) higher on *Achievement of Lesson Aim* and *Behavioral Climate*, respectively. Principals rated teachers who received coaching 0.23 *SD* (*p* = .153) higher on the *Overall Effectiveness Composite*. Students rated teachers who received coaching 0.30 *SD* (*p* = .008) higher on the *Challenge* composite. Finally, we find that MTC increased the probability that students felt that they learned a lot in class every day by 8 percentage points (*p* = .017; see Kraft & Blazar, 2014, for further details). Exploratory analyses also indicate positive effects on all other domains of teaching practice evaluated by students on the TRIPOD survey (see Table A2 in online appendix). This suggests that the positive effects of coaching extend beyond the specific classroom practices that coaches targeted. For Cohort 2, we find no statistically significant effects of coaching on any of our outcome measures. Except for *Behavioral Climate*, magnitudes of coefficients are negative, suggesting that these null effects are unlikely to be due to issues related to statistical power. We can detect statistically significant differences between these treatment-by-cohort coefficients for the *Summary Index* (F = 4.72, p = .035, including when we reweight this measure; see Table A1 in online appendix), as well as *Challenge* and *Learn a Lot* at the .05 level and *Achievement of Lesson Aim* and *Behavioral Climate* at the .10 level. We also detect statistically significant differences between cohorts, either at the .05 or .10 level, on all additional TRIPOD domains (see Table A2 in online appendix).

Robustness Checks

Next, we provide evidence that these estimates are robust to model specification and possible threats to internal validity due to missing data. In Table 5, we only present estimates of the effect of coaching on the *Summary Index* to facilitate a more parsimonious discussion of our findings given that trends are similar across other outcome measures (results available upon request). One concern may be that we do not account fully for contextual factors within individual schools in instances where randomization blocks include multiple school sites. Although this is true only for a small subset of blocks (3 of 40), in column 1 of Table 5 we find that trends in estimates and statistical significance of cross-cohort differences are preserved when we replace randomization block indicators with school-by-cohort fixed effects. We adopt these models rather than controlling for school characteristics given the limited variation in school demographics across our sample. A related concern may be that we do not control appropriately for teacher characteristics, which could drive our results. Although we are cautious about oversaturating our model by controlling for all possible background characteristics, we do find that results are robust to inclusion of a select set of teacher covariates, including interest in coaching, teaching experience, pathway to teaching certification, competitiveness of undergraduate institution, and whether or not the teacher earned a graduate degree (see Table 5, column 2).

Finally, we examine the robustness of findings to missing data due to attrition and incomplete data collection.[1] First, we look for differential attrition between the treatment and control groups. Of the 153 total teachers, 22 dropped from the study—including 5 control-group teachers and 2 treatment teachers from Cohort 1, and 7 control-group teachers and 8 treatment teachers from Cohort 2. Over half of these teachers dropped because they left teaching (see Table 6 for all reasons for dropping), which is reflective of the 27% annual turnover rate among teachers across the Recovery School District in the 2011–2012 school year (Cowen Institute, 2012). However, we do not find differential attrition between treatment and control groups when pooling across cohorts (p = .534) or when testing within each cohort (Cohort 1: p = .220; Cohort 2: p = .920). Furthermore, when we account for missing data by multiply imputing baseline and outcome measures using teacher characteristics and an indicator for treatment status (see Rubin, 1987), we find that results are unchanged (see Table 5, column 3). We interpret the findings of these robustness checks as clear evidence that results are not driven by model specification or missing data.

Explanations for Differences

The stark differences in treatment effects across cohorts could be due to two broad reasons. One explanation could be that elements of the research design in the second cohort may have masked the true effectiveness of the MTC program. For example, in Cohort 2, there may have been spillover effects or other changes in the counterfactual that would attenuate findings. It may also be the case that changes in the samples of participating teachers across cohorts led to differences in treatment effects; this would be true if MTC is differentially effective for specific groups of teachers. Finally, there may be outliers in one cohort or another that drive results.

Alternatively, treatment effect differences may reflect substantive changes in the coaching model from Cohort 1 to Cohort 2. Changes in teacher-to-coach ratios, the number of weeks of coaching each teacher received, coach personnel,

TABLE 5

Robustness Tests of the Effect of MATCH Teacher Coaching on a Summary Index of Teachers' Practices

	School-by-cohort fixed effects	Teacher controls	Multiple imputation
Panel A: Pooled estimates			
Treat	0.090	0.100	0.158
	(0.192)	(0.186)	(0.182)
Panel B: Disaggregated estimates			
Treat × Cohort 1	0.478†	0.608**	0.600*
	(0.250)	(0.224)	(0.265)
Treat × Cohort 2	−0.151	−0.212	−0.131
	(0.248)	(0.208)	(0.217)
Test between cohort coefficients			
F statistic	3.190	8.150	4.412
p value	.081	.007	.044
n (Teachers Cohort 1)	52	52	59
n (Teachers Cohort 2)	83	83	94

Note. Estimates in each panel and column are from separate regression models. Standard errors clustered by school year in parentheses. Teacher controls include interest in coaching; experience dummies (first- or second-year teaching, third- or fourth-year teaching); and indicators for earning a graduate degree, alternative certification, and for attending an undergraduate school with Barron's ranking of "very competitive" or higher. Imputation analyses account for missing data due to teachers who dropped from the study or student surveys that were lost in the mail. We use all available teacher characteristics and an indicator for treatment status to impute missing values across 10 replication data sets.

†$p < .1$. *$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 6

Number of Teachers Who Dropped from the Study for Different Reasons, by Cohort and Treatment Group

	Cohort 1 (*n* = 7)		Cohort 2 (*n* = 15)	
	Treatment	Control	Treatment	Control
Left teaching	1	4	2	5
Personal reason (e.g., health, lack of time)	1		1	
Wanted coaching		1		1
Did not want coaching			3	
Did not want to participate in data collection			2	1

and the focus of coaching might have led to an intervention that was less impactful at changing teachers' practices. We explore both explanations below.

Research Design Explanations

Differences in the Treatment–Control Contrast. One possible explanation for differential treatment effects related to the research design may be that the treatment remained constant across cohorts but the counterfactual experiences of control-group teachers changed across years. This might be true if there were general improvements in professional development programming provided to teachers across cohorts, or if spillover effects meant that control-group teachers in Cohort 2 had access to strategies utilized in the MTC program that those in Cohort 1 did not. Both would result in the same reduced treatment–control contrast.

One strategy that others have used to examine this form of bias is to compare the two control groups on baseline measures or on baseline-to-spring gain scores (e.g., Angrist, Pathak, & Walters, 2013). Differences between the two groups might suggest that one control group started at a higher level than the other or that one control group made larger gains over the course of a year than the other. Comparing control-group teachers' baseline scores across cohorts, we find no difference on either MTC observation rubric dimension. Specifically, control-group teachers in Cohort 1 scored 4.88 and 4.48 on *Achievement of Lesson Aim* and *Behavioral Climate*, respectively, at baseline, compared with 5.02 and 5.25 for those teachers in Cohort 2 (p = .771 and p = .123 for the two dimensions, respectively). This indicates that both sets of control-group teachers—which are mutually exclusive—were roughly equivalent with regard to two dimensions of instructional practice specifically targeted by the MTC program. Although we do observe that Cohort 1 control-group teachers scored 0.81 raw points higher than those in Cohort 2 on the principal survey *Overall Effectiveness Composite* (6.63 compared with 5.82, p = .004), we note that the magnitude is small and oppositely signed from differences at baseline on MTC dimensions described above and, therefore, is unlikely to explain differential treatment effects described above. We do not compare baseline-to-spring gain scores given some evidence that gain scores are not comparable across cohorts due to different sets of raters.[2]

We also examine potential differences in the treatment–control contrast through analyses of possible spillover effects. During the course of the study, several treatment teachers reported that their administrators adopted strategies taught by MTC as part of their school-wide professional development training, which could have reduced the treatment–control contrast in Cohort 2. However, using two sets of data, we argue that spillover is unlikely to drive our null findings from Cohort 2. In an end-of-year survey, we asked control-group teachers whether they were exposed to strategies taught by MTC during the coaching year. In Cohort 1, 2 teachers out of 25 control-group respondents (8%) indicated that they learned about strategies discussed in coaching but did not use them in their classes; another 8 (32%) indicated that they did utilize these strategies. In Cohort 2, a similar percentage of teachers indicated learning about but not using these strategies (9%), but a smaller percentage (23%) indicated utilizing these strategies in their classes. This suggests that effects of spillover may have been smaller in Cohort 2 than in Cohort 1, which would not explain the differential treatment effects described above. In addition, when we control for indicators for whether or not control-group teachers learned about or utilized coaching strategies in their classrooms, results remain unchanged (see Table A3 in online appendix). Specifically, when we pool data across cohorts, we find a null effect on the *Summary Index*; disaggregating by cohort, we find a statistically significant effect at the .05 level on the same outcome of 0.70 *SD* in Cohort 1 and a null effect in Cohort 2.

A second way that we explore spillover effects is by examining the lasting effects of coaching for Cohort 1 in the follow-up year. If control-group teachers from Cohort 1 had access to MTC strategies in the follow-up year, then we would expect the Cohort 1 treatment effect to be attenuated or to disappear. This is because the control-group teachers would benefit more from the improved professional development in the follow-up year than would treatment teachers who already had access to the MTC program and who no longer were receiving coaching. However, we find that this is not the case. Of the 59 teachers who participated in Cohort 1, we were able to collect an additional year of data on 33 teachers (21 from the treatment group and 12 teachers from the treatment group) who were still working as classroom teachers in New Orleans and who agreed to continue their participation in the study. Even though treatment and control teachers participated in the follow-up year at different rates, we find that these teachers do not differ on observable characteristics included in Table 1 from those who did not participate (p = .840; see Table A4 in online appendix). Data collection for the follow-up year coincided with the first year of coaching for teachers in Cohort 2. All but two of these Cohort 1 control-group teachers worked in the same schools as Cohort 2 teachers. Drawing on these data, we find that the magnitude of treatment effects is preserved in the follow-up year (effect size of 0.48 *SD* on the *Summary Index*), though imprecisely estimated due to the smaller sample of participating teachers (see Table A5 in online appendix). Consistent findings for Cohort 1 at the end of the coaching and follow-up years suggest that spillover did not drive attenuated results in Cohort 2.

Differences in Participating Teachers. Above, we present evidence that results are robust to inclusion of teacher control variables. However, as this model includes cohort-specific randomization blocks, teachers are compared within cohorts; it is possible that treatment effects are moderated by teacher characteristics that vary across cohorts. In the last column of Table 1, we compare observable characteristics of teachers across cohorts and find that teachers are similar on almost all characteristics, including their initial interest in coaching, gender, race, pathway into teaching, and level of education. The only observable characteristic for which we observe a difference between cohorts is years of teaching experience. Specifically, teachers in Cohort 1 have more teaching experience than those in Cohort 2, with 27% of teachers in Cohort 1 in their first or second year of teaching, compared with 63% of teachers in Cohort 2 ($p < .001$). This difference could account for the differential treatment effects across cohorts if more experienced teachers benefit more from coaching than less experienced teachers.

In Table 7, we explore this possibility by disaggregating results by cohort and experience level. To do so, we replace our treatment-by-cohort indicators in Equations 1 and 2 with treatment-by-cohort-by-experience dummies (first- or second-year teacher, third- or fourth-year teacher, or fifth- or higher year teacher). If differences in experience were driving the results, then we would expect to see larger effect sizes for teachers with more experience in both cohorts; in addition, teachers with fewer years of experience in both cohorts would exhibit smaller effect sizes. However, this is not the case. Treatment effect estimates across all three experience bins are generally positive for Cohort 1 and generally negative for Cohort 2. Furthermore, we observe differential treatment effects across cohorts, even for teachers in the same experience level. For example, on the *Summary Index*, teachers with five or more years of teaching experience from Cohort 1 have a treatment effect of 0.69 *SD*, whereas teachers with the same experience level in Cohort 2 have a treatment effect of −0.46 *SD*; the difference between these estimates is statistically significant (p = .035). We can also detect a marginally significant difference in treatment effects on the *Summary Index* across cohorts for third- and fourth-year teachers (p = .068). This suggests that null effects in Cohort 2 are unlikely to be explained solely by the increased proportion of teachers who are less experienced.

School Outliers. Finally, we examine whether there are outlier randomization blocks that drive treatment effects in one cohort or another. For example, changes in leadership or school culture across school years may have impacted the success of the program. In Figure 1, we plot treatment effect estimates on the *Summary Index* by randomization block and cohort. Many school-level blocks are in both cohorts; those that are only in one cohort, or those blocks that include teachers from multiple schools, automatically lie on the *x*- or *y*-axis. Markers are labeled with the number of teachers in each block. We identify as outliers those blocks with treatment effect estimates ±2 *SD* or higher, which are marked with a circle.

Across cohorts, there is some variation in treatment effect estimates; however, visual inspection suggests that there does not appear to be any clear outlier that would drive the results. In Cohort 1, most treatment effect estimates are positive. Of the 13 total blocks, all fall within 2 *SD*. In Cohort 2, most treatment effects are clustered between 0 and −1 *SD*, which is also close to our overall estimate that is negative in magnitude but indistinguishable from 0. We do identify three blocks whose treatment effects fall beyond ±2 *SD*. However, two are negative and one is positive in magnitude. Therefore, as expected, when we exclude these three blocks from our primary analyses, treatment effect estimates are slightly smaller, but general patterns are unchanged. On the *Summary Index*, we estimate a treatment effect for Cohort 1 of 0.39 *SD* (p = .133) compared with 0.56 *SD* with the full sample, and an effect for Cohort 2 of −0.09 *SD* (p = .718) compared with −0.17 *SD*. In addition, when we limit our analyses just to those schools that are in the sample for both cohorts, reestimated results are even closer to our original findings: 0.51 *SD* (p = .047) and −.13 *SD* (p = .650) for Cohorts 1 and 2, respectively. Together with the data examining potential spillover effects and differential treatment effects for more-experienced teachers, these findings indicate that differences in the research design are unlikely to account for large differences in treatment effects.

TABLE 7

Parameter Estimates of the Effect of MATCH Teacher Coaching on Teachers' Practices Disaggregated by Teaching Experience

		MATCH rubric		Principal survey	TRIPOD student survey		
	Summary index	Achievement of lesson aim	Behavioral climate	Overall effectiveness composite	Challenge	Control	Learn a lot
Treat × Exp	0.327	0.403	0.251	0.045	0.081	−0.179	−0.002
1–2 × Cohort 1	(0.444)	(0.355)	(0.345)	(0.388)	(0.166)	(0.233)	(0.036)
Treat × Exp	−0.133	−0.208	−0.036	−0.139	−0.145	−0.075	−0.044
1–2 × Cohort 2	(0.323)	(0.298)	(0.263)	(0.350)	(0.147)	(0.155)	(0.055)
Treat × Exp	0.634†	0.679	0.827†	0.349	0.378***	−0.008	0.090†
3–4 × Cohort 1	(0.323)	(0.445)	(0.447)	(0.225)	(0.102)	(0.144)	(0.049)
Treat × Exp	−0.227	−0.255	0.133	−0.377	−0.084	−0.101	−0.029
3–4 × Cohort 2	(0.327)	(0.430)	(0.359)	(0.322)	(0.170)	(0.204)	(0.042)
Treat × Exp	0.692*	0.569†	0.824**	0.199	0.396*	0.597***	0.147***
5 plus × Cohort 1	(0.309)	(0.312)	(0.260)	(0.493)	(0.178)	(0.160)	(0.044)
Treat × Exp	−0.462	−0.189	−0.076	−0.648	−0.226*	−0.067	0.060*
5 plus × Cohort 2	(0.426)	(0.549)	(0.598)	(0.702)	(0.109)	(0.178)	(0.030)
p values for differences between coefficients							
Treat × Exp 1–2 × Cohort 1 = Treat × Exp 1–2 × Cohort 2	.421	.202	.514	.710	.308	.708	.530
Treat × Exp 3–4 × Cohort 1 = Treat × Exp 3–4 × Cohort 2	.068	.133	.237	.071	.020	.711	.068
Treat × Exp 5 plus × Cohort 1 = Treat × Exp 5 plus × Cohort 2	.035	.239	.188	.334	.003	.005	.106
n (teachers)	135	134	134	132	115	115	115
n (students)	—	—	—	—	3,404	3,399	3,334

Note. Standard errors clustered by school year in parentheses. See Table 3 for further details. Exp = experience.
†*p* < .1. **p* < .05. ***p* < .01. ****p* < .001.

Coaching Content and Delivery Explanations

Thus far, we have provided evidence that differential treatment effects between Cohorts 1 and 2 are unlikely to be accounted for by the research design. The alternative explanation is that these differences are due to substantive changes in the program model across years, including larger teacher-to-coach ratios, a decrease in the number of weeks of coaching, changes in coach personnel, and an increased focus on behavior management. Because all of these changes were made as part of the program design prior to the beginning of the second cohort, rather than a reaction to teachers recruited to participate in the second year, we argue that between-cohort differences reflect exogenous variation in program characteristics. Therefore, we attribute differential treatment effects to these changes and other possible unobserved differences in implementation between cohorts. Below, we draw on qualitative data to describe these changes and then explore

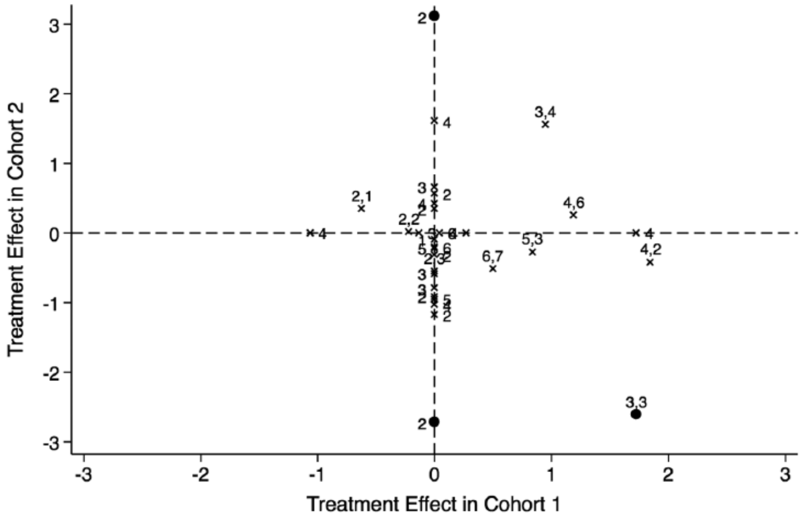

FIGURE 1. *Treatment effects on Summary Index, by randomization block and cohort, labeled with the number of teachers in each sample (Cohort 1, Cohort 2).*
Note. Blocks with treatment effects of ±2 *SD* or higher are marked with a circle.

quantitatively the extent to which specific changes in treatment might be related to program effectiveness.

Variation in Treatment Implementation. A rich set of qualitative data from coaching logs and discussions with program leaders allows us to describe variation in the program model across cohorts. One substantive change in Cohort 2 was the total number of coached teachers and, therefore, larger teacher-to-coach ratios. In Cohort 1, 30 teachers were assigned to receive coaching, of whom 28 took up the offer. Two coaches each worked with 10 teachers, and the third worked with 15. These numbers do not sum to 28, as some teachers worked with more than one coach. In Cohort 2, 49 teachers were assigned to coaching, of whom 45 took up the offer; 41 teachers completed the full year of coaching. Two coaches worked with 18 or 21 teachers, whereas the third coach worked with only 9 teachers to devote additional time to administrative and managerial duties as director of the program. As in Cohort 1, some teachers worked with more than one coach.

Another related change to the program model was the total number of weeks of coaching that teachers received. As planned, almost all teachers in Cohort 2 (88%) received 3 weeks of coaching, compared with almost all teachers in Cohort 1 (86%) who received 4 weeks of coaching. In a few instances, teachers received an additional week of coaching based on their coach's discretion and availability. This reduction in the dosage of coaching offset the need to increase coaches' weekly workload substantially. Coaches in Cohort 1 worked with between 1.5 and 1.9 teachers, on average, per week, compared with coaches in Cohort 2 who worked with between 2 and 2.2 teachers, on average, per week.

Across cohorts, there was also turnover in program personnel, with only one out of three coaches returning for the second year. All coaches were former teachers with professional experience in education nonprofits or charter school management organizations. They were also trained by the program director/lead coach using the same overarching model. As part of this process, coaches jointly observed classroom instruction and normed scores on the MATCH rubric. In addition, the program director shadowed the other coaches throughout the school year, providing direct feedback on how coaches interacted with teachers and observing how coaches implemented this feedback. Conversations with the program director indicated that training was more formalized in Cohort 2 and that feedback cycles were more frequent. In particular, additional trainings over the

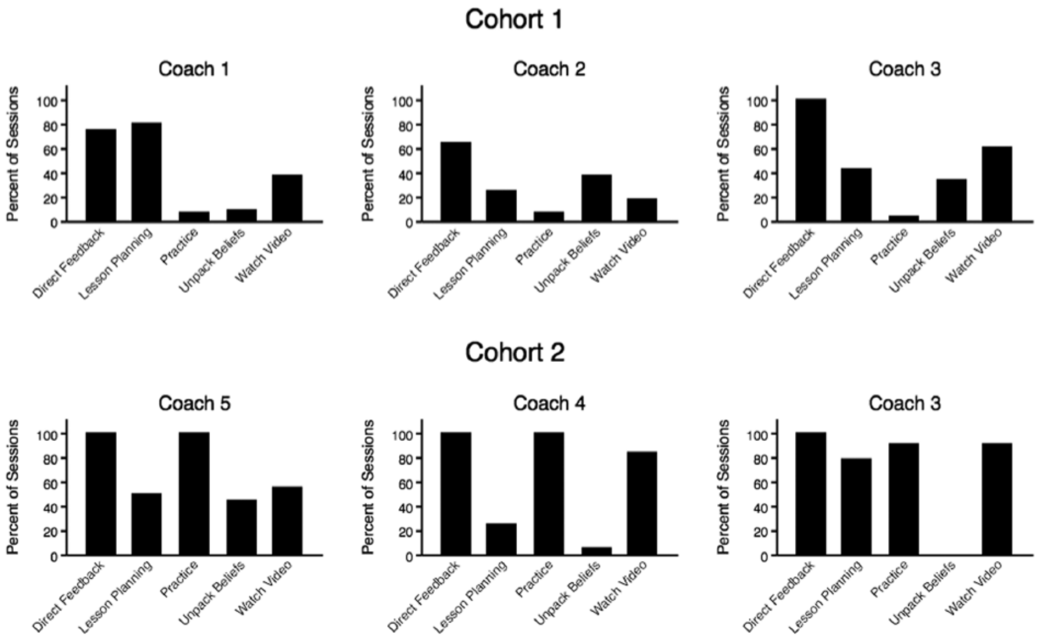

FIGURE 2. *Techniques used in debriefing sessions with teachers, by coach and cohort.*

summer set clear procedures for how coaches in Cohort 2 should debrief observations with teachers and write up action steps.

At the same time, given the individualized nature of the teacher–coach relationship, coaches had the freedom to utilize a variety of techniques while working and debriefing with teachers. Data from coach logs allow us to explore how frequently coaches used these techniques. In Figure 2, we show that, in both cohorts, coaches most often provided direct feedback to teachers about what they observed while watching a given lesson. In Cohort 2, coaches also relied heavily on having teachers practice a specific skill and watching video recordings of instruction. Unfortunately, our data do not allow us to capture other elements of the teacher–coach relationship, such as rapport, that likely play a role in individual coaches' success.

A final difference in implementation between the two cohorts was a programmatic change to increase the focus on behavior management in Cohort 2 over instructional delivery and student engagement. Specifically, in Cohort 2, coaching was formally organized such that coaches prioritized behavior management early in the coaching process and only moved on to other focus areas after teachers mastered this skill. This decision was made by the head of MATCH and the program director, who together felt that teachers' success in the classroom depended first on mastering behavior management. Coaches determined teachers' baseline mastery of behavior management through scores on the MTC rubric evaluated in the spring prior to receiving coaching. In contrast, in Cohort 1, there was no such formal programmatic approach to sequencing the topics of coaching. Instead, coaches made decisions about which areas to focus on based on their own interpretation of teachers' strengths and weaknesses and on conversations with these teachers.

We illustrate the ways that these decisions played out in practice in Figure 3, which shows a histogram of the percent of these weeklong sessions that teachers focused on each of the three focus areas of coaching, and Figure 4, which shows the percent of sessions that focused on one area versus another at different points over the course of the school year (four in Cohort 1 and three in Cohort 2). In line with our conversations with MTC leadership, in Cohort 1, teachers worked on each of the three areas of coaching to varying degrees; some teachers worked on a

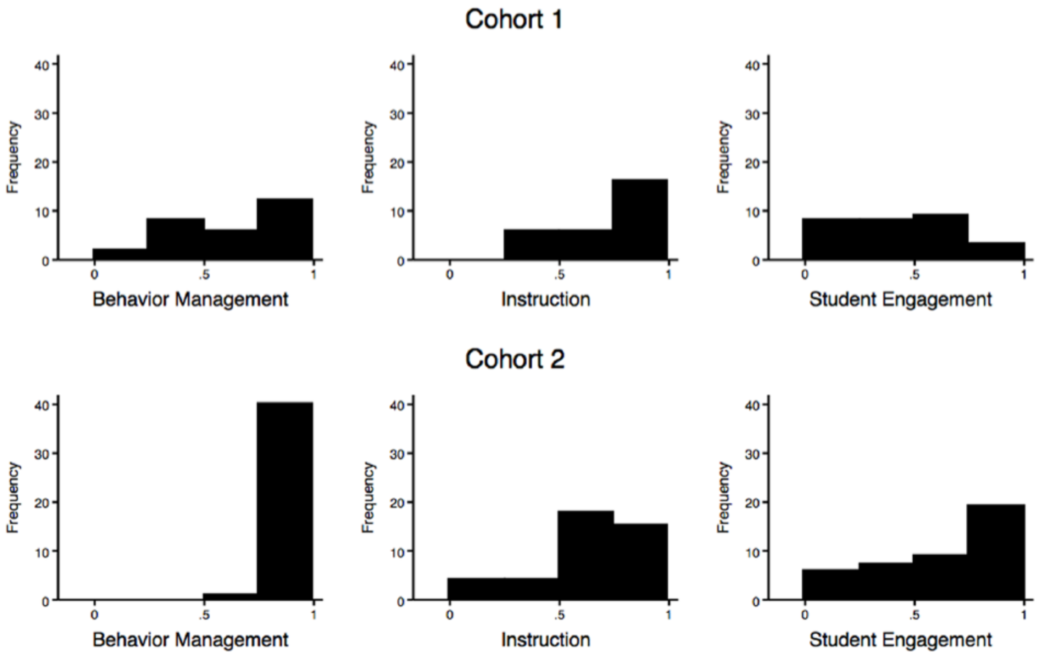

FIGURE 3. *Distributions of the percentage of sessions that each teacher worked on a given focus area, by cohort.*

given area during all of their sessions, whereas others did not work on a focus area at all (see Figure 3). Interestingly, we also observe that the focus of coaching shifted over the course of the year from behavior management to instruction (see Figure 4). During the first week of coaching, 62% of sessions covered behavior management, and during the last week of coaching, 59% covered instruction. While this progression was not built into the formal design of the coaching process, one reason for this may be that coaches wanted to work with teachers on a range of classroom practices during their time together.

In contrast, in Cohort 2, almost all teachers focused on behavior management in every session, with some teachers never working on instruction or student engagement (see Figure 3). This likely indicates that coaches felt that teachers had not mastered behavior management. Relatedly, we observe that this focus on behavior management was also largely maintained throughout the course of the coaching year (see Figure 4). The fact that we see a substantive percent of total sessions that focus on instruction or student engagement suggests that some teachers worked on behavior management and other practices at the same time.

Exploring Components of Effective Coaching. Our research design does not allow us to disentangle changes in teacher-to-coach ratios from the number of weeks of coaching, turnover in coach personnel, or changes in the focus of coaching. Instead, we conduct exploratory analyses to examine the relationship between some of these features of coaching and improvements in teachers' practices. Due to very limited variation in the total number of weeks of coaching within a given cohort, our data do not provide much evidence on the relationship between the number of weeks coached and outcomes. We also note that teacher-to-coach ratios are collinear with coaches in all but one instance; therefore, we discuss these two features together under a broad umbrella of coach effectiveness.

These exploratory analyses derive from slight modifications to the regression models described above. The revised teacher- and student-level models that describe the relationships between coaching characteristics and each of our outcome

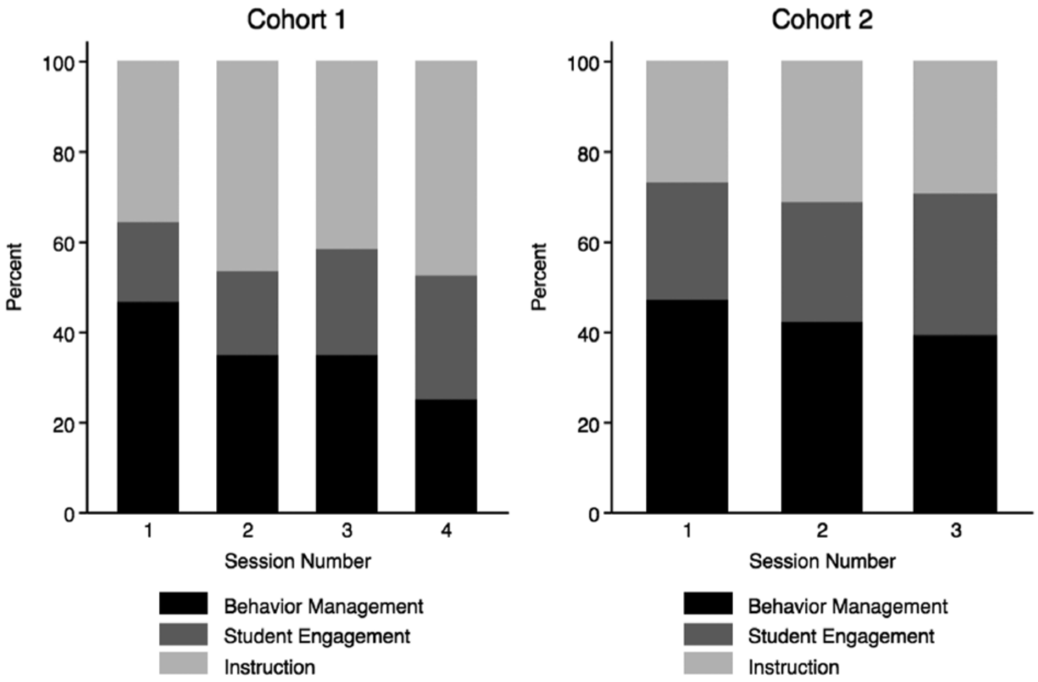

FIGURE 4. *Changes in the content of coaching across coaching sessions, by cohort.*

measures are given in Equations 3 and 4, respectively:

$$Y_{jht} = Y_{j,t-1} + \beta COACHING\_CHARACTERISTIC_j + \delta_h + \varepsilon_{jht}, \qquad (3)$$

$$A_{ijch} = \beta COACHING\_CHARACTERISTIC_j + \delta_h + \left(\nu_j + \varphi_c + \varepsilon_{ijch}\right). \qquad (4)$$

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 remove fixed effects for randomization blocks 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 add a cohort indicator, δ_h, to hold constant any difference in outcomes across years due to, for example, differences in classroom raters across years.

Coach effectiveness. In Table 8, we disaggregate treatment effects by coach and find that there are substantive and statistically significant differences between them. In Cohort 1, there were three coaches, numbered 1 through 3. Coach 3, the head coach, continued to work in Cohort 2, but Coaches 1 and 2 were replaced by Coaches 4 and 5. In a few instances, coached teachers worked with two different coaches over the course of the school year; therefore, we weight coach indicators in Equations 3 and 4 by the fraction of time a teacher spent with one coach versus another. Substantively, these estimates represent treatment effects attributable to each coach.

We observe statistically significant and positive coach effects on our *Summary Index* for all three coaches in Cohort 1, upward of 0.87 *SD*. For Cohort 2, we again observe some statistically significant and positive effects for working with the head coach (Coach 3) on the MATCH observation rubric; these estimates are largely

TABLE 8

Parameter Estimates of the Effect of MATCH Teacher Coaching on Teachers' Practices Disaggregated by Coach

	MATCH rubric			Principal survey	TRIPOD student survey		
	Summary index	Achievement of lesson aim	Behavioral climate	Overall effectiveness composite	Challenge	Control	Learn a lot
Coach 1 (Cohort 1)	0.599*	0.999**	1.140***	−0.267	0.418***	0.402†	0.104***
	(0.256)	(0.358)	(0.301)	(0.296)	(0.110)	(0.220)	(0.027)
Coach 2 (Cohort 1)	0.548*	0.536†	0.448†	0.392	0.045	−0.001	0.013
	(0.231)	(0.275)	(0.259)	(0.322)	(0.174)	(0.223)	(0.065)
Coach 3 (Cohort 1)	0.867*	0.614	0.735	0.703*	0.403***	0.047	0.093†
	(0.422)	(0.487)	(0.482)	(0.300)	(0.093)	(0.246)	(0.052)
Coach 3 (Cohort 2)	0.188	0.442†	0.668***	−0.013	−0.070	−0.086	0.061
	(0.227)	(0.220)	(0.186)	(0.405)	(0.121)	(0.124)	(0.042)
Coach 4 (Cohort 2)	−0.271	−0.408	−0.189	−0.205	−0.084	−0.232	−0.037
	(0.257)	(0.334)	(0.22)	(0.282)	(0.099)	(0.172)	(0.040)
Coach 5 (Cohort 2)	0.128	−0.046	0.220	0.172	0.059	−0.030	−0.007
	(0.267)	(0.239)	(0.235)	(0.331)	(0.130)	(0.127)	(0.050)
p values for differences between coefficients							
Coach 1 = Coach 2	.864	.127	.022	.212	.009	.130	.168
Coach 1 = Coach 3 (Cohort 1)	.472	.339	.268	.033	.895	.278	.814
Coach 1 = Coach 3 (Cohort 2)	.228	.192	.174	.620	.003	.053	.384
Coach 1 = Coach 4	.024	.007	.001	.880	.001	.023	.003
Coach 1 = Coach 5	.209	.020	.020	.334	.035	.088	.051
Coach 2 = Coach 3 (Cohort 1)	.425	.87	.544	.454	.021	.879	.295
Coach 2 = Coach 3 (Cohort 2)	.273	.789	.491	.411	.590	.739	.536
Coach 2 = Coach 4	.020	.031	.065	.160	.523	.412	.520
Coach 2 = Coach 5	.254	.114	.521	.617	.948	.911	.813
Coach 3 (Cohort 1) = Coach 3 (Cohort 2)	.161	.749	.896	.151	.002	.629	.631
Coach 3 (Cohort 1) = Coach 4	.027	.092	.092	.030	.000	.353	.047
Coach 3 (Cohort 1) = Coach 5	.149	.231	.341	.225	.032	.781	.167
Coach 3 (Cohort 2) = Coach 4	.101	.010	.002	.648	.916	.409	.046
Coach 3 (Cohort 2) = Coach 5	.836	.100	.069	.667	.345	.595	.223
Coach 4 = Coach 5	.19	.305	.145	.241	.179	.202	.537
n (Teachers)	135	134	134	132	115	115	115
n (Students)	—	—	—	—	3,404	3,399	3,334

Note. Standard errors clustered by school year in parentheses. Coach indicator variables weighted by the amount of time a teacher spent with one coach versus another. See Table 3 for further details.

$^{\dagger}p < .1$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

indistinguishable from those for this coach in Cohort 1. Conversely, we find no significant effects of being coached by Coach 4 or 5 in Cohort 2. Coefficients for Coach 4 are negative in magnitude for all outcome measures, though imprecisely estimated. For Coach 5, estimates for half of the outcome measures are negative in magnitude, and half are positive. Generally, these patterns suggest positive effects for coaches who worked in Cohort 1 and null effects for those coaches who were replaced in Cohort 2.

Because we cannot disentangle coach effects from other changes in the program model, it is possible that comparing coach effect estimates across cohorts may just be a proxy for other differences across cohorts. At the same time, we also observe some differences in coach effects within cohorts. For example, Coach 1 shows larger effects than Coach 2 on *Behavioral Climate* and *Challenge*; Coach 3 (Cohort 2) shows larger effects than Coaches 4 and 5 on both *Achievement of Lesson Aim* and *Behavioral Climate*.

In addition, coach effects do not appear to be explained fully by other programmatic changes. Differences in training of coaches between cohorts are unlikely to account for our findings given that training was more intensive in the second year. Furthermore, results are unlikely to be driven by differences in teacher-to-coach ratios across cohorts. Above, we note a modest increase in the number of total teachers and teachers per week with whom each coach worked. Intuitively, a heavier workload in Cohort 2 might have affected these coaches' (particularly Coaches 4 and 5) ability to individualize feedback to each teacher. Although we cannot rule out this as one contributing factor, the increases across cohorts in total teachers coached and teachers coached per week are not proportionate to the large differences in coach effects. In particular, we observe some of the largest coach effects for Coach 1 in Cohort 1, who worked with 15 total teachers and 1.9 teachers per week. This workload is not substantively different from those for coaches in Cohort 2 who had the smallest treatment effects. Coach 4 worked with 18 teachers and 2.2 teachers on average per week, and Coach 5 worked with 21 teachers and 2.1 teachers on average per week. This suggests that other differences between coaches likely play a key role.

Finally, we note that, although teachers were not randomly assigned to coaches, it is unlikely that teacher–coach matching would bias our estimates substantially. In both cohorts, coaches made decisions about who would work with each teacher based mostly on past teaching experience; that is, coaches who had experience in elementary school tended to coach teachers at this level and similarly for those with experience at other grade levels. Geographical proximity also played a role in some matches. Together, these findings point to evidence of coach effects as one likely explanation for the differential findings we observe across cohorts.

Focus of coaching. Next, we explore whether differences in treatment effects may be attributable to time spent on a given focus area—behavior management, instructional delivery, or student engagement. In the regression model that explores this relationship, we include variables that describe the number of weeks teachers spent on each focus area. Because teachers often worked on more than one focus area in a given week, variables are not mutually exclusive. As described above, we include baseline measures of our teacher-level outcomes; this is important, given that the number of weeks of coaching that teachers focused on a given instructional domain likely is related to their incoming level of quality in this area. For similar reasons, we control for the total number of weeks of coaching received.

Our findings suggest that an additional week spent on instructional delivery is consistently associated with positive and mostly statistically significant improvements in teachers' practices, including 0.39 *SD* on *Achievement of Lesson Aim* (p = .006; see Table 9). Conversely, time spent on behavior management is associated with negative and often statistically significant decrements in teachers' practices, including direct measures of a teacher's behavior management skills (−0.18 *SD* on *Behavioral Climate,* p = .052; −0.13 *SD* on *Control,* p = .014). These negative coefficients remain when we rerun models only with Cohort 1 (−0.26 *SD* on *Behavioral Climate,* p = .095; −0.15 *SD* on *Control,* p = .171, not shown in Table 9), indicating that results are not confounded with cohort. Finally, when we formally compare coefficients for time spent on behavior management versus time spent on

TABLE 9

Parameter Estimates of the Effect of MATCH Teacher Coaching on Teachers' Practices Disaggregated by the 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	−0.151	−0.240*	−0.182†	0.000	−0.021	−0.135*	−0.009
	(0.096)	(0.114)	(0.091)	(0.101)	(0.044)	(0.055)	(0.012)
Instructional delivery	0.318*	0.387**	0.475***	0.096	0.143**	0.122	0.028
	(0.130)	(0.134)	(0.118)	(0.127)	(0.050)	(0.080)	(0.019)
Student engagement	−0.052	−0.096	−0.074	0.001	−0.052	−0.032	0.000
	(0.095)	(0.119)	(0.090)	(0.085)	(0.039)	(0.038)	(0.015)
Number of weeks of coaching	0.022	0.057	−0.014	−0.025	−0.011	0.040	−0.001
	(0.106)	(0.118)	(0.098)	(0.106)	(0.052)	(0.069)	(0.020)
p values for differences between coefficients							
Behavior management = instruction	.002	.001	.001	.465	.018	.007	.056
Behavior management = student engagement	.536	.450	.458	.990	.623	.123	.657
Instruction = student engagement	.068	.026	.002	.615	.002	.114	.201
n (Teachers)	135	134	134	132	115	115	115
n (Students)	—	—	—	—	3,404	3,399	3,334

Note. Standard errors clustered by school year in parentheses. Focus area variables indicate the number of sessions that a teacher worked on a given area; these are always coded as 0 for control-group teachers. See Table 3 for further details.
$^{\dagger}p < .1$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

instructional delivery, we find that they are statistically significantly different from each other when predicting each of our outcome measures except for *Overall Effectiveness Composite*. This indicates that time spent on the latter dimension over the former may contribute to program effectiveness and the differential treatment effects we observe.

Given that the focus of coaching is endogenous, these results should be interpreted cautiously. That is, the program design suggests that teachers who spent more time on behavior management likely were those most in need of support. Indeed, this appears to be the case, with the number of weeks spent on behavior management negatively correlated with baseline *Behavioral Climate* score at −0.21 ($p = .068$). Therefore, we might expect to see a negative relationship between this variable and teachers' overall effectiveness. At the same time, it may be easiest to realize large initial gains in behavior management practices with teachers most in need of coaching for behavior management.

It is also noteworthy that the shift toward behavior management in Cohort 2 was not matched with targeted improvements in this specific area of teaching practice. In fact, we observe a statistically significant and negative relationship between time spent on behavior management and outcomes that directly measure management skills, *Behavioral Climate* and *Control*. It is surprising that, in a group of novice teachers where the focus of coaching was to improve their behavior management, neither outside observers nor students identified teachers as getting better at this skill. The fact that neither behavior management nor instructional delivery improved for teachers in Cohort 2—whereas in Cohort 1, teachers improved on both dimensions—could be a further indication of the importance of coach effects.

Conclusion

A variety of theoretical and empirical evidence points to teacher coaching as a high-quality professional development opportunity that can improve teachers' practices and student achievement (Allen et al., 2011; Campbell & Malkus, 2011; Desimone, 2009; Neuman & Cunningham, 2009; Powell et al., 2010; Russo, 2004; Sailors & Price, 2010). In this study, we find inconsistent effects on teachers' practices of a coaching program focused on behaviors common across grades and subjects. Our analyses suggest that differential effects across the two teacher cohorts we study are unlikely to be explained by idiosyncrasies in the research design. The large increase in the proportion of first- or second-year teachers in Cohort 2 may be a contributing factor but is unlikely to explain the large differences in treatment effects that we observe. Instead, our analyses suggest that these differences are due to one or more substantive changes in the coaching model—namely, teacher-to-coach ratios, the total number of weeks of coaching received, turnover of coaches, and shifts in the focus of coaching.

Our research design cannot disentangle which of these program changes, or which combination, led to differences in program effects across cohorts. At the same time, exploratory analyses provide some evidence that changes in the focus of coaching and, in particular, coach effectiveness across cohorts may have played a leading role. The fact that individual coaches likely differ in their effectiveness is not altogether surprising. A large body of research finds substantial variability in teacher effectiveness, as well as a fairly steep learning curve in the first few years on the job (Harris & Sass, 2011; Kane, Rockoff, & Staiger, 2008; Papay & Kraft, in press; Rockoff, 2004). At the same time, these results are particularly salient at a time when little is known about particular skill sets that translate into being a good coach and the necessary conditions that make for a positive teacher–coach relationship. In our study, the intensity of training does not seem to explain coach effectiveness. Atteberry and Bryk (2011) suggested that there is a threshold effect where coaches who work with more than 12 teachers provide weaker implementation due to higher demands. Coaches' overall and weekly workload may play a role in our findings. However, these factors do not appear to be the primary driver of coach effectiveness given vastly different treatment effects for coaches who worked with relatively similar numbers of teachers/teachers per week. Some current descriptive work explores the teacher–coach dynamics but still argues that we need to examine learning processes within this relationship (see, for example, Coburn & Woulfin, 2012). To confirm our findings and expand on them, future research may consider randomly assigning teachers to coaches. This would allow analysts to test for variation at the coach level, as well as to look at the characteristics of effective coaches and effective teacher–coach relationships.

More surprising are our results around the focus of coaching. Prior work exploring effective classroom practices provides suggestive evidence on the importance of behavior and classroom management, often above other classroom features (Grossman, Loeb, Cohen, & Wyckoff, 2013; Kane & Staiger, 2012; Stronge, Ward, & Grant, 2011). If anything, these findings lead us to suspect that focusing specifically on classroom management would yield a greater increase in perceived teacher quality. This is particularly true in the context of this study, which took place in charter schools that often employ "no-excuses" policies toward classroom behavior. If, in fact, coaching teachers on instructional delivery is more beneficial than coaching on behavior management, this would be an important finding. Because of the endogeneity of time spent on any given classroom practice in this study and the way that it is confounded with coach effectiveness, future research is needed to explore differential effects between coaching for instruction versus behavior management.

Finally, identifying cost-effective coaching designs will be an imperative for researchers as schools and districts look to invest in coaching as a key component of professional development efforts. MTC costs between US$5,500 and

US$9,000 per teacher, driven largely by personnel costs and teacher-to-coach ratios. Although one way to lower costs would be to reduce the time a coach spends with any individual teacher, we find that when MTC cut back the number of weeks of coaching received, the program was no longer effective. Although we are unable to disentangle this result from the other changes in the coaching model, these findings align with prior work emphasizing the importance of high dosages or high density in professional development programs (Garet et al., 2001; Ramey et al., 2011). Another component to consider for cost-effectiveness that we do not explore is the mode of coaching. Compared with in-person interactions, online coaching might enable coaches to reach a broader group of teachers and decrease commuting time. Current evidence indicates that Web-based coaching around teacher–student interactions can raise student achievement (Allen et al., 2011). But, it is not clear if this mode is equally effective, more effective, or less effective than in-person coaching.

The potential to improve the quality of the teacher workforce via teacher coaching will depend on the efforts of researchers and practitioners to identify the specific design features of effective coaching programs.

Acknowledgments

We thank Michael Goldstein, Orin Gutlerner, and the members of the MATCH Teacher Coaching staff for their tireless work in implementing the coaching program. We also thank Larry Katz and Michael Hansen for their feedback on initial drafts.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by funding from New Schools for New Orleans.

Notes

1. In Cohort 1, we are missing data for all 7 teachers who dropped from the study. In Cohort 2, we are missing data for 11 teachers, including 6 who left teaching and 5 who dropped participation. We are also missing some data for individual outcome measures due to maternity leaves at the end of the year, principals who did not complete the survey, and student surveys that were lost in the mail.

2. The two raters who observed Cohort 2 teachers at the end of the year provided scores on *Achievement of Lesson Aim* and *Behavioral Climate* that were 1.2 and 1.3 raw points higher, on average, than those from other sets of raters who observed Cohort 1 teachers both the prior and concurrent springs. One of the observers for Cohort 2 rated more than 25% of teachers in the top score point on both of the rubric items. This does not affect our treatment estimates for Cohort 2 as raters were fully crossed with treatment conditions. However, it could lead to artificial differences in gain scores across cohorts due to rater effects.

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.

Atteberry, A., & Bryk, A. S. (2011). Analyzing teacher participation in literacy coaching activities. *The Elementary School Journal*, *112*, 356–382.

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*, 430–454.

Coburn, C. E., & Woulfin, S. L. (2012). Reading coaches and the relationship between policy and practice. *Reading Research Quarterly*, *47*, 5–30.

Cowen Institute. (2012). *The state of public education in New Orleans: 2012 report*. New Orleans, LA: Tulane University.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession*. Washington, DC: National Staff Development Council.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*, 181–199.

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*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915–945.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., . . .Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., . . .Britton, E. (2008). *Impacts of comprehensive teacher induction: Results from the first year of a randomized controlled study*. Washington, DC: U.S. Department of Education.

Gregory, A., Allen, J. P., Mikami, A. Y., Hafen, C. A., & Pianta, R. C. (2014). Effects of a professional development program on behavioral engagement of students in middle and high school. *Psychology in the Schools*, *51*, 143–163.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*, 445–470.

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 http://www.urban.org/UploadedPDF/1001431-what-makes-for-a-good-teacher.pdf?RSSFeed=UI_EducationPolicyCenter.xml

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics*, *95*, 798–812.

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.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*, 615–631.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Measures of Effective Teaching Project, Bill & Melinda Gates Foundation.

Kaufman, D., Johnson, S. M., Kardos, S. M., Liu, E., & Peske, H. G. (2002). "Lost at sea": New teachers' experiences with curriculum and assessment. *Teachers College Record*, *104*, 273–300.

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

Kraft, M. A., & Blazar, D. (2014). *Improving teachers' practice across grades and subjects: Experimental evidence on individualized teacher coaching* (Working paper). Brown University. Retrieved from http://scholar.harvard.edu/files/mkraft/files/mtc_nola_sep2014.pdf

Lake, R., Bowen, M., Demeritt, A., McCullough, Haimson, J., & Gill, B. (2012). *Learning from charter school management organizations: Strategies for student behavior and teacher coaching*. New York, NY: Mathematica Policy Research.

Maier, A., Cellini, K., & Grogan, E. (2012, March). *Fast start: Jumpstarting early teacher career effectiveness through targeted training and coaching*. Paper presented at the Association for Education Finance and Policy annual conference, New Orleans, LA.

Marsh, J. A., McCombs, J. S., Lockwood, J. R., Martorell, F., Gershwin, D., Naftel, S., . . .Crego, A. (2008). *Supporting literacy across the Sunshine State: A study of Florida middle school reading coaches*. Santa Monica, CA: RAND.

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal*, *46*, 532–566.

Papay, J. P., & Kraft, M. A. (in press). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0047272715000304

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology*, *102*, 299–312.

Ramey, S. L., Crowell, N. A., Ramey, C. T., Grace, C., Timraz, N., & Davis, L. E. (2011). The dosage of professional development for early childhood professionals: How the amount and density of professional development may influence its effectiveness. *Advances in Early Education and Day Care*, *15*, 11–32.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*, 247–252.

Rubin, D. (1987). *Multiple imputation for nonresponsive in surveys*. New York, NY: Wiley.

Russo, A. (2004). School-based coaching. *Harvard Education Letter*, *20*(4), 1–4.

Sailors, M., & Price, L. R. (2010). Professional development that supports the teaching of cognitive reading strategy instruction. *The Elementary School Journal*, *110*, 301–322.

Sawchuk, S. (2009, September). Growth model. *Education Week*, *29*(3), 1–4.

Smith, S. (2013). *In one California school district, teachers help teachers get better* (The Hechinger Report). Available from http://hechingerreport.org

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*, 339–355.

Wildman, T. M., Magliaro, S. G., Niles, R. A., & Niles, J. A. (1992). Teacher mentoring: An analysis of roles, activities, and conditions. *Journal of Teacher Education*, *43*, 205–213.

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 a doctoral candidate in quantitative policy analysis in education at the Harvard Graduate School of Education. His research focuses on teacher and teaching quality, and the effects of policies aimed at improving both.

MATTHEW A. KRAFT is an assistant professor of education at Brown University. His research interests include the economics of education, education policy analysis, and applied quantitative methods for causal inference. His primary work focuses on policies to improve educator and organizational effectiveness in K–12 urban public schools.

Manuscript received September 3, 2014
Revision received December 17, 2014
Accepted January 26, 2015