Does Test Preparation Mean Low-Quality Instruction?

David Blazar[1] and Cynthia Pollard[2]

Critics of test-based accountability warn that test preparation has a negative influence on teachers' instruction due to a focus on procedural skills. Others advocate that the adoption of more rigorous assessments may be a way to incentivize more ambitious test preparation instruction. Drawing on classroom observations and teacher surveys, we do find that test preparation activities predict lower quality and less ambitious mathematics instruction in upper-elementary classrooms. However, the magnitudes of these relationships appear smaller than the prevailing narrative has warned. Further, our findings call into question the hypothesis that test rigor can serve as a lever to elevate test preparation to ambitious teaching. Therefore, improving the quality of mathematics instruction in the midst of high-stakes testing likely will require that policymakers and school leaders undertake comprehensive efforts that look beyond the tests themselves.

Keywords: accountability; ambitious instruction; econometric analysis; educational policy; elementary schools; high-stakes testing; instructional practices; instructional quality; mathematics education; regression analyses; survey research; test preparation

Increased test-based accountability over the past several decades has led to the growth and prominence of standardized testing in U.S. schools and subsequently to increased test preparation activities in classrooms (Au, 2007; Popham, 2001). Critics of test-based accountability have argued that test preparation detracts from students' classroom experiences by crowding out high-quality forms of instruction—often referred to as "inquiry-oriented," "ambitious," or "rich" instruction—in favor of routine practices aimed at boosting students' test scores (Amrein & Berliner, 2002; Diamond, 2007; Koretz, 2008). Some further hypothesize that low-quality test preparation instruction results from the low cognitive demand of many tests used for accountability purposes, which do not create incentives for teachers to engage students around ambitious instruction (Darling-Hammond & Adamson, 2014). Comparatively, "teaching to the test" could promote instructional quality if the test is aligned to rigorous content and teachers in turn align their instruction to these standards (Polikoff & Porter, 2014). We test these hypotheses by drawing on survey and video data to examine whether observed quality of instruction differs between teachers and classrooms with varying levels of engagement in test preparation. We also examine whether these relationships differ across districts where teachers prepare students to take high-stakes tests that vary considerably in their level of cognitive demand.

Like all other work on this topic, our data are descriptive in nature. However, we attempt to gain some insight into the underlying causal relationship between test preparation and instructional quality in two key ways. First, we condition our estimates on a rich set of characteristics that capture many of the factors most likely to incent teachers to engage in these activities. Second, for a subsample of teachers who contributed lessons that explicitly aimed at preparing students for state tests and ones that did not, we compare instructional quality scores between lessons from the same teacher. This approach helps avoid bias due to an individual teacher's underlying quality and his or her propensity to engage in different levels of test preparation. These analyses provide a comprehensive picture of the nature and quality of test preparation instruction in upper-elementary mathematics as well as potential mechanisms that may drive test preparation instruction to be less ambitious than typically desired in U.S. classrooms.

[1]University of Maryland College Park, College Park, MD
[2]Harvard Graduate School of Education, Cambridge, MA

Educational Researcher, Vol. 46 No. 8, pp. 420–433
DOI: 10.3102/0013189X17732753

Literature Review

Teaching to the Test or Creating Coherence? The Alignment Debate

One theory of action underlying test-based accountability presumes that holding schools and teachers accountable for students' test scores will create incentives to improve instructional quality to increase student learning (Reeves, 2004). However, the relationship between test-based accountability and student outcomes depends in large part on the way in which teachers adapt their practices to the tests themselves and the standards they aim to assess (Polikoff, 2012). Thus, aligning classroom practice to tests could have either positive or negative effects on teaching quality (Koretz, McCaffrey, & Hamilton, 2001).

On the negative end, aligning instruction to tests may encourage the teaching of superficial and procedural knowledge likely to be measured on many assessments. Aligning instruction to high-stakes tests may also result in reallocating instructional resources to the narrow subset of content that appears on these assessments at the expense of the rest of the domain of interest (Koretz, 2005). Some researchers have raised concern that high-stakes testing creates incentives for teachers to focus their efforts on moving "bubble" students from one side of a proficiency threshold to another, taking away attention from other students (Booher-Jennings, 2005; Bulkley, Christman, Goertz, & Lawrence, 2010).

Alternatively, aligning instruction to tests could have desirable effects on teaching practice and student outcomes *if* both the test and the standards to which they are aligned are high quality and rigorous (Darling-Hammond & Adamson, 2014). Cognitively demanding assessments designed around authentic tasks that replicate how students will encounter content in the real world may motivate teachers to shift their instruction toward a student-centered pedagogy (Au, 2007; Newmann, Bryk, & Nagaoka, 2001). This line of thinking spurred the standards movement of the 1990s (e.g., Resnick & Resnick, 1992) and more recently the widespread adoption of the Common Core State Standards. Scholars of the first standards movement often argued that the adoption of more rigorous standards and tests was an insufficient approach to improving classroom instruction (McLaughlin & Shepard, 1995; National Research Council, 1999), particularly when standards and tests were not accompanied with implementation supports for teachers and schools (Carnoy, Elmore, & Siskin, 2003). Still, proponents of new testing programs aligned to the Common Core State Standards have expressed optimism that these assessments—widely perceived to be more rigorous than many of those previously used for accountability purposes—will motivate more ambitious instruction in U.S. classrooms, in part by decreasing the motivation for test preparation focused on narrow curricular goals (Peery, 2013; Phillips & Wong, 2010).

Test Preparation and Ambitious Instruction: The Evidence

Despite staunch narratives about the role of high-stakes testing in schools, there has been relatively little empirical investigation into the relationship between test preparation activities and the extent to which teachers engage in ambitious, inquiry-oriented instruction. Studies that have investigated this topic generally are small in scale and descriptive in nature (for one example and a review of similar studies, see Firestone, Mayrowetz, & Fairman, 1998).

We are aware of three studies that explore quantitatively the relationship between test preparation and instructional quality. Using three years of classroom observation data and interviews from 70 teachers from 11 schools, Valli, Croninger, and Buese (2012) found that as schools experienced pressure to make annual yearly progress on state exams, teachers' instruction was characterized by less cognitive demand. Teachers less frequently evoked student reasoning, required higher-order thinking, or provided challenging content. Declines in conceptual instruction and increases in procedural instruction were most pronounced in the weeks leading up to state exams. In an analysis drawing on a subset of the data used in this paper, Hill, Blazar, and Lynch (2015) identified a negative but relatively weak relationship between upper-elementary teachers' reports of their engagement in test preparation and outside observers' assessment of the quality of their mathematics instruction. As this finding was part of a larger analysis on predictors of the quality of elementary mathematics instruction, the authors did not explore this relationship in depth. The most comprehensive study of test preparation and ambitious instruction that we found used self-reports of the frequency of test preparation activities (i.e., having practice sessions with test-like items, teaching test-taking mechanics) from a stratified random sample of 247 fourth-grade math teachers throughout New Jersey (Firestone, Monfils, & Schorr, 2004). The study also collected classroom observations and interview data from a nonrandom subset of 78 teachers from seven districts in the state. With both survey and observational data, the authors found variation in the type and quality of test preparation math instruction; some lessons were characterized by procedural instruction and others by more inquiry-based practices.

The evidence to date aligns to some extent with hypotheses about the negative relationship between test preparation activities and instructional quality. Yet, it is difficult to draw broad conclusions from this work given small samples, self-reported instructional quality, and the endogeneity of test preparation activities. We also argue that a key question has been left fully unexplored: whether more complex and demanding tests create opportunities for higher quality and more ambitious forms of test preparation instruction relative to non–test preparation instruction.

Sample

This study drew on data from 328 fourth- and fifth-grade teachers from five school districts (henceforth Districts 1–5) on the East Coast of the United States. Data collection occurred between the 2010–2011 and 2012–2013 school years. Although teachers volunteered to participate, descriptive statistics suggest that their makeup reflects national patterns (Snyder, 2014). The vast majority are White females who earned their teaching credential through traditional certification programs (see Table 1). Analyses of these data in other work indicate that teachers who

Table 1
Sample Descriptive Statistics

	Full Sample	Survey Sample	Lesson Sample
Male (%)	15.31	16.00	10.00
African American (%)	20.74	21.71	5.08**
White (%)	65.94	64.47	86.44**
Novice teacher (%)	10.73	11.33	8.77
Traditionally certified (%)	84.97	84.78	94.55~
Master's degree (%)	74.61	75.66	74.55
Math content knowledge (standardized)	0.00	0.01	–0.02
N teachers	328	308	60

~$p < .10$. **$p < .01$ on difference between subsample (Columns 2 or 3) and full sample (Column 1).

agreed to participate in this study had similar value-added scores as other teachers in their district (Blazar, Litke, & Barmore, 2016).[1] These value-added scores were derived from high-stakes tests that many think may be manipulated by test preparation (Koretz, 2008).

Our analyses focus on two subsamples from this larger group. One sample includes teachers from Districts 1 through 4 (N = 308) who completed a survey asking about their engagement in test preparation activities. Teachers from District 5 participated in a separate randomized control trial of a mathematics professional development program that collected some of the same data as the first project but not the survey of test preparation activities. Descriptive statistics look similar between the group of 308 teachers and the full sample, without any statistically significant differences. The second sample includes 60 teachers from all five participating districts for whom we have two types of videotaped lessons: lessons that explicitly aimed at preparing students for state tests and lessons in which test preparation was not a focus of instruction. Other teachers in our study only contributed lessons where test preparation activities were not observed. Focusing on the subsample of teachers with both types of lessons allows us to compare the instructional quality of test preparation and non–test preparation lessons from the same teacher, thus avoiding several likely sources of selection and omitted variables bias. Limiting the sample in this way, though, creates a tradeoff with regard to external validity. Teachers who contributed both test preparation and non–test preparation lessons to the project were more likely to be White and traditionally certified than those who only contributed non–test preparation lessons.

Related analyses from these same data provide additional information on these districts and the high-stakes tests administered. Districts 1 and 2 came from the same state and took the same high-stakes assessment. Lynch, Chin, and Blazar (2017) coded test items for format and cognitive demand using the Surveys of Enacted Curriculum framework (Porter, 2002) and found that it was a relatively demanding assessment. Items often asked students to solve nonroutine problems, including looking for patterns and explaining their reasoning. Roughly 40% of items required short or open-ended responses; the rest were multiple choice. In District 3, the high-stakes assessment was basic skills oriented, asking students to answer mostly procedural, multiple-choice questions. Twelve percent of items were short response, and 2% were open ended. The assessments in Districts 4 and 5 also had low levels of cognitive demand, and all items were multiple-choice format.

Data

Mathematics Lessons

Mathematics lessons were captured by videotape over the course of three years, with an average of three lessons per year for teachers in Districts 1 through 4 and six lessons per year for teachers in District 5. Teachers were allowed to schedule recordings during times that were convenient for them and when students were not taking a test. Videos were recorded using a three-camera, unmanned unit; site coordinators turned the camera on prior to the lesson and off at its conclusion. Most lessons lasted between 45 and 60 minutes.[2]

These lessons were scored on the *Mathematical Quality of Instruction* (*MQI*, 2014) observation instrument, which captures the cognitive demand of the math activities that teachers provide to students, teachers' interaction with students around that content, and the accuracy of the mathematical material delivered (Hill et al., 2008).[3] We focus our analyses on the Ambitious Mathematics Instruction dimension because it captures the type of complex practices that many think are crowded out of classrooms due to a focus on test preparation. In other work, teachers' Ambitious Mathematics Instruction scores have been shown to relate to students' academic performance on a low-stakes math test (Blazar, 2015), supporting the importance of this type of instruction for student learning. One item, *linking and connections*, captures instances where the teacher makes explicit connections between multiple mathematics representations (e.g., a numeric fraction and a fraction pie). *Explanations* captures instances in which teachers give meaning to ideas, procedures, steps, or solution methods. A third item, *teacher uses student productions*, captures instances where teachers build their instruction off of student ideas, including appropriately identifying mathematical insight in students' questions, comments, or work.

Two certified and trained raters scored each of the aforementioned three items and seven additional items on a scale from 1 (*not present*) to 3 (*present and sustained*) for each 7.5-minute segment of a given lesson. In our analyses, we use lesson- and teacher-level Ambitious Mathematics Instruction scores. We

Table 2
Univariate Descriptive Statistics of Main Dependent and Independent Variables

	All Teachers	District 1	District 2	District 3	District 4	District 5
Dependent variables						
Ambitious Mathematics Instruction	1.27	1.38	1.24	1.20	1.26	1.25
(teacher level)	(0.11)	(0.12)	(0.07)	(0.10)	(0.08)	(0.06)
Ambitious Mathematics Instruction	1.25	1.31	1.24	1.19	1.24	1.25
(lesson level)	(0.14)	(0.20)	(0.15)	(0.12)	(0.13)	(0.14)
Independent variables						
Test preparation composite	2.56	2.48	2.66	2.63	2.53	NA
	(0.56)	(0.53)	(0.54)	(0.72)	(0.60)	
Use items	2.26	2.27	2.39	2.50	2.11	NA
	(0.79)	(0.82)	(0.84)	(1.23)	(1.09)	
Incorporate formats	2.56	2.47	2.61	2.81	2.49	NA
	(0.72)	(0.93)	(0.96)	(1.19)	(1.03)	
Test-taking strategies	2.54	2.54	2.78	2.36	2.51	NA
	(0.87)	(0.98)	(0.96)	(1.16)	(1.02)	
Reallocate time	2.49	2.33	2.57	2.65	2.49	NA
	(0.83)	(0.88)	(0.87)	(1.22)	(1.08)	
Focus on bubble students	2.94	2.79	2.93	2.80	3.07	NA
	(0.79)	(1.07)	(0.91)	(1.20)	(0.97)	
N teachers	328	70	56	49	133	20

Note. Standard deviations in parentheses below means. Items within the Ambitious Mathematics Instruction dimension are on a scale from 1 (*not present*) to 3 (*present and sustained*). Survey items are on a scale from 1 (*never or rarely*) to 4 (*daily*). Sample sizes at the bottom of the table refer to the full sample of teachers in each district. For lesson-level Ambitious Mathematics Instruction scores, sample sizes are smaller: 60 total teachers who contributed both test preparation and non–test preparation lessons to the project, with 4 from District 1, 15 from District 2, 1 from District 3, 20 from District 4, and 20 from District 5.

calculated lesson-level scores by first averaging scores for each item across raters, across each 7.5-minute segment, and finally across the 10 items within this dimension. We calculated teacher-level scores by averaging the lesson-level scores across all available lessons; for teachers who participated in the study for multiple years, this means averaging scores across years. These scores adequately capture the quality of instruction, with adjusted intraclass correlations of 0.89 and 0.71 for the lesson- and teacher-level scores, respectively.[4] Average interrater agreement is 0.74 across the Ambitious Mathematics Instruction items. As shown in Table 2, the teacher- and lesson-level scores have similar means (1.27 and 1.25, respectively) and standard deviations (*SD*; 0.11 and 0.14, respectively). Average Ambitious Mathematics scores are highest in District 1 and lowest in District 3. For our final analyses, we standardized both teacher- and lesson-level scores within the full sample to have a mean of 0 and a *SD* of 1.

We used additional information from MQI scoring to identify the subset of lessons that explicitly aimed at preparing students for state tests. These were lessons in which test preparation was a major focus of instruction; however, this did not preclude teachers from engaging in other activities as well.[5] After watching each lesson, raters identified the topic of the lesson, wrote a two- to three-paragraph narrative of activities that occurred in the lesson, and listed specific strengths and weaknesses. We searched the text of these summaries with a list of 70 terms compiled from the glossaries of the Educational Testing Service, the assessment pages of participating districts' websites, and other terms that we learned were associated with test preparation in certain districts (see Appendix Table A1; we do not include in this list terms that threaten the anonymity of sample districts). After flagging summaries containing one or more of these terms, two researchers read each summary to determine if it should be categorized as a test preparation lesson. In instances where summaries did not provide enough information, we reviewed the lesson transcripts and videos. This process resulted in 73 lessons from 60 teachers being categorized as engaging in explicit test preparation. All 60 of these teachers also contributed videotaped lessons to the project in which test preparation was not an explicit focus of instruction, allowing us to compare these lessons to each other.

Teacher Survey

Our second main data source was a survey administered in the fall of each school year that asked teachers about their engagement in five types of test preparation activities. All five items were developed based on Koretz's (2005) framework on inappropriate test preparation activities thought to boost test scores at the expense of building students' generalized content knowledge. The survey asked about the extent to which teachers used standardized test items in their instruction (use items); incorporated item formats (incorporate formats); taught test-taking strategies, such as process of elimination or plugging in answers (teach test-taking strategies); set aside time to review concepts most likely to be found on the state test (reallocate time); or focused their instruction on students expected to score just below a given performance level on the state test (focus on bubble students). Teachers answered each of these questions on a 4-point scale from 1 (*never or rarely*) to 4 (*daily*), capturing the number of days in a typical week that included these activities.[6] As described earlier, engaging in test preparation during one day or lesson likely did not preclude teachers from engaging in other activities as well. To limit the threat due to multiple hypothesis testing, we also created a composite measure of test preparation by averaging teachers' responses across items (internal consistency reliability = 0.80).

We present univariate and bivariate descriptive statistics for these items in Table 2. Histograms of survey items indicate that all five are roughly normally distributed (see Appendix Figure A1). On average across districts, teachers reported engaging in each activity between once or twice in a typical week (2 on the 4-point scale) and three or four times a week (3 on the 4-point scale). Teachers reported the least time spent on using test items (mean = 2.26, *SD* = 0.79) and the most time focusing on bubble students (mean = 2.94, *SD* = 0.79). On average, teachers in District 3 reported using items, incorporating formats, and reallocating time more frequently than teachers in other districts. Comparatively, teachers in District 2 more frequently reported engaging in test-taking strategies, while teachers in District 4 more frequently reported focusing on bubble students. Pairwise correlations between items range from 0.29 (between use items and focus on bubble students) to 0.63 (between use items and incorporate formats), suggesting that teachers who engaged in one type of test preparation activity sometimes engaged in others (see Appendix Table A2). We standardized items to have a mean of 0 and a *SD* of 1. We did so within each school year to account for slight differences in wording of two items in the third year of the study. For teachers who took the survey in multiple years, we averaged these standardized scores across years.[7]

Analyses

The underlying question of interest to policymakers and practitioners is causal in nature: Does test preparation result in low-quality and unambitious mathematics instruction? The hypothesized model that describes this relationship is straightforward:

$$\begin{aligned} &\textit{Ambitious Mathematics Instruction}_j = \\ &\beta \textit{Test Preparation Activities}_j + \varepsilon_j. \end{aligned} \tag{1}$$

The outcome of interest captures the degree of ambitious teaching from teacher *j* and is predicted by his or her engagement in different test preparation activities. The coefficient of interest, β, describes the relationship between teachers' engagement in test preparation and their instructional quality. Based on prior theory and research, we hypothesized that this relationship would be negative.

However, in all analyses on this topic, including our own, test preparation is not randomly assigned to teachers. Those most likely to engage in high levels of test preparation may be systematically different from those who do not. Teachers who lack the personal resources necessary for teaching may rely on test preparation materials as a form of scripted curricula (Au, 2011). Teachers whose students have low incoming test scores may be more inclined to engage in test preparation to boost these scores (Diamond & Spillane, 2004). Teachers with a large proportion of students with special needs or limited English proficiency may engage in less test preparation if these students' scores are unreported when disaggregated sample sizes do not meet minimum reporting requirements set by state accountability guidelines (Booher-Jennings, 2005). School-level pressures related to students' achievement and background characteristics may create additional incentives for teachers to engage in test preparation (Diamond & Spillane, 2004). Finally, district-level differences, including the extent to which test scores are used to evaluate teachers, may create different incentives for teachers to engage in test preparation (Herlihy et al., 2014).

We addressed these concerns with two approaches. First, we modified Equation 1 to include a rich set of control variables that aim to account for the most likely reasons that teachers vary in the extent to which they engage in test preparation activities: teachers' own resources for teaching (i.e., experience in the classroom, certification pathway, whether or not they had a master's degree, mathematical content knowledge; for information on these measures, see Hill et al., 2015), background characteristics of teachers' students as captured in district administrative records (i.e., baseline achievement on high-stakes math tests, percentage of students eligible for special education services, percentage of students identified as limited English proficient), background characteristics of teachers' schools (i.e., the same student characteristics aggregated to the school level), and district fixed effects. Indeed, in Appendix Table A3, we show that many of these characteristics predict teachers' self-reported engagement in test preparation activities. In results presented in the following, we examine the sensitivity of our estimates of the relationship between test preparation and Ambitious Mathematics Instruction to different combinations of these control variables.

Given the presence of other unobserved characteristics that may be omitted from our models and thus could lead to bias, we took a second approach in which we fit a model comparing the instructional quality of lessons explicitly aimed at preparing students for state tests to other lessons from the same teacher not explicitly focused on test preparation. This approach avoids bias due to an individual teacher's underlying quality and his or her inclination to engage in different levels of test preparation. To do so, we regressed the Ambitious Mathematics Instruction score for lesson *l* and teacher *j* on an indicator for whether or not that lesson focused on test preparation, $\textit{Test Preparation}_{lj}$, and teacher fixed effects, τ_j:

$$\begin{aligned} &\textit{Ambitious Mathematics Instruction}_{lj} = \\ &\delta \textit{Test Preparation}_{lj} + \tau_j + \varepsilon_{lj}. \end{aligned} \tag{2}$$

We restricted this analysis just to those teachers who provided test preparation and non–test preparation lessons to the project (*N* = 60 teachers and 537 lessons), given that δ was estimated only from these teachers. We clustered standard errors at the teacher level to account for the fact that we have multiple lessons per teacher.

In our final analysis, we examined whether the relationship between test preparation and instructional quality varied across

Table 3
Relationship Between Ambitious Mathematics Instruction and Test Preparation Activities, Varying the Control Set

	Model 1	Model 2	Model 3	Model 4	Model 5
Test preparation composite	–0.20***	–0.14**	–0.14*	–0.13*	–0.10*
	(0.06)	(0.05)	(0.06)	(0.05)	(0.04)
Background teacher characteristics		X	X	X	X
Background class characteristics			X	X	X
Background school characteristics				X	X
District fixed effects					X
N teachers	308	308	308	308	308

Note. Estimates are standardized effect sizes. Robust standard errors clustered at the school level in parentheses. Background teacher characteristics include: experience in the classroom, certification pathway, whether or not they had a master's degree, and mathematical content knowledge. Background characteristics of teachers' students include: average incoming level of achievement on high-stakes math tests, percentage of students eligible for special education services, and percentage of students identified as limited English proficient. Background characteristics of teachers' schools aggregate the three student characteristics to the school level.
*$p < .05$. **$p < .01$. ***$p < .001$.

districts and thus might be related to the high-stakes tests that teachers were preparing students to take. This analysis uses items from the teacher survey as our measure of test preparation and thus excludes District 5 where this survey was not administered. It was not possible to run this moderation analysis using test preparation versus non–test preparation lessons as sample sizes of teachers in each district with both types of lessons were too small. We hypothesized that the negative relationship between teachers' engagement in test preparation activities and the ambitious nature of their mathematics instruction would be most pronounced in districts where the state test was oriented around basic skills (i.e., in Districts 3 and 4) rather than cognitively demanding activities (i.e., in Districts 1 and 2). For example, use of multiple-choice items that asked students to perform basic procedures may result in less ambitious instruction than use of open-ended or short-response test items that reviewed students' understanding of concepts.

To test these hypotheses, we modified Equation 1, interacting our test preparation survey items with dummy variables for each district, specified in the model as district fixed effects, γ_d :

$$\begin{aligned} &\textit{Ambitious Mathematics Instruction}_{jsd} = \\ &\pi \textit{Test Preparation Activities}_j {}^{*}\gamma_d + \omega \textit{Background} \\ &\qquad \textit{Teacher Characteristics}_j \\ &+\varphi \textit{Background Student Characteristics}_j \\ &+\theta \textit{Background School Characteristics}_{js} \\ &+\gamma_d + \varepsilon_{jsd} \end{aligned} \tag{3}$$

The outcome of interest is the Ambitious Mathematics Instruction score for teacher j in school s and district d. Our main parameters of interest are in the vector, $\boldsymbol{\pi}$, which describe the relationship in each district between teachers' reported engagement in test preparation activities and their observed quality of instruction. To examine cross-district differences in the relationship between test preparation and instructional quality, we conducted a series of post hoc tests comparing the magnitude of our interaction variables. Ultimately, we leveraged variation across a small number of districts and assessments. Thus, we consider this approach as illustrating patterns rather than providing conclusive evidence about the relationship between test rigor, test preparation, and the ambitious nature of teachers' mathematics instruction.

Results

In Tables 3 and 4, we examine the relationship between Ambitious Mathematics Instruction and test preparation using our teacher survey items. The first of these tables focuses on the composite measure of test preparation activities as our main predictor and examines the sensitivity of the relationship between this variable and Ambitious Mathematics Instruction to different sets of control variables that we hypothesized may lead to biased estimates if omitted. This analysis informs the full set of control variables that we include in subsequent analyses. Here and later in the paper, estimates are presented as standardized effect sizes. Robust standard errors clustered at the school level are in parentheses.

Large differences in estimates across the models presented in Table 3 indicate that, indeed, there likely are several omitted variables that bias the relationship between teachers' engagement in test preparation activities and Ambitious Mathematics Instruction. Compared to Model 1, which includes no controls, the estimate in Model 5, which includes all theoretically driven controls, is roughly half as large. The largest differences emerge after controlling for background teacher characteristics (Model 2) and district fixed effects (Model 5), suggesting that excluding such controls would lead us to substantially overstate the negative effect of test preparation on the ambitious nature of teachers' mathematics instruction. After controlling for background teacher characteristics, inclusion of background characteristics of teachers' students and their schools does not appear to change the inferences we would draw. From a perspective of bias, this is heartening and suggests that other related characteristics that are unobserved in our data but very well could be thought of as "omitted variables" may not in fact lead to large degrees of bias. From a construct validity perspective, this also suggests that even though different types of students do appear to receive different

Table 4
Relationship Between Ambitious Mathematics Instruction and Test Preparation Activities

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Test preparation composite	−0.10*					
	(0.04)					
Use items		−0.10*				
		(0.05)				
Incorporate formats			−0.02			
			(0.04)			
Test-taking strategies				−0.07		
				(0.05)		
Reallocate time					−0.10*	
					(0.05)	
Focus on bubble students						−0.06
						(0.05)
N teachers	308	308	308	308	308	308

Note. Estimates are standardized effect sizes. Robust standard errors clustered at the school level in parentheses. All models control for background teacher characteristics (i.e., experience in the classroom, certification pathway, whether or not they had a master's degree, and mathematical content knowledge), background characteristics of teachers' students (i.e., average incoming level of achievement on high-stakes math tests, percentage of students eligible for special education services, and percentage of students identified as limited English proficient), background characteristics of teachers' schools (i.e., the same three student characteristics aggregated to the school level), and district fixed effects.
$^*p < .05$.

amounts of test preparation (see Appendix Table A3), practically speaking, we are not conditioning out the unique experiences of historically disadvantaged students by including these controls in our models.

We move to Table 4 to examine substantively the relationship between test preparation activities and teachers' Ambitious Mathematics Instruction. Based on results from Table 3, we continue to control for background characteristics of teachers and district fixed effects. We also include other controls describing teachers' students and their schools in case these characteristics lead to bias when examining specific test preparation activities rather than the survey composite utilized in Table 3. Our first model focuses on the composite measure of test preparation activities and thus is identical to the result presented in Model 5 of Table 3. Here, we find that test preparation is a significant and negative predictor of teachers' Ambitious Mathematics Instruction score. A 1 *SD* increase in the number of days that included test preparation activities is associated with a 0.10 *SD* decrease in the ambitious nature of their mathematics instruction. We also disaggregate results by survey item to examine whether specific activities drive this relationship. Given moderate to strong correlations between survey items (see Appendix Table A2), we fit models with each item entered separately.[8] In Models 2 and 5, we see statistically significant and negative relationships for use items (effect size = −0.10 *SD*) and reallocate time (effect size = −0.10 *SD*). Converting these effect sizes back to the raw scale suggests that teachers who reported engaging in these activities roughly one day more per week than the average teacher scored 0.10 *SD* lower on our measure of instructional quality.[9] These results highlight two specific activities that may drive test preparation instruction to be less ambitious than desired.

Although results presented in Table 3 suggest that we have accounted for several variables that could lead to bias, results in Table 4 still are observational in nature. We aim to reduce lingering biases with our model in Table 5, where we compare the instructional quality of math lessons explicitly aimed at preparing students for state tests to other lessons from the same teacher that did not do so. In a naïve model that simply compares mean differences between these two types of lessons across teachers, we observe that lessons that explicitly focused on test preparation scored roughly a third of a *SD* lower on Ambitious Mathematics Instruction than those that did not do so. When we add in teacher fixed effects, thus restricting our comparison to lessons from the same teacher, we see a marginally significant negative relationship that is smaller in magnitude (−0.25 *SD*). This pattern is similar to what we observed in Table 3, where failure to account for factors related to the nonrandom selection of test preparation to teachers led us to overstate the relationship between these activities and teachers' Ambitious Mathematics Instruction.[10] However, we are cautious in overinterpreting the

Table 5
Differences in Ambitious Mathematics Instruction Lessons That Explicitly Focus on Test Preparation and Those That Do Not

	Model 1	Model 2
Test preparation (dummy)	−0.33*	−0.25~
	(0.13)	(0.14)
Teacher fixed effects		X
N teachers	60	60
N lessons	537	537

Note. Estimates are standardized effect sizes. Robust standard errors clustered at the teacher level in parentheses.
$\sim p < .10$. $^*p < .05$.

Table 6
Relationship Between Ambitious Mathematics Instruction and Test Preparation Activities by District

	Model 1: IV = Test Preparation Composite	Model 2: IV = Use Items	Model 3: IV = Incorporate Formats	Model 4: IV = Test-Taking Strategies	Model 5: IV = Reallocate Time	Model 6: IV = Focus on Bubble Students
Test Preparation Activity × District 1	−.33*	−.38*	−.27*	−.31**	−.27*	−.01
	(.14)	(.15)	(.13)	(.12)	(.13)	(.14)
Test Preparation Activity × District 2	−.05	−.02	.05	−.08	−.13~	−.01
	(.08)	(.10)	(.08)	(.07)	(.07)	(.11)
Test Preparation Activity × District 3	−.02	−.02	−.01	.02	.08	−.18~
	(.08)	(.08)	(.08)	(.12)	(.09)	(.09)
Test Preparation Activity × District 4	−.06	−.08	.04	−.00	−.11	−.05
	(.05)	(.06)	(.05)	(.06)	(.08)	(.05)
p value on test of differences between districts						
District 1 = District 2	**.081**	**.044**	**.038**	**.067**	.364	.961
District 1 = District 3	**.059**	**.031**	**.079**	**.045**	**.030**	.305
District 1 = District 4	**.074**	**.065**	**.032**	**.013**	.299	.747
District 2 = District 3	.807	.958	.574	.431	**.081**	.258
District 2 = District 4	.925	.617	.878	.329	.844	.753
District 3 = District 4	.707	.469	.615	.837	.115	.252
N teachers	308	308	308	308	308	308

Note. Estimates in each column come from the same regression model of Ambitious Mathematics Instruction on district by test preparation activity interactions. The test preparation activity in each model is listed in the column header. Estimates are standardized effect sizes. Robust standard errors clustered at the school level in parentheses. All models control for background teacher characteristics (i.e., experience in the classroom, certification pathway, whether or not they had a master's degree, and mathematical content knowledge), background characteristics of teachers' students (i.e., average incoming level of achievement on high-stakes math tests, percentage of students eligible for special education services, and percentage of students identified as limited English proficient), background characteristics of teachers' schools (i.e., the same three student characteristics aggregated to the school level), and district fixed effects. In bottom panel, *p* values equal to or less than .10 are bolded. IV = independent variable.
$\sim p < .10$. $^{*}p < .05$. $^{**}p < .01$.

difference between Models 1 and 2, as 95% confidence intervals around these two estimates overlap.

Finally, in Table 6, we present estimates of the relationship between test preparation activities and Ambitious Mathematics Instruction by district, which allows us to examine whether test preparation activities aligned to less rigorous state tests might also be related to lower quality instruction. As in Table 4, we begin with a model that focuses on a single composite measure of test preparation to mitigate threats of observing false positives due to multiple hypothesis testing. In subsequent models, we disaggregate results by test preparation activity and find that cross-district differences are similar.[11]

Results from these models provide little support for our hypothesis about the moderating role of test rigor. The negative relationship between test preparation activities and Ambitious Mathematics Instruction is driven predominantly by District 1 (effect size = −0.30 *SD*), even though this district had the most cognitively demanding assessment in our sample. We hypothesized that a state test with 40% of items that were short response or open ended and often asked students to solve nonroutine problems, such as identifying patterns, may elevate the quality of instruction from teachers who engage in high levels of test preparation relative to teachers who engage in lower levels of test preparation. Teachers in District 1 reported engaging in lower levels of test preparation than teachers in other districts, but only to a small degree; on average, teachers in this district still engaged in test preparation activities several times each week (see Table 2). Comparatively, we see no such relationship in Districts 3 or 4, which both had low-demand tests that we thought might drive the instruction of teachers who engaged in large degrees of test preparation activities to be lower quality, on average, than teachers who engaged in these activities to a lesser degree (p = .059 and .074 for comparisons of coefficients between Districts 1 and 3 and between Districts 1 and 4, respectively, from Model 1). We also consistently observe that the relationship between test preparation and Ambitious Mathematics Instruction is weaker (i.e., closer to 0 *SD*) in District 2 than in District 1, even though we hypothesized that we should see no difference given that teachers were preparing students to take the same high-stakes tests. Moreover, this difference is most pronounced in Model 2, where the independent variable captures the test preparation activity (use items) arguably focused most narrowly on the format and content of the state test.

We recognize that this analysis is limited by a small number of districts and state tests. That said, our descriptive analyses consistently point away from prior hypotheses surrounding the moderating role of test rigor. This suggests that alternative factors may be equally or more important in explaining our results. We turn to some of these factors in our conclusion.

Discussion and Conclusion

Prevailing sentiments surrounding test-based accountability have held test preparation partly responsible for uninspired teaching in U.S. schools (Amrein & Berliner, 2002; Diamond, 2007; Koretz, 2008). Many have deemed more rigorous Common Core–aligned assessments as worth teaching to and optimistically viewed their widespread adoption as a lever to elevate test preparation to ambitious teaching (Peery, 2013; Phillips & Wong, 2010). Our results support only part of this narrative.

We find that test preparation is a significant and negative predictor of the ambitious and inquiry-oriented nature of upper-elementary teachers' mathematics instruction. This is true across analyses that use two different measures of test preparation—self-reports by teachers and coding of lessons that teachers contributed to our study—as well as across models that account in different ways for many of the most likely sources of selection and omitted variables bias. Our findings align with previous work (Koretz, 2005) suggesting that coaching that focuses on particularities of test items rather than the content they aim to measure may be especially detracting from ambitious instruction.

At the same time that these findings warrant some concern about the relationship between test preparation and the ambitious nature of teachers' mathematics instruction, we question whether these relationships are as large as the prevailing narrative has warned. Results using our teacher survey indicate that small to moderate decrements in instructional quality (roughly 0.10 *SD*) emerge only after a substantive increase in the number of days that teachers reported engaging in test preparation (roughly one day per week, every week). In our preferred analyses comparing the ambitious nature of test preparation lessons to non–test preparation lessons from the same teacher, we find that the former score 0.25 *SD* lower than the latter. These effects on instructional quality are considerably smaller than effects of other instructional interventions, such as math or science professional development or teacher coaching, in the range of 0.60 *SD* (for two meta-analyses, see Kraft, Blazar, & Hogan, 2017; Scher & O'Reilly, 2009). The modest effect size we observe relative to other interventions is especially noteworthy as we consider test preparation to be a more intensive intervention than coaching or other development efforts. On average in our sample, teachers reported engaging in test preparation activities roughly two to three days each week, every week. Comparatively, coaching and other development efforts tend to be spread out over a handful of weeks during the school year.

Improving the ambitious nature of teachers' classroom instruction is, in our view, a worthy goal. Our findings suggest that as researchers, policymakers, and practitioners consider this goal in light of high-stakes testing, some avenues may be more promising than others. One solution proposed by some is to get rid of high-stakes testing altogether (Ravitch, 2011). The passage of the Every Student Succeeds Act in December 2015, though, means that this is not a viable option at least in the near future as states still are required to conduct annual testing and hold schools accountable for these scores. The small to moderate relationships between test preparation and ambitious mathematics instruction described previously also suggest that this would not be a silver bullet. Eliminating testing may help elevate instructional quality to some degree but is unlikely to improve the average mathematics lesson observed in classrooms across the United States, which to date have been described as mostly procedural in nature (Blazar et al., 2016; Kane & Staiger, 2012; Stigler, Gonzales, Kwanaka, Knoll, & Serrano, 1999).

Another straightforward (though not simple) approach may be to improve what is being tested. Those who propose using authentic and challenging assessments to drive educational reform hypothesize that these assessments will help elevate the overall quality of instruction in part by creating less of a need or motivation for test preparation focused on narrow curricular goals (Peery, 2013; Phillips & Wong, 2010). However, our results are inconsistent with this hypothesis. For example, we observe that the negative relationship between test preparation activities and Ambitious Mathematics Instruction is driven primarily by District 1, where teachers were preparing students to take the most demanding high-stakes test in our sample. We did not see this same relationship in District 2, even though teachers were preparing teachers for the same high-stakes test; teachers also reported engaging in similar amounts of test preparation. Nor did we observe test preparation activities to detract from the quality of instruction in Districts 3 or 4, both of which had basic skills–oriented tests. This suggests that challenging assessments on their own are unlikely to drive the relationship between test preparation and instructional quality. One alternative explanation for these findings may be that, on average, Ambitious Mathematics Instruction scores were substantively higher in District 1 than in the other three. Thus, when average instructional quality is high, test preparation may be particularly distracting. The patterns we describe in this moderation analysis are descriptive in nature yet provide an important challenge to the hypothesis articulated by some policymakers and scholars that the test itself can serve as a lever for improving instructional quality.

Our findings contribute evidence that testing and test designs are neither a primary source of nor the sole solution to procedural instruction. Thus, we see a critical need to think more comprehensively about the complex relationship between testing, tests, and instruction. This call is not new. Previously, we discuss literature in which researchers describe how creating coherence between standards, assessments, and instruction is a complex process that requires a multifaceted approach (Polikoff & Porter, 2014). In their analysis of the effect of high-stakes testing during the standards movement of the 1990s, Carnoy et al. (2003) concluded that the success of accountability and testing reforms depended on "internal capacity" (p. 10) at the school level. Relatedly, in their exploration of test-based accountability in California and attempts to improve the rigor of mathematics instruction, Cohen and Hill (2008) found that efforts to improve standards and tests were insufficient conditions for increasing teaching quality. They argued that improving teaching and learning required coherence among the tests and several other policy instruments, including curricula and opportunities for high-quality professional development.

Before concluding, it is worth reiterating that our data are descriptive in nature and therefore subject to potential biases.

Although we aimed to limit concerns due to the nonrandom sorting of test preparation to teachers through use of teacher fixed effects and other strategic controls, it is possible that there were additional unobserved variables that we could not account for. It also is possible that there are other ways in which test preparation may be (negatively) related to instructional quality that we could not observe in this study. For example, it is possible that instructional quality might suffer in other subject areas such as science or social studies. In our analysis of the moderating role of test rigor, we were limited by a small number of districts. Our colleagues did identify large differences in test format and cognitive demand across these districts' tests (Lynch et al., 2017). However, with additional districts, including those that have adopted assessments aligned to the Common Core, there may be greater variation in the cognitive demand of state tests that in turn drives larger differences in instruction. Future research may examine these patterns across a larger number of districts to increase sample size and maximize variation in the cognitive demand of state tests as well as strengthen external validity of results.

Practitioners and scholars alike have long argued that high-stakes testing is antithetical to instructional improvement. Our work contributes empirical evidence about the negative relationship between test preparation and ambitious mathematics instruction. However, it also implies a looseness between this relationship as well as the relationship between test rigor and mathematics instructional quality. Thus, positioning testing and new assessments as the primary solution to instructional ills may distract from more comprehensive efforts to elevate teaching in U.S. classrooms. While current testing debates rightfully note the importance of alignment between standards and assessments, we argue that just as important may be the alignment of professional development and other supports to help all teachers and students meet the ideals set out by instructional reforms.

NOTES

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education (Grant R305C090023) to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. Additional support comes from the National Science Foundation (Grant 0918383). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

[1]These analyses focus on Districts 1 through 4. In District 5, student test score data were not available for teachers who did not participate in the research study.

[2]Although the nonrandom sample of lessons is a limitation of this study, analyses from the Measures of Effective Teaching project indicate that teachers were ranked almost identically when they chose lessons themselves compared to when lessons were chosen for them (Ho & Kane, 2013).

[3]Exploratory and confirmatory factor analyses indicate that the 13 total items on the *Mathematical Quality of Instruction* (*MQI*) instrument cluster together to form two unique factors: Errors and Imprecisions (N = 3 items), which focuses on "teacher errors or imprecision of language and notation, uncorrected student errors, or the lack of clarity/precision in the teacher's presentation of the content" (MQI, 2014); and Ambitious Mathematics Instruction (N = 10 items), a phrase used by others to refer to instruction that is "intellectually ambitious, uncertain, and contested" (Cohen & Ball, 1999, p. 6; see also Cohen, 2011; Lampert, 2001). See Blazar, Braslow, Charalambous, and Hill (2017) for information on items and scoring procedures.

[4]The lesson-level intraclass correlation (ICC) calculates the amount of variance in scores attributable to the lesson. Following a generalizability study framework (Hill, Charalambous, & Kraft, 2012), this ICC is adjusted for the median number of segments per lessons. The teacher-level ICC is adjusted for the median number of lessons per teacher.

[5]Our analysis of these lessons indicated that test preparation was a major focus of instruction. However, without rewatching all lessons, it was not possible to identify the amount or percentage of instructional time spent on test preparation activities versus other activities. Developing a coding scheme to capture this breakdown also would be challenging given that test preparation often was not mutually exclusive with other instructional elements. For example, a teacher might introduce new material on multiplying fractions and mention throughout the lesson the connection between these new concepts and standards likely to be assessed on the state test. In other lessons, this distinction may be clearer, where a teacher begins class by going over practice items from an upcoming state test and then transitions to other activities.A related concern is that measurement error in our identification of test preparation lessons could bias estimates of the relationship between test preparation and instructional quality. Because we classified lessons as focused on test preparation when explicit evidence was present, it is likely that lessons with subtler test preparation practices escaped identification. For example, a teacher may have reviewed material for a high-stakes test without stating this explicitly to students or the raters observing this lesson. However, in these instances, we likely would find even smaller differences in the quality of instruction between test preparation and non–test preparation lessons given that on average, the latter had slightly higher Ambitious Mathematics Instruction scores than the former.

[6]In the third year, the scale for the items also changed to 1 (*not at all*) to 5 (*very much*). For descriptive statistics presented in Table 2, we rescaled to 1 to 4 for consistency with scales from the other years. Eleven teachers were missing data on one of these five items. Here, we imputed values to the mean of the full sample.

[7]We averaged survey scores across years, where applicable, for two reasons. First, in analyses that leverage the teacher survey, we were interested primarily in drawing inferences about individual teachers. This is similar to the way that the *MQI* instrument, in particular, has been used in other research settings (Hill, Charalambous, Blazar et al., 2012; Hill, Charalambous, & Kraft, 2012; Kane & Staiger, 2012). Second, we were concerned about measurement error in teacher by year scores. Year-to-year differences in survey responses or observation instrument scores may capture true underlying differences in teachers, or such differences may reflect measurement error (for a discussion of measurement error in survey responses see, e.g., Bound, Brown, & Mathiowetz, 2001; for a discussion of this topic in relation to teacher observation scores, see, e.g., Garrett & Steinberg, 2015). Measurement error in either or both of our measures would attenuate the correlation between them (Spearman, 1904). This is what we observe in a robustness test linking teacher survey responses and videos from the same year (results available on request).

[8]Point estimates are similar when all items are included in the same model, though standard errors are larger, indicating some degree of multicollinearity.

[9]We arrive at this estimate by converting a 1 SD increase in test preparation to units on the raw scale—where 1 SD is equivalent to 0.6 scale points for the test preparation composite and between 0.7 and 0.9 scale points for individual items—and then to days per week—where 1 point on the survey scale is roughly equivalent to 1.2 days per week.

[10]In light of the small sample size relative to the number of variables included in the model (i.e., test preparation indicator and 60

teacher fixed effects), we test the robustness of findings to models that exclude teacher fixed effects and control instead for the same observable characteristics in models from Table 4. Results are similar.

[11]In all models in Table 6, we continue to include district fixed effects to account for differences at the district level that may incent teachers to engage in different levels of test preparation as well as differences in average Ambitious Mathematics Instruction scores across districts. Because this approach makes comparing relationships across districts less direct—in essence, comparing gaps across districts—we also fit models that exclude district fixed effects. We find that district by test preparation activity coefficients and patterns of cross-district differences are similar in these models.

REFERENCES

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing and student learning. *Education Policy Analysis Archives*, *10*(18), 1–74.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267.

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies*, *43*(1), 25–45.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29.

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71–94.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, *53*(2), 324–359.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. *Handbook of Econometrics*, *5*, 3705–3843.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education*, *85*(2), 186–204.

Carnoy, M., Elmore, R. F., & Siskin, L. S. (Eds.). (2003). *The new accountability: High schools and high-stakes testing*. New York, NY: Routledge Falmer.

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement* (CPRE Research Report Series RR-43). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania Graduate School of Education.

Cohen, D. K., & Hill, H. C. (2008). *Learning policy: When state education reform works*. New Haven, CT: Yale University Press.

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. Hoboken, NJ: John Wiley & Sons.

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education*, *80*(4), 285–313.

Diamond, J., & Spillane, J. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *The Teachers College Record*, *106*(6), 1145–1176.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis*, *20*(2), 95–113.

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: Inquiry-oriented and didactic responses. *Assessment in Education*, *11*(1), 67–88.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1–28.

Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open*, *1*(4), 1–23.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., . . . Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, *17*(2–3), 88–106.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching*. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores (CSE Report 655)*. Retrieved from https://www.cse.ucla.edu/products/reports/r655.pdf

Koretz, D. M. (2008). *Measuring up*. Cambridge, MA: Harvard University Press.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions*. Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Kraft, M. A., Blazar, D., & Hogan, D. (2017). *The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence*. Manuscript submitted for publication.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

Lynch, K., Chin, M., & Blazar, D. (2017). Relationship between observations of elementary teacher mathematics instruction and student achievement: Exploring variability across districts. *American Journal of Education*, *123*(4), 615–646.

Mathematical Quality of Instruction. (2014). Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&pageid=icb.page394761

McLaughlin, M. W., & Shepard, L. A. (1995). *Improving education through standards-based reform. A report by the National Academy of Education Panel on Standards-Based Education Reform*. Palo Alto, CA: National Academy of Education, Stanford University.

National Research Council. (1999). *Testing, teaching, and learning: A guide for states and school districts*. San Diego, CA: National Academies Press.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence? Improving Chicago's schools*. Chicago, IL: Consortium on Chicago School Research.

Peery, A. B. (2013). Reading for the future: How the Common Core will change instruction. *The NERA Journal*, *48*(2), 1–9.

Phillips, V., & Wong, C. (2010). Tying together the common core of standards, instruction, and assessments. *Phi Delta Kappan*, *91*(5), 37–42.

Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education*, *118*(3), 341–368.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399–416.

Popham, W. J. (2001). Teaching to the test? *Educational Leadership*, *58*(6), 16–21.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14.

Ravitch, D. (2011). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Reeves, D. B. (2004). *Accountability for learning: How teachers and school leaders can take charge*. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 37–75). Boston, MA: Kluwer Academic Publishers.

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness*, *2*(3), 209–249.

Snyder, T. D. (2014). *Digest of education statistics, 2013* (NCES 2014-086). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292.

Stigler, J. W., Gonzales, P., Kwanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. Washington, DC: National Center for Education Statistics.

Valli, L., Croninger, R. G., & Buese, D. (2012). Studying high-quality teaching in a highly charged policy environment. *Teachers College Record*, *114*(4), 1–33.

AUTHORS

DAVID BLAZAR, EdD, is an assistant professor of education policy and economics at the University of Maryland College Park, Department of Teaching and Learning, Policy and Leadership, 2311 Benjamin Building, 3942 Campus Dr, College Park, MD 20740; *dblazar@umd.edu*. His research focuses on efforts to improve teacher and teaching quality, including professional development, the organizational context of schools and districts, and accountability policy.

CYNTHIA POLLARD, EdM, is a doctoral student at the Harvard Graduate School of Education, 13 Appian Way, Cambridge, MA, 02138; *cynthiapollard@g.harvard.edu*. Her research focuses on characteristics of teachers and teaching that are especially effective for students from historically marginalized groups.

Manuscript received November 30, 2016
Revisions received April 25, 2017, and August 28, 2017
Accepted August 30, 2017

Appendix

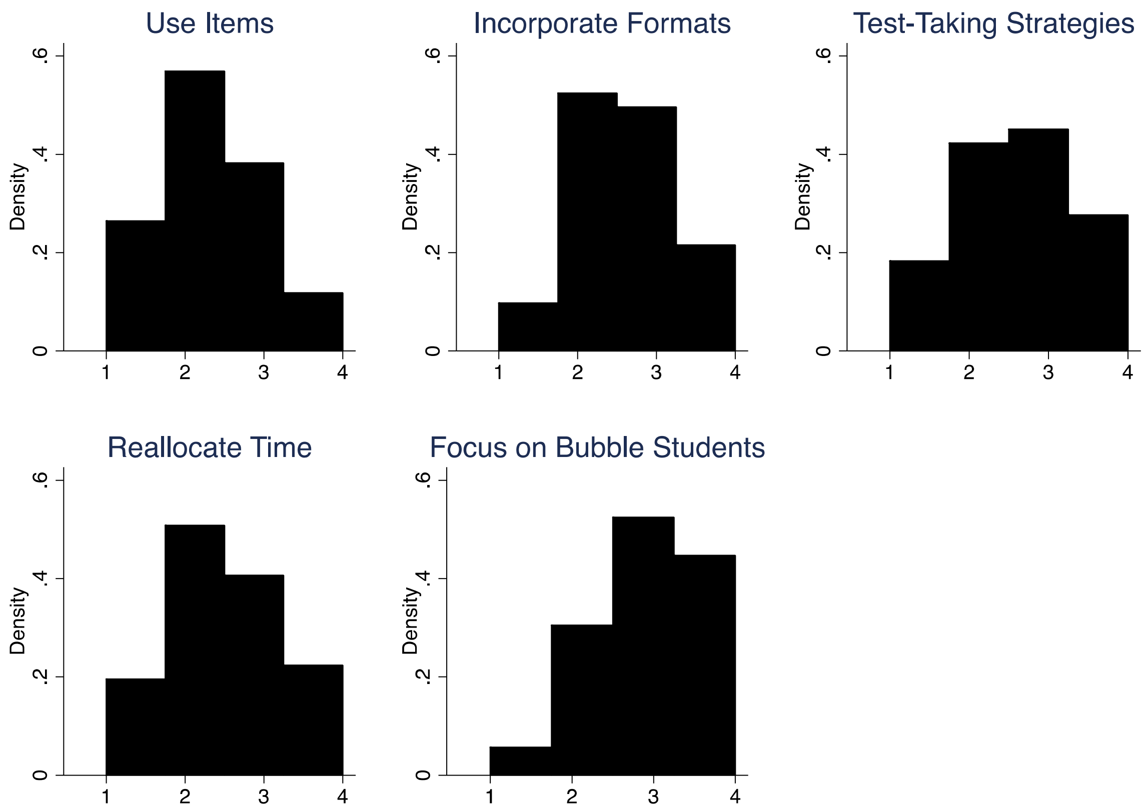

FIGURE A1. *Distributions of responses to test preparation survey items across teachers*

Table A1
List of Test Preparation Search Terms

Advanced	Open response
Assess	Open-response
Assessment	Percentile
Basic	Performance assessment
Below basic	Performance level
Calendar math	Practice test
Closed response	Preparation
Closed-response	Prepare
Constructed response	Proficient
Constructed-response	Rubric
Criterion Referenced Competency Tests	Scale score
criterion-referenced	Scale-score
Cutscore	Score
Cut-score	Score-band
Distractors	Score band
Does not meet	Scoring
Domain	Selected response
Exam	Selected-response
Examination	Standardized
Exceeds	Stanine
High stakes	State test
High-stakes	State-test
Item	Test
Meets	Test review
Multiple choice	Test prep
Multiple-choice	Test-prep
Needs improvement	Upcoming test
Open ended	Warning

Table A2
Pairwise Correlations Between Test Preparation Survey Items

	Use Items	Incorporate Formats	Test-Taking Strategies	Reallocate Time	Focus on Bubble Students
Use items	1.00				
Incorporate formats	0.63***	1.00			
Test-taking strategies	0.49***	0.46***	1.00		
Reallocate time	0.54***	0.50***	0.43***	1.00	
Focus on bubble students	0.29***	0.35***	0.39***	0.41***	1.00

Note. Sample includes 308 teachers.
*** $p < .001$.

Table A3
Predictors of Test Preparation Activities

	Use Items	Incorporate Formats	Test-Taking Strategies	Reallocate Time	Focus on Bubble Students
Background teacher characteristics					
Novice teacher (dummy)	−0.05	−0.55**	−0.34~	−0.41~	−0.17
	(0.21)	(0.18)	(0.19)	(0.21)	(0.19)
Traditionally certified (dummy)	0.21	−0.24	−0.16	0.07	0.04
	(0.16)	(0.19)	(0.19)	(0.20)	(0.19)
Master's degree (dummy)	0.07	−0.04	−0.02	−0.18	−0.09
	(0.14)	(0.15)	(0.15)	(0.14)	(0.15)
Math content knowledge (standardized)	−0.25***	−0.09	−0.23***	−0.05	−0.06
	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)
Background class characteristics					
Class average prior math achievement (standardized)	0.06	−0.24	0.14	−0.01	−0.66***
	(0.16)	(0.15)	(0.17)	(0.16)	(0.18)
Class proportion SPED (standardized)	−0.13*	−0.16**	−0.11	−0.06	−0.08
	(0.06)	(0.06)	(0.07)	(0.06)	(0.08)
Class proportion LEP (standardized)	0.02	−0.11	0.07	0.16~	−0.13
	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)
Background school characteristics					
School average prior math achievement (standardized)	0.01	0.13	0.01	0.02	0.17~
	(0.09)	(0.10)	(0.10)	(0.10)	(0.10)
School proportion LEP (standardized)	−0.09~	−0.11*	−0.06	−0.11~	0.06
	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)
School proportion SPED (standardized)	−0.74**	−0.29	−0.42~	−0.14	0.98***
	(0.28)	(0.25)	(0.25)	(0.26)	(0.26)
p value on joint *F* Test					
District fixed effects (coefficients not shown above)	0.006	0.021	0.090	0.042	0.093
N teachers	308	308	308	308	308

Note. Estimates are standardized effect sizes. Robust standard errors clustered at the school level in parentheses. SPED = students eligible for special education services; LEP = students identified as limited English proficient.
~$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.