FACTORS ASSOCIATED WITH ALIGNMENT BETWEEN TEACHER SURVEY REPORTS AND CLASSROOM OBSERVATION RATINGS OF MATHEMATICS INSTRUCTION

ABSTRACT

We investigated the alignment between a teacher survey self-report measure and classroom observation measure of ambitious mathematics instructional practice among teachers in two urban school districts using two different standards-based mathematics curricula. Survey reports suggested mild differences in teachers' instructional practices between the two districts, whereas observation ratings indicated starker differences. That said, teachers' survey and observer ratings were significantly correlated in both districts. Factors significantly predicting the extent of survey-observation alignment included teachers' grade level, Mathematical Knowledge for Teaching, cognitive demand, and—for one district—teachers' adherence to the surface-level aspects of their curriculum. Qualitative analyses suggested that teachers' survey-observation alignment could be a function of their interaction with colleagues who provided instructional models against which they could gauge the extent of their standards-based instruction. Our study has implications for research on instructional practice, as well as for school district policies to support and evaluate teachers.

Julia Heath Kaufman
RAND CORPORATION

Mary Kay Stein
UNIVERSITY OF PITTSBURGH

Brian Junker
CARNEGIE MELLON UNIVERSITY

TEACHERS' self-report measures of their instructional practices are used in a wide variety of education studies, from research on how professional development shapes teachers' instructional practices (Penuel, Fishman, Yamaguchi, & Gallagher, 2007) to studies that take into account "fidelity" to particular curricula in order to gauge curricular effectiveness (Agodini et al., 2009) or

THE ELEMENTARY SCHOOL JOURNAL VOLUME 116, NUMBER 3

tie implementation to student achievement (Gamoran, Porter, Smithson, & White, 1997). Although other methods of teacher self-report like teacher logs likely capture more about teachers' practice than surveys (Rowan & Correnti, 2009; Stecher et al., 2006), surveys are a popular method for gathering data about instruction because of their low cost and ease of administration.

Despite their continued use, surveys are not always an accurate measure of what teachers do. Studies present a particularly mixed picture about the accuracy of surveys in the case of teachers' reports about their ambitious instructional approaches associated with mathematics standards (Mayer, 1999; Ross, McDougall, & Hogaboam-Gray, 2003; Spillane & Zeuli, 1999; Stecher et al., 2006). Such approaches—exemplified by National Council of Teachers of Mathematics (NCTM) *Principles and Standards for School Mathematics* (2000)—are embodied in a variety of widely used standards-based mathematics curricula that emphasize conceptual understanding, mathematical reasoning and communication, and problem solving (e.g., Connected Mathematics, Investigations, Everyday Mathematics). In this article, we refer to these instructional approaches as "ambitious," "reform" or "standards-based" practices to distinguish them from conventional approaches that have typically focused on procedures and exercises for students to practice those procedures (Stein, Remillard, & Smith, 2007).

Research indicates that teachers tend to overestimate the extent of more ambitious, standards-based instructional practices (Burstein et al., 1995; Mayer, 1999; Ross et al., 2003). Additional studies suggest that inaccurate self-reports may be related to the challenges teachers face in understanding the deeper content and pedagogy associated with standards-based practices (Cohen, 1990; Spillane & Zeuli, 1999) and differences in understanding and use of mathematical language among researchers and teachers (Hill, 2005). Rather than teaching algorithms and monitoring students' procedural uptake and practice of those algorithms, standards-based instruction calls for the use of cognitively challenging instructional tasks and requires teachers to attend carefully to student reasoning and work to scaffold students' partial understandings toward increasingly sophisticated mathematical ideas (Lampert, 1990; Stein, Engle, Smith, & Hughes, 2008). Understanding and enacting such instruction requires intensive training and support (Ball & Cohen, 1996; Cohen & Hill, 2001).

The responsibility for ensuring that teachers are prepared to teach in these new and demanding ways has fallen primarily to districts. District and school leaders make a wide array of decisions about teacher training and support, including the provision of professional development and coaching, curricula, and time for teachers to collaborate with colleagues regarding instruction (Hightower, Knapp, Marsh, & McLaughlin, 2002; Marsh, 2000; Spillane et al., 2002). These decisions shape the quality of teachers' learning opportunities (Stein & Coburn, 2008; Supovitz 2006) and the depth to which they understand and can implement the main tenets of standards-based reform (Stein & Kaufman, 2010). Because district context likely shapes teacher understanding about standards-based reforms, and because that understanding has been associated with the accuracy of teachers' self-reports (Cohen, 1990; Spillane & Zeuli, 1999), we hypothesize that district context plays a role in shaping how closely teachers' survey responses correspond to observations of their actual instruction.

Purpose of This Study

The purpose of this study is to examine the alignment between elementary teachers' survey reports and observation ratings of ambitious mathematics instructional practices in two urban school districts that provided very different opportunities for teacher learning. By situating our investigation of teachers' survey responses in differing district contexts, we aim to examine the veracity of the above hypothesis: that district context shapes teachers' capacities to respond accurately to survey measures. To our knowledge, no other studies have contrasted accuracy of survey responses in differing school district contexts.

Our results suggest differing patterns in survey-observation alignment for each district and provide both quantitative and qualitative evidence regarding the factors associated with that alignment. These results have implications for researchers gathering data about teachers' instruction, as well as district policymakers, and our study underscores the need for better self-report measures of teachers' instruction.

Review of Relevant Literature

Teachers Approximate the Frequency, but Not Quality, of Standards-Based Instructional Practices

Teachers have been shown to accurately approximate the frequency of instructional practices associated with ambitious, standards-based mathematics reform in a survey self-report (Mayer, 1999; Ross et al., 2003). Using small samples of teachers, both Mayer (1999) and Ross et al. (2003) found that teachers with higher self-reports on the frequency of standards-based practices tended to engage in those practices more often in the classroom. Although Mayer (1999) found that teachers overestimated their alignment with "NCTM-like" approaches to mathematics teaching, they "did so systematically, maintaining their position relative to one another" (p. 449), which enabled identification of teachers who engaged in NCTM-like approaches more than others. Burstein et al. (1995) similarly found that although teachers overreported standards-based instructional goals like "raising questions and formulating conjectures," they were often within one survey response category (e.g., almost every day, once or twice a week, once or twice a month) of the actual time they spent on given instructional goals.

While teachers give somewhat accurate, if inflated, reports about the frequency of their ambitious instructional practices, they are less likely to provide accurate reports about the quality of these practices (Spillane & Zeuli, 1999; Stecher et al., 2006). For example, Spillane and Zeuli (1999) studied 25 teachers who reported frequent use of standards-based practices and found that only four of those teachers taught in ways that approximated the "spirit of the mathematics reforms" (p. 19). In Stecher et al. (2006), teachers' survey reports of their reform mathematics practices had little correlation with observation measures of the quality of those practices.

How Teachers' Understanding and Environmental Constraints Could Play a Role in Survey-Observation Alignment

What accounts for teachers' tendency to overestimate the extent of their standards-based mathematics practices? One obvious explanation is that such prac-

tices are socially desirable, especially in the context of district standards-based reforms, and thus teachers deliberately overreport such practices. Research suggests two additional possibilities: teachers may give inaccurate self-reports because (*a*) they lack a deep understanding of tenets of the standards-based reform, or (*b*) they report what they'd prefer to be doing, rather than what they actually are doing.

Some studies indicate that teachers may believe—and thus report—that they are engaging in standards-based practices when they are doing so superficially or not at all because they do not have a complete understanding of the reform (Cohen, 1990; Spillane & Zeuli, 1999) or do not understand the reform in the same way as researchers who are measuring it (Hill, 2005). A now-classic example is Cohen's 1990 case study of "Mrs. Oublier." While Mrs. Oublier communicated her enthusiastic engagement in standards-based mathematics reforms, observations of her teaching indicated only surface adherence to standards-based reforms, with class discussions consisting of rapid-fire, teacher-led exchanges with little sharing of student thinking. Similarly, the 21 out of 25 teachers in Spillane and Zeuli's (1999) study who reported high use of standards-based practices used terms like "problem solving" to describe routine, traditional practices in their classroom, which—according to the authors—suggested teachers' difficulties understanding the "epistemological regularities" of standards-based instruction.

A second possibility is that some teachers may have extensive knowledge about the meaning of standards-based practices but feel constrained or unsupported to engage in those practices given their district and school environment. Such a possibility is reflected in case study research, where teachers' concerns about testing and behavioral management influenced how much they used problem-solving activities and asked students to explain their thinking (Mayer, 1999; Peterson, 1990; Wilson, 1990). Wilson (1990), for example, noted that the press for high test scores and multiple, conflicting policy messages in one teacher's school appeared to affect both his knowledge of standards-based instructional strategies and the nature of his instruction. In such an environment, it is not unreasonable to speculate that teachers' survey reports may reflect their intentions to engage in reform practices more than their actual use of such practices.

If differences in teacher knowledge and environment are related to the accuracy of teacher self-reports, then the district context is implicated because district leaders make many decisions about the policy messages, training, support, and curricular materials that affect teacher learning and instruction. However, research comparing surveys and observations has focused on teachers as individuals isolated from district context. One possible exception is Spillane's (1999) study of teachers in different districts who reported high use of standards-based practices but engaged in those practices at differing levels. However, no work has been done to examine survey-observation alignment quantitatively or over time. Nor has any work considered what specific district-level and teacher-level factors are associated with that alignment.

In this article, we consider teacher survey self-report and observation measures for one aspect of standards-based instruction: inquiry-based practices. The Scaling Up Mathematics project from which we draw our data is particularly well suited for comparing survey-observation alignment in different district contexts because project researchers have documented significant differences in curriculum implementation and instruction in two intensively studied urban districts in two different states:

Region Z and Greene (pseudonyms). Specifically, Greene was noted to have more curriculum-specific professional development and coaching (Stein & Coburn, 2008), more frequent discussions with others about mathematics (Coburn & Russell, 2008), and higher-quality instruction (Stein & Kaufman, 2010) compared to Region Z. Additionally, Greene's curriculum—Investigations—offered more open-ended tasks and more guidance to teachers compared to the Everyday Mathematics curriculum in Region Z (Stein & Kim, 2009).[1] All of these district-level differences could hypothetically influence survey-observation alignment.

Method

Sample

Region Z, a school district in New York City, included 10 elementary schools with 400 to 800 students each. Approximately 60% of the students were African-American and 35% were Hispanic, with 88% free or reduced-priced lunch students and 10% English language learners (ELLs). Greene, an urban school district in the southwestern United States, included 16 elementary schools and served about 20,000 students. About 87% of Greene's students were Hispanic, 86% were eligible for free or reduced-price lunches, and 50% were ELL.

This study includes data from a 2-year time period which we refer to as Y1 and Y2. Our participants included (1) a subsample of 47 K–5 teachers—23 in Region Z and 24 in Greene—for which we have survey self-report, classroom observation, and interview data in Y1 and/or Y2 and (2) a larger pool of over 700 K–5 teachers in both districts for whom we have only survey data in Y1 and/or Y2.

47-teacher subsample. Our findings focus on survey-observation alignment and the factors associated with that alignment for the subsample of 47 teachers. These teachers were selected from among 58 teachers at eight case study schools (four in each district) for whom we collected survey, classroom observation, and interview data. We used purposive sampling (Strauss & Corbin, 1998) to select case study schools that varied in terms of teacher professional community and teacher knowledge, which were the two areas of focus for the larger study.

We only included case study teachers for whom we had at least one year of survey and observation data, with one year including a single survey and four to six classroom observations ($M = 5.7$ observations per year for Region Z teachers and 5.9 observations per year for Greene teachers). We administered surveys in the spring of each year, and we typically conducted observations on three consecutive lessons in the fall and three consecutive lessons in the spring of each year. For 27 of the 47 teachers (13 in Region Z and 14 in Greene), we have survey and observation data from both years of our study. For the other 20 teachers (10 in each district), we have only one year of survey data because they left their school in Y1 or were new to the school in Y2. Altogether, our data for the 47 teachers include 422 classroom observations and 74 surveys. For 12 out of the 47 teachers, we also drew on qualitative data from four to five semistructured interviews conducted with each teacher in each year.

Larger sample of over 700 teachers. The larger pool of all K–5 mathematics teachers in each district received the same survey as the 47-teacher subset once each year. In Y1, 798 teachers responded to the survey; 756 responded in Y2. The average response rate over both years was 71% in Region Z and 94% in Greene. In our method

section, we discuss use of the survey responses from those 700 teachers for our factor analyses in order to arrive at a survey measure of inquiry-based practice. In our results, we draw comparisons between survey results for the larger sample of 700 teachers and our 47-teacher subsample to discuss the generalizability of our findings.

Measures

Inquiry-based practice (IBP). Because standards-based instruction can be defined in multiple ways, we focused on one core aspect of standards-based instruction: inquiry-based mathematics instructional practice (IBP). We defined teachers' IBP as the extent to which teachers engaged in two activities that have been described as key to standards-based practices in the education literature: (*a*) attention to student thinking: attending to students' mathematical thinking by observing students and prompting them to share their strategies for solving problems (Schoenfeld, 1998; Shifter, 2001; Stein et al., 2008), and (*b*) justifications using mathematics: encouraging students' use of mathematics and mathematical reasoning to judge the accuracy of their and others' approaches to problems (Engle & Conant, 2002; Hamm & Perry, 2002; Lampert, 1990).

Survey measure of IBP. We developed our survey composite measure of IBP through preliminary factor analysis of all the items in our survey intended to measure IBP and then through further factor analyses to test for the presence of additional factors within that composite. The 17 items in our final survey composite are listed in Table 1.

IBP survey measure development. Twenty-seven items from our survey addressed inquiry-based practices, including some items borrowed from other surveys (Ross et al., 2003; Weiss, Banilower, McMahon, & Smith, 2001) and some that we developed based on own work (Stein et al., 2008; Stein, Grover, & Henningsen, 1996) and that of others (Engle & Conant, 2002; Hamm & Perry, 2002; Shifter, 2001). While some of the items employed a disagree/agree response scale, other items used a response scale measuring frequency of inquiry-based practices. Many survey items addressing inquiry-based practice simultaneously focused on attention to student thinking and justifications using mathematics (e.g., "When two students solve the same math problem correctly using two different strategies, I have them share the steps they went through with each other"). We therefore combined all survey items related to inquiry-based practices together for our factor analysis.

Factor analysis. We arrived at our final 17-item survey composite of IBP ($\alpha = .88$) after factor analysis on all 27 survey items intended to measure inquiry-based practice, using maximum-likelihood method and no rotation. This preliminary factor analysis utilized survey data from the 798 elementary teachers in both districts who took the survey in Y1. The factor analysis suggested removal of 10 items—out of the original 27—that appeared unrelated to the remaining items in terms of low factor loadings and conceptual focus.

Using Y1 survey data, we conducted further factor analysis on our 17-item composite to identify any other factors and determine whether additional items should be dropped. We employed maximum-likelihood method and oblimin rotation; Tabachnick and Fidell (2007) suggest oblimin rotation when factors are correlated, and the two main factors in our analysis were moderately correlated ($r(798) = -.63$). Table 1 lists the 17 survey items along with the extracted communalities and factor

Table 1. Factor Loadings and Extracted Communalities for Factors with Eigenvalues Greater than One

	Extracted Communalities	Factor 1	Factor 2
Extent of agreement with statements about how you teach mathematics (1 = strongly disagree; 5 = strongly agree):			
I like to use math problems that can be solved in many different ways[a]	.26	.36	.35
When two students solve the same math problem correctly using two different strategies, I have them share the steps they went through with each other[a]	.27	.31	.41
How frequently do you do the following in your mathematics instruction (1 = never; 5 = always):			
I pose open-ended questions[b]	.32	.28	.49
I engage the whole class in discussion	.39	.23	.58
I require students to explain their reasoning when giving an answer[b]	.51	.31	.65
I ask students to explain concepts to one another[b]	.54	.27	.68
I ask students to consider alternative methods for solutions[b]	.51	.27	.66
Before students begin work on a task, I tell them that they will be able to check the accuracy of their work by checking with me as soon as they've finished. (RC)	.34	.55	−.20
When assigning a set of problems, I tell my students which procedure they should use. (RC)	.48	.70	−.01
Before turning an open-ended project over to my students, I walk them through an example of how to successfully attack the problem. (RC)	.51	.68	−.22
Before turning an open-ended project over to my students, I give them a detailed roadmap to follow through the project. (RC)	.57	.74	−.17
I provide students with more steps to follow than what appears in the curriculum that I use. (RC)	.35	.58	−.14
When students get stuck on a multistep problem, I walk them through the steps they need to perform. (RC)	.59	.74	−.22
After students have worked on a particularly challenging assignment, I provide opportunities for them to see how others have approached the assignment.	.37	.11	.59
When students are uncertain about how to get started on an open-ended project, I tell them how to do the first step. (RC)	.64	.76	−.23
When a student is unable to complete a task on his/her own, I give him/her a set of steps to follow. (RC)	.65	.78	−.20
When students construct their own ways of doing a problem, I have students themselves share their approaches with the rest of the class using their own ways of expressing themselves.	.35	.21	.55

Note.—RC = reverse coded.

[a] All items using this response scale are taken from Ross et al.'s (2003) survey of elementary teachers' commitment to mathematics education reform.

[b] Item taken from the 2000 National Survey of Science and Mathematics Education (Weiss et al., 2001).

loadings for each the two extracted factors with eigenvalues greater than 1.0. The first factor had an eigenvalue of 5.09 and explained 27% of the variance, and the second factor had an eigenvalue of 3.61 and explained 21% of the variance.

The extracted communalities for the two items that used an agree/disagree scale were somewhat lower than for those items that used a frequency scale (i.e., never/always). However, eliminating the agree/disagree items lowered the factor loadings for five of the remaining items to below .25. We therefore decided to keep the two items in our factor analysis although they used a different scale. The second factor in our analysis indicated that teachers' responses to the reverse-scored items were more

similar to one another than to the items that were not reverse scored. The different response pattern for reverse-scored items could reflect a positive bias in teachers' responses to items that were negatively worded, or it could suggest that positively versus negatively worded items measured different aspects of inquiry-based practice. We did not find evidence of separate factors reflecting each of the two practices that are part of our IBP definition—attention to student thinking and justifications using mathematics—which is understandable given that the two practices were confounded in many survey items.

Factor analysis with the same 17 items using Y2 survey data yielded results that were fairly similar to the factor analysis using Y1 data summarized in Table 1. Specifically, the factor with the highest eigenvalue (4.67) loaded positively on almost all items (with a slightly negative factor loading of −.02 for only one item) and explained 27% of the variance, while a second factor with an eigenvalue of 3.5 explained 20% of the variance and indicated a different response pattern for the negatively versus positively worded items.

In our findings, we used regression analysis to examine the variables predicting survey-observation alignment, with the alignment score calculated using the average of the entire 17-item survey composite. However, we also ran regressions with survey-observation alignment calculated separately for reverse-scored versus non-reverse-scored survey items to take into account the second factor that emerged from our analysis. The results of those regression analyses—summarized in our findings and in the Appendix—were strikingly similar to the regressions that utilized the 17-item composite. Thus, for the remainder of this method section and the first part of our findings, we focus on the average of the 17-item IBP composite.

Missing data. Among the 47 teachers for whom we had survey data in one or both years, eight surveys were missing one item for the 17-item IBP composite. To impute missing responses, we regressed the remaining items in the composite on the missing item for all teachers completing the survey across both districts for the year in which the item was missing. We then used teachers' responses in the resulting prediction equation to estimate their response on the missing item. Use of prediction equation to estimate item responses for teachers with no missing data yielded correct predictions for 75% to 89% teachers, depending on the item being predicted.

Survey measure reliability and validity. We assessed test-retest reliability of our composite using survey responses of the same teachers in Y1 and Y2. Given the yearlong gap between survey administrations, the correlation between teachers' responses in each year was surprisingly high ($r(482) = .70$, $p < .001$), and suggests that more conventional test-retest reliability for a shorter interval between test and retest might be even higher. In our results, we also point out the relatively strong correlation between survey and observation IBP for all teachers, which provides further evidence about the validity of our survey composite.

Observation measure of IBP. Our observation measure of IBP is derived from two research phases: (*a*) trained observers documented teacher-student interaction throughout lessons and responded to specific prompts about teachers' inquiry-based practice, and then (*b*) coders studied observer documentation and used rubrics to provide ratings for attention to student thinking and justifications using mathematics in each lesson.

IBP observation measure development. Table 2 lists the prompts provided to classroom observers, which addressed similar aspects of inquiry-based practice to those

Table 2. Prompts and Codes for Observation Measures of Inquiry-Based Mathematics Practice

Classroom Observation Protocol Prompts
What, if anything, did the teacher do to uncover student thinking?
How did the teacher listen to student thinking? What evidence was there that the teacher tried to understand student thinking?
Describe how the teacher assisted student thinking. To what extent did the teacher help students to identify and articulate the key ideas in their work or thinking? How did she help students represent their thinking and keep track of their work? How did she ask questions that pushed student thinking?
Describe how the teacher made student thinking available for the entire class. Did shared student work come primarily from volunteers or did the teacher appear to have a purpose for whose methods were displayed and in what order? Once students displayed their work, what did the teacher do with it? How did the teacher help students to explain their thinking to the entire class? How did she facilitate class discussions about student work?
How and under what conditions did the teacher encourage links between students' informal reasoning and more formal, canonical or sophisticated mathematical thinking?
What did the teacher expect or allow students to discover on their own and under what circumstances? How to approach problems? The concepts that underlie problems? How to organize and record their work? How to justify their conjectures?
What knowledge did the teacher impart or teach to the students and under what circumstances? The steps required to do the mathematical problems? The concepts that underlie the problems? How to attach mathematical notation to their work? The justification of a particular mathematical move?
How and by whom was the correctness of a mathematical answer or approach determined? By the answer in the resource materials? By the flawless execution of a procedure? By a calculator? By mathematical logic?
Attention to Student Thinking Scoring Rubric
0 The teacher did no work to uncover student thinking; he/she did most of the talking in the lesson and/or asked questions with short or one-word answers.
1 The teacher did some work to uncover student thinking by asking some open-ended questions, by asking for some explanations, by arranging for public sharing of student responses, and/or by listening respectfully.
2 In addition to #1, the teacher purposefully selected certain students to share their work during whole-class discussion because she wanted the whole class to hear about the mathematical approach the student took, and/or the teacher connected or sequenced students' responses in a meaningful way.
Justifications Using Mathematics Scoring Rubric
0 The teacher fostered little or no student construction of mathematical ideas, thinking, and/or reasoning. Judgments about correctness were derived from the text or the teacher, with no appeal to mathematical reasoning.
1 The teacher fostered some student construction of mathematical ideas, thinking, and/or reasoning. However, judgments about correctness were mostly derived from the text or the teacher. Nevertheless, some appeals to mathematical reasoning were made.
2 The teacher fostered student construction of mathematical ideas, thinking, and/or reasoning. Additionally, judgments about correctness were primarily (most of the time) derived from mathematical reasoning and discussion during the class.

reflected survey items, including extent to which teachers asked questions intended to uncover student thinking, explanation, and reasoning; gave students the opportunity to share their thinking with others; and allowed students to determine correctness of problems on their own (versus giving them steps to follow). Table 2 also includes the rubrics that coders used to rate lessons based on classroom observer documentation and responses to prompts. The observation prompts and rubrics were developed using the same literature base described for development of survey IBP items.

IBP observation measure procedures. We hired 10 classroom observers who had a master's or doctoral degree in an education-related field with expertise in mathemat-

ics education or qualitative research methods. Our observer training (*a*) provided background about inquiry-based mathematics curriculum and instruction, (*b*) addressed how to measure cognitive demand of mathematical tasks, (*c*) discussed qualitative fieldwork methods, and (*d*) asked trainees to critique field notes written by one another as practice for classroom observations.

After observing a lesson, observers provided detailed field notes that documented classroom setting, instructional materials, and teacher and student talk through the lesson, and responded to specific prompts about teachers' inquiry-based practice listed in Table 2. Each observer was periodically joined by a project team researcher during classroom observations to ensure the field notes were collected in as similar a manner as possible.

In the months following the completion of classroom observations, we developed rubrics for coding attention to student thinking and justifications using mathematics, and we tested the rubrics by coding observer field notes from our project alongside four experienced educators with master's or doctoral degrees in mathematics education. We met together with these coders four times to discuss and revise ratings, create and modify decision rules for our ratings, and agree on consensus codes for 10 lessons that we coded together. We then assigned coders 15–20 randomly chosen lessons to code each month and met together monthly to share codes for randomly selected lessons to prevent coding drift and further refine our coding document.

Observation measure validity and reliability. Two mathematics educators—from among the four coders—rated 10% of the same lessons; those two educators rated 81% of all the lessons in our sample. The two coders gave the same attention to student thinking rating for 74% of the double-coded lessons and the same justifications using mathematics rating for 79% of the lessons. Teachers' attention to student thinking ratings were highly correlated with their justifications using mathematics ratings ($r(74) = .75$, $p < .001$). Given this high correlation, and that our IBP survey measure combined those activities, we also averaged the two observation ratings to create one measure of observation IBP with five possible values (0, .5, 1, 1.5, and 2).

While we only have six observations for each year to compare to teachers' once-a-year survey report, the intraclass correlation coefficient (ICC) among observation IBP scores within teachers is .53 across both years of our sample (.54 in Year 1 and .46 in Year 2). Landis and Koch (1977) describe an ICC between .4 and .6 as constituting a moderate correlation, and such a correlation provides some evidence that teachers received relatively consistent observation scores for IBP across their lessons.

Survey-observation alignment measure. We first standardized teachers' survey and observation IBP, then calculated survey-observation alignment by subtracting teachers' standardized survey IBP from their standardized observation IBP. Thus, teachers with positive survey-observation alignment scores provided a higher survey IBP self-rating relative to their observation IBP rating. In contrast, teachers with negative survey-observation alignment scores provided a lower survey self-rating relative to their observation IBP rating. An alignment score of zero indicates more alignment in terms of scores for survey and observation IBP that are the same standardized distance from their respective means (unstandardized survey IBP $M = 3.56$ on a 1–5 scale; unstandardized observation IBP $M = .80$ on a 0–2 scale).

Additional survey and observation measures. We included a variety of survey and observation measures in our regression analysis that we hypothesized could be

related with our survey-observation alignment construct. Additional survey measures included teachers' experience and education, attitude toward the main mathematics curriculum (composite measure, $\alpha = .71$), hours of mathematics professional development, and tie strength of teachers' mathematics talk with others.[2] We also used a set of survey items developed by the Study of Instructional Improvement (Hill, Schilling, & Ball, 2004) to measure teachers' Mathematical Knowledge for Teaching (MKT).

Observation measures beyond IBP included in our analysis include (*a*) congruence of teachers' instruction and (*b*) teacher's maintenance of cognitive demand throughout a lesson. Our congruence measure does not reflect the quality of teachers' inquiry-based instruction but rather reflects whether a teacher incorporated the pedagogical methods and activities recommended by the curriculum—however superficially—into their instruction (e.g., group work, manipulatives, use of multiple representations, open-ended questioning strategies). To rate congruence for each lesson, coders used checklists of instructional approaches that we considered to be congruent (e.g., use of multiple representations) or incongruent (e.g., explicit test preparation) with the approach of Everyday Mathematics or Investigations. Following their use of these checklists, coders would assign a holistic yes/no congruence rating to the entire lesson, weighing the number of congruent versus incongruent aspects of instruction noted in the checklist and their perception of teachers' overall instruction. The two mathematics educators who double-coded lessons gave the same congruence rating for 67% of the lessons. This low interrater reliability suggests that any results for congruence should be regarded as tentative.

Our cognitive-demand measure is a simplified version of what we used to rate maintenance of high cognitive demand in other project studies (Stein & Kaufman, 2010; Stein, Kaufman, & Tekkumru-Kisa, 2013). This measure captures the extent to which teachers maintained the cognitive demand of the major mathematical task at each of three phases: within curriculum materials, as set up by the teacher in the classroom, and as enacted by the teacher and students. We considered high-cognitive-demand tasks to be those featuring (*a*) procedures with connection to concepts, meaning, or understanding or (*b*) open-ended tasks without an immediately obvious pathway toward a solution. Low-cognitive-demand tasks featured no mathematical activity, unsystematic or nonproductive exploration, memorization, or procedures without connection to concepts, meaning, or understanding.

Cognitive demand was rated on the following three-point scale for each lesson: three points when (*a*) the major mathematical task received the same high-cognitive-demand rating from materials to enactment or (*b*) the cognitive demand rating increased from low in materials to high in set-up and enactment; two points when (*a*) the cognitive demand of the major mathematical task was maintained at a high level but shifted from one high-level task to another or (*b*) the cognitive demand increased from low in materials to high in either set-up or enactment (but not both); one point when the cognitive demand of the major mathematical task was high at least through set-up but dropped to a low level in enactment; and zero points when cognitive demand of the major mathematical task started low in materials and was low in both set-up and enactment. The two mathematics educators who double-coded lessons in our sample gave the same cognitive-demand score for 83% of the lessons.

Analytic Approach

Our analysis takes a mixed-methods approach (Tashakkori & Teddlie, 1998) that investigates the quantitative variables associated with survey-observation alignment and potential qualitative explanations for alignment.

Quantitative analysis. We first conducted exploratory data analysis to investigate patterns in survey-observation alignment for both districts and then employed regression analysis to explore the survey and observation variables associated with survey-observation alignment. Given the nested nature of our data (i.e., observations nested in teachers), we ran hierarchical linear models (HLM) using the lme4 package (Bates, Maechler, & Bolker, 2011) in R (R Development Core Team, 2011).[3]

Our final HLMs included mixed effects that provided us with the best model fit, based on deviance statistics, and we excluded variables with nonsignificant effects that did not make a difference for model fit. Our full two-level model is listed below, with Y*ij* representing the alignment between teacher *j*'s standardized survey response and standardized observation rating) at occasion *i*.

Level 1 (occasion):

$$\begin{aligned} Y_{ij} = \beta_{oj} &+ \beta_1(\text{Greene District}_j) + \beta_2(\text{grade 3–5 teacher}_j) + \beta_3(\text{MKT score}_j) \\ &+ \beta_4(\text{congruence}_{ij}) + \beta_5(\text{maintenance of high cognitive demand}_{ij}) \\ &+ \beta_6(\text{Greene District} \times \text{congruence}_{ij}) + \beta_7(\text{survey IBP score}_j) \\ &+ r_{ij}, r_{ij} \sim N(0, \sigma^2). \end{aligned}$$

Level 2 (teacher):

$$\beta_{oj} = \gamma_{oo} + \mu_{oj},\ \mu_{oj} \sim N(0, \tau^2).$$

As we also report in our findings, we ran additional HLMs with subsets of the observation ratings and factors within the full survey composite to confirm the relationships between fixed/random effects and survey-observation alignment that we observed in our main HLMs.

Qualitative analysis. To further explore potential explanations for patterns in our quantitative models and additional factors that could be influencing alignment of survey and observation IBP, we conducted qualitative analysis with a subsample of 12 teachers with a difference of more than +/− .9 standard deviations between their standardized survey and standardized observation rating. We chose the .9 threshold because it provided us with examples of at least a few teachers from each district who either had high survey IBP relative to their observation IBP or vice versa. We only drew on interview data from the year(s) that each teacher had a +/− .9 difference between their survey and observation inquiry-based practice rating.

In our qualitative coding of interview data for the 12 teachers, we focused on themes that could be related to survey-observation alignment and the factors in our regressions that were significantly associated with that alignment. Themes that particularly stood out included policy messages about curriculum implementation, professional development and coaching, interaction with others regarding the curriculum lessons, and teachers' feelings of competence. We elaborated on those themes through codes and subcodes that emerged through the analysis (Lincoln & Guba, 1985; Strauss & Corbin, 1998).

Table 3. Means and Standard Deviations for Teacher Survey and Observation Variables

	Region Z		Greene		
	n	*M* (*SD*)	*n*	*M* (*SD*)	*t*-value
Observation variables:					
Standardized observation IBP:					
Year 1	19	−.73 (.71)	21	.78 (.79)	−6.3***
Year 2	17	−.58 (.56)	17	.39 (.99)	−3.5**
Congruence:					
Year 1	19	.45 (.31)	21	.90 (.22)	−5.3***
Year 2	17	.55 (.34)	17	.78 (.30)	−2.1*
Cognitive demand:					
Year 1	19	.90 (.69)	21	2.05 (.64)	−5.4***
Year 2	17	.93 (.79)	17	1.67 (1.10)	−2.3*
Survey-observation alignment:					
Year 1	19	.35 (.79)	21	−.28 (.74)	2.6*
Year 2	17	.21 (1.00)	17	−.22 (.87)	1.6
Survey variables:					
Standardized survey IBP ($\alpha = .88$):					
Year 1	19	−.37 (.85)	21	.50 (.93)	−3.1**
Year 2	17	−.36 (.82)	17	.16 (1.16)	−1.5
Years teaching:					
Year 1	19	7.6 (3.1)	21	5.4 (3.2)	2.2*
Year 2	17	7.9 (2.7)	17	6.8 (2.8)	1.2
Education:[a]					
Year 1	19	4.3 (.7)	21	3.9 (1.4)	1.4
Year 2	17	4.4 (.6)	17	4.4 (1.2)	0.0
Hours of mathematics PD/year:					
Year 1	16	13.8 (19.4)	20	27.7 (24.8)	−1.9
Year 2	15	7.7 (11.6)	5	20.0 (25.3)	−1.5
Tie strength:					
Year 1	19	5.8 (4.5)	20	7.4 (3.4)	−1.2
Year 2	13	3.4 (1.7)	14	5.7 (5.6)	−1.5
Positive attitude toward curriculum ($\alpha = .71$):					
Year 1	17	3.1 (.8)	19	3.5 (.7)	−1.7
Year 2	16	3.1 (.9)	17	3.5 (.6)	−1.7
Mathematical Knowledge for Teaching:					
Year 1	18	6.3 (2.5)	21	7.0 (2.7)	−.8
Year 2	17	5.7 (2.6)	16	5.7 (2.9)	0.0

[a] Education was calculated by giving teachers one point each for degree, credits beyond degree, and certificate in ESL, bilingual, math or special education.

*$p < .05$ in independent *t*-tests comparing teachers in Region Z and Greene.

**$p < .01$.

***$p < .001$.

Results

Survey-Observation IBP and Alignment in the Two Districts

Table 3 includes means and standard deviations for teachers' survey and observation IBP scores, survey-observation alignment, and all other survey and observation variables used in our analysis. In order to assess the generalizability of our 47-teacher subsample, we compared survey responses of teachers in that subsample to the over 700 teachers in each district who were surveyed in Y1 and Y2. Independent *t*-tests indicated that our 47-teacher subsample did not differ significantly from the larger sample of all teachers regarding survey IBP and most other survey variables, includ-

ing education, professional development hours, and tie strength in either year of our study. That said, our 47 teachers had higher scores than all teachers who completed the survey for MKT in Y1 ($t(796) = 2.88$, $p = .004$). These minimal differences suggest our findings are generalizable to all teachers in those districts.

Compared to their counterparts in Region Z, Greene teachers had much higher observation mean ratings for inquiry-based practice (IBP), as well as much higher congruence and cognitive-demand mean ratings. While we found large by-district differences in observation IBP, by-district differences in survey IBP were milder, with Greene teachers having significantly higher survey IBP compared in Region Z teachers in Y1 but not Y2. Given the centrality of survey and observation IBP to our analysis, we also conducted skewness tests for the two variables, which indicated that neither variable was significantly skewed (survey IBP skewness = −.33, $SE = .28$; observation IBP skewness = .23, $SE = .12$).

Region Z teachers also had more positive survey-observation alignment, meaning that their survey IBP rating tended to be higher than their observation IBP rating. Greene teachers, on the other hand, had more negative survey-observation alignment, meaning that their survey IBP rating tended to be lower than their observation IBP rating. Greene teachers also had significantly fewer years of teaching experience in Y1 but not Y2 compared to Region Z teachers.

We also calculated Spearman's rho for the correlation between survey and observation IBP for each year across all teachers and for Region Z versus Greene teachers. The correlation between survey and observation IBP was quite significant among all teachers in each year (Y1, $r(40) = .67$, $p < .001$; Y2, $r(34) = .60$, $p < .001$). However, for Region Z teachers, the correlation between survey and observation IBP was high and significant only in the first year (Y1, $r(19) = .55$, $p = .014$; Y2, $r(17) = .11$, $p = .668$).[4] For Greene teachers, the correlation was high and significant for both years (Y1, $r(21) = .52$, $p = .016$; Y2, $r(17) = .69$, $p = .002$).

Factors Associated with Survey-Observation Alignment

In our HLMs, we tested multiple independent survey and observation variables to ascertain their effect on survey-observation alignment. We omitted any variables from our models that did not have a significant association on survey-observation alignment, including teacher education, years of teaching experience, years of teaching experience with Everyday Mathematics (for Region Z teachers) and Investigations (for Greene teachers), year of study (Y1 or Y2), attitude toward the curriculum, professional development hours, tie strength regarding mathematics talk, and several survey measures regarding the strength of teachers' social networks. We also omitted random effects for school, classroom observer, and coder, as each random effect predicted little to none of the unexplained variation in the model.

To include both survey and observation IBP scores as independent variables in our models would produce a tautological model that left no variation in the dependent variable for other predictors to explain. However, we included survey IBP score as an independent variable to better understand its role in survey-observation alignment; we included survey IBP rather than observation IBP given the multicollinearity between observation IBP and the other observation variables in our full model.

Our final HLMs are presented in Table 4. We kept the same number of cases across our models so that AIC (Akaike information criterion) could be used to assess model fit. The variables in the full model explained 53% of the variation in survey-observation agreement, following Snijders and Bosker's (1994, 1999) method for calculating R^2 in two-level models. The residuals in the final model displayed no apparent pattern, both on their own in a scatterplot and plotted against any other variables included in the models, suggesting that the final models represented a good fit with the data. Because our HLMs suggested mainly teacher-level factors predicting survey-observation alignment, and because a random parameter for observer or coder explained almost no variation in our models, we refer to teachers with higher survey IBP compared to observation IBP as "overestimators" and teachers with higher observation IBP compared to survey IBP as "underestimators."

Two variables had a significant negative relationship with survey-observation alignment in our final model: (*a*) being a congruent teacher from Greene ($p < .001$), and (*b*) maintaining the cognitive demand of mathematical tasks ($p < .001$). Thus, teachers with higher congruence and higher cognitive-demand ratings were more likely to be underestimators of their IBP.

In contrast, MKT ($p < .05$), being a third- through fifth-grade teacher ($p < .10$), and survey IBP ($p < .001$) all had a significant negative relationship with survey-observation alignment. Thus, those with higher MKT, those who taught third through fifth grade, and those with higher survey IBP were more likely to be overestimators of their IBP.

Teacher-level variance explained almost as much as the residual unexplained variance in the first model. However, by the final model, teacher-level variance decreased considerably, especially with the addition of survey IBP, which explained much of the variance within teachers' observations.

We engaged in three additional tests to confirm our models. First, we examined differences for models when we only included data from three lessons observed in the fall versus three lessons observed in the spring of both Y1 and Y2. Second, we compared differences for models when we calculated survey-observation alignment using only survey items that were not reverse scored for one model and only survey items that were reverse scored in another model. Third, we examined differences for models when we calculated survey-observation alignment using only survey IBP items with the never-always response scale (and not the agree/disagree scale) for another model.

The Appendix includes results for all tested models. Each model yielded very similar effects; all coefficients ran in the same general direction as in the full model.Teachers' MKT score and grade level became less strong or nonsignificant predictors of survey-observation alignment in some models. Additionally, the interaction between congruence and being a Greene teacher was not significant when using the spring observation data only in the survey-observation alignment calculation. The other coefficients were similarly significant across tested models compared to the full model.

Potential Explanations for Factors Associated with Survey-Observation Alignment

Our qualitative analyses included interview data for 12 teachers with the largest differences between standardized survey and observation rating. We included teachers in these analyses if their survey-observation difference was greater than .90 or less

Table 4. Fixed-Effect Estimates (Top) and Variance-Covariance Estimates (Bottom) for Hierarchical Linear Models of the Predictors of Survey-Observation Alignment

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Fixed effects:							
Intercept	.16 (.15)	−.18 (.17)	−.79 (.21)***	−.54 (.22)*	−.48 (.22)*	−.60 (.22)**	.22 (.16)
Greene district	−.27 (.21)	−.24 (.19)	−.29 (.17)+	−.14 (.18)	−.02 (.19)	.55 (.25)*	.16 (.18)
Grades 3–5		.63 (.18)**	.39 (.18)*	.35 (.18)*	.29 (.18)	.25 (.18)	.20 (.11)+
MKT score			.12 (.03)***	.12 (.03)***	.13 (.03)***	.14 (.03)***	.05 (.02)*
Congruence				−.46 (.11)***	−.23 (.11)*	.01 (.13)	−.05 (.12)
Cognitive demand					−.22 (.04)***	−.21 (.04)***	−.24 (.03)***
Greene × congruence						−.75 (.21)***	−.73 (.19)***
Standardized survey IBP							.68 (.06)***
Random parameters:							
Intercept for teacher (γ_{00})	.46 (46)	.35 (39)	.28 (34)	.30 (38)	.33 (42)	.34 (44)	.06 (13)
Residual	.55 (54)	.55 (61)	.53 (66)	.50 (62)	.46 (58)	.44 (56)	.41 (87)
Decline in variance from empty model (%)	1	12	20	20	22	23	53
AIC	959	952	942	928	904	895	824
Cases	383	383	383	383	383	383	383
Teachers	47	47	47	47	47	47	47

Note.—Standard errors are in parentheses for fixed effects. Percentage of total variance explained is in parentheses for random parameters.

+ $p < .10$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

Table 5. Teachers with Survey-Observation Alignment Scores of +/− .90

	Region Z		Greene	
	A < −.90 Obs. > Survey	A > .90 Survey > Obs.	A < −.90 Obs. > Survey	A > .90 Survey > Obs.
N	2	3	5	2
Grade	1.5	3.3	1.4	4
Observation IBP	Raw = .6	Raw = .5	Raw = 1.0	Raw = .8
	Z = −.3	*Z* = −.5	*Z* = .3	*Z* = −.1
Survey IBP	Raw = 2.6	Raw = 4.0	Raw = 3.0	Raw = 4.2
	Z = −1.8	*Z* = .8	*Z* = −1.1	*Z* = 1.2
Congruence	.6	.8	.9	.8
Cognitive demand	1	1	1.7	1.5
MKT score	2.3	9	4.1	9.5

than −.90. As indicated in Table 5, the summary statistics for these teachers followed some of the trends that we located in our HLMs: overestimators tended to have higher MKT scores. Also, as suggested by our HLMs, no clear pattern emerged regarding congruence for Region Z teachers, whereas Greene teachers who underestimated their instruction had slightly higher congruence than Greene teachers who overestimated their instruction. Although the HLMs indicated that overestimators tended to have lower cognitive demand than underestimators, that pattern was not as strong among the extreme over- and underestimators in our qualitative sample. Interestingly, the observation IBP rating for teachers with greater than a +/− .90 survey-observation difference varied little within districts. Teachers' survey rating thus appeared to drive the large differences between survey and observation among these teachers.

By-district trends in policy messages and curricular support. What teachers shared regarding district policy messages and curricular support was more related to the district where they taught than their survey-observation alignment. Region Z teachers reported few to no policy messages about the use of Everyday Mathematics (EM) and almost no professional development and coaching regarding implementation of EM. Of the five Region Z teachers in our subsample, four noted only the district policy message that they should "use" EM; one teacher reported no policy messages regarding EM. One Region Z teacher provides a representative example: "We don't really hear from the district. But I think the curriculum is Everyday Math and make sure you do it every day." Most Region Z teachers also reported little support for their teaching of EM. Four reported meeting with a coach once every few months, although one of the five teachers noted attending one daylong EM workshop.

In contrast, almost all of the seven Greene teachers in our subsample reported receiving the district policy message that they should implement the inquiry-based strategies within their Investigations curriculum. Five of the seven teachers received intensive once-a-week coaching in Y1 or leading up to Y1. Five Greene teachers attended at least one weeklong training, and sometimes two weeklong trainings, on the use of Investigations. Unlike in Region Z, Greene teachers also reported consistent talk about Investigations at grade-level meetings and professional development. That said, several teachers complained about the absence of district policy messages about Investigations in Y2 of our study compared to Y1.

Little or anticurriculum teacher interaction among overestimators. Across both districts, overestimators reported little to no positive interaction with other teachers regarding use of their mathematics curriculum. The three Region Z overestimators reported little interaction with others in their school about engaging in inquiry-based practice with EM. One of those overestimators, who taught children with special needs, remarked that "it's very, very difficult to kind of plan for math with the other second grade teachers." When asked if she talked with others about mathematics lessons, she responded, "Really, we just don't have that opportunity unfortunately." The two other overestimators in Region Z only reported speaking to other teachers about how to modify EM to make it easier for students. One teacher explained, "[EM] goes [into] tenths, like the eighth of an inch or the sixteenth of an inch . . . whereas [for the state test] they need to know down to the half an inch or quarter of an inch . . . so we [teachers] said maybe [EM] is pushing them too hard for their actual level." Another Region Z teacher said, "We have made modifications where we start a [EM] lesson and we say, 'Oh, there is no way they can do this.' We have stopped it and done something totally different."

One overestimator in Greene also reported little interaction with peers regarding the teaching of Investigations: "I mean, we [teachers] used to talk more, but we don't really . . . we don't all buy into [Investigations] the same. So, it's kind of hard to . . . share with someone who doesn't really do [Investigations] as much as I do." In contrast, the other overestimator in Greene reported frequent interaction with his coaches in regard to his inquiry-based practices, although he reported much less substantial interaction with his grade-level team about those practices.

Frequent talk about curriculum and struggle to use the curriculum among underestimators. In contrast with overestimators, underestimators in both districts reported more frequent, positive interaction with others regarding curricula and/or feelings of uncertainty or struggle to use their curriculum well. In Region Z, one of the two underestimators reported frequent discussion with another teacher at her school and grade level who she perceived to be especially good at using EM: "Lane's so good . . . I mainly just ask her, 'How did you do this?' or 'What did you do this time?' She says to me, 'I'll tell you what I did. I did this. I did this. I had them doing this.'. . . She'll have all the [EM manipulatives] out and she'll make it very interactive." The other Region Z underestimator reported little teacher talk about EM. But, unlike her overestimating peers, she felt unsure about her instruction and wished she could improve: "I taught kindergarten for 18 years, so there are some things [with EM] that I'm not comfortable with as I should be."

All but one of the underestimators in Greene reported intensive interaction with other teachers and their coach in the current year or the prior year in order to implement Investigations well. One kindergarten teacher, for example, commented, "My coach, Ms. Zevon, I talk to all the time . . . she really helps me reflect on what I'm doing and what I could do differently . . . then, Mrs. Horace, who is the other math coach, like I said, I taught with her, so I feel comfortable enough to go ask her questions." Another first-grade teacher at Greene noted, "Last year, I was always, 'Hey I did my lessons this way. What did you do? Did that work?' We did that a lot last year. So this year, we don't do as much as we did last year . . . but yet, of course, we are always collaborating. Always, always, always."

The one underestimator who reported no intensive interaction was new to Greene in Y2. She had not attended a weeklong workshop on Investigations, as did the other

underestimators, and she reported struggling to implement the curriculum well: "It's really challenging sometimes because I am used to the traditional way of, you know, standing in front of the classroom and [saying] 'This is what you are going to learn' . . . that's been really challenging for me to allow the kids to have that freedom to explore math or whatever on their own."

Discussion and Conclusions

For this article, we investigate alignment between survey self-reports and observation ratings among teachers in two urban districts regarding one key element of standards-based mathematics instruction: inquiry-based practices. We engaged in mixed-methods analysis to consider the factors associated with survey-observation alignment and some explanations for those associations.

Review of Quantitative Results

Our quantitative analysis pinpointed district and teacher-level factors significantly related to teachers' survey-observation alignment. According to that analysis, underestimators tended to be teachers who also received high cognitive-demand scores for their lessons. Furthermore, underestimators also tended to be those teachers who were congruent with general pedagogical tenets of their curriculum, but only if they were from Greene school district. Congruence in Region Z did not matter for survey-observation alignment. These findings suggest that teachers who provided higher-quality standards-based instruction were also more measured in their assessment of how well they were implementing inquiry-based practice. It is possible that the more these teachers learned about inquiry-based practice, the more that they were able to recognize and thus acknowledge what they did not know.

Conversely, third- through fifth-grade teachers and those with higher MKT tended to overestimate the extent of their inquiry-based practice. While difficult to interpret, these results could imply that teachers with higher MKT overestimated their IBP because they had higher feelings of competence regarding their mathematics instruction in general, which is reinforced by some of our qualitative findings.

The tests of our main models used versions of our survey-observation alignment variable that were calculated with fall versus spring observation data or subsets of the survey items. The coefficients in these alternative models were very similar to those in the full models; the small differences across these models are likely attributable to use of only half the observation data in some models and use of shorter survey scales in others. These tests thus affirm use of the entire 17-item survey scale and all the observation data in our main models.

Review of Qualitative Results

In our qualitative analysis of a subset of teachers with particularly large gaps between their survey and observation IBP, those who overestimated their IBP reported not receiving or seeking out interaction with others to improve their inquiry-based practices. On the other hand, those who underestimated their IBP discussed much more intensive interaction with others to improve their inquiry-based practices, and they also communicated less confidence and certainty about their ability to engage in that instruction.

By-district trends could be further influencing interaction with others, in that Greene teachers in our subsample more often reported regular opportunities to talk about Investigations with their coach and peers, whereas Region Z teachers often reported little interaction and little or no policy messages encouraging that interaction. These by-district differences in social interaction and support are also discussed in Coburn and Russell (2008) and Stein and Coburn (2008).

Our qualitative data suggest that teachers' survey estimates of their IBP could be a function of their exposure to strong models of inquiry-based practice by which to judge their own practice accurately. Comments from overestimators additionally suggest that testing may have played some role in the negative or low interaction Region Z teachers had with others about their inquiry-based practices. This finding might also explain significant effects of being a third- through fifth-grade teacher on overestimation, as those are the tested grades in both districts.

Whether aspects of the standards-based curricula used in Greene and Region Z could be influencing survey-observation alignment is beyond the bounds of this article. However, some of our other research has suggested that the more open-ended nature of Investigations tasks may be closely aligned with IBP (Stein & Kaufman, 2010; Stein & Kim, 2009). For this reason, Greene teachers using Investigations may have been better equipped to understand those practices, and thus overestimate them less often. These curriculum differences might also explain why congruence to curriculum predicted underestimation among Greene teachers but not Region Z teachers. But these hypotheses would have to be confirmed through further research.

Limitations

We should note three main limitations for this work. First, observer biases could play a role in survey-observation alignment that we could not detect in our analysis. For example, the overall differences in curriculum and instruction in Greene compared to Region Z could have encouraged an upward bias across all IBP observation ratings for Greene teachers. Second, while our factor analysis yielded an inquiry-based practice composite that was significantly correlated with observation IBP across teachers, some of the factor loadings were lower than what is typically judged as satisfactory (see, e.g., recommendations from Costello & Osborne, 2005).

Third, our study is an exploratory, observational study focused on a single aspect of mathematics instruction among a small sample of teachers in only two district settings. As such, we can make no causal claims based on our findings. Much more research across varied contexts and on various aspects of instruction is necessary to understand more about survey-observation alignment of instruction and the factors that might be influencing survey-observation alignment. Furthermore, our small sample size of 47 teachers, together with the large number of variables in the full model, could lead to inflated Type I error rates, as well as overfitting. Babyak (2004), for example, suggests 10–15 cases per predictor to provide good estimates. However, the similarities among coefficients in models using fewer versus more predictors in our HLMs reported in Table 4 are some reassurance of our results.

Implications for Research and School District Policy

Our study has some implications for research and school district policy. First, our study suggests that research relying only on survey self-report to measure mathematics instruction could draw faulty conclusions about that instruction and its relationship to other variables. Researchers with limited capacity to gather observational data from teachers on their instruction might consider observing a subset of teachers on measures closely aligned with a survey in order to gather more evidence about that survey's validity. Researchers might also consider incorporating measures of teacher knowledge and interaction into their survey, as well as other factors that could be associated with the accuracy of teachers' self-reports. Such measures could then be included in regressions to adjust for survey self-report biases.

This study also implies the need for much better survey methods to understand teacher practice across varying contexts, given that surveys remain the most cost-effective way to measure what teachers do. Some researchers have explored the use of hypothetical situations for better measuring teachers' reform practices (Ruiz-Primo & Li, 2002; Stecher et al., 2006). Researchers in political science and health fields have also explored the use of "anchoring vignettes" in surveys to calibrate the use of response scales among respondents who hold differing understandings about the extent of their perceptions or beliefs (Kapteyn, Smith, & Van Soest, 2007; King, Murray, Salomon, & Tandon, 2004). Huntley (2012) has suggested similar "vignettes" to measure variation in teachers' implementation of mathematics textbooks, with both observers and teachers using the vignettes to report on quality and quantity of textbook use. This research, or a search for similar ways to calibrate teachers' responses, holds promise for gathering more accurate survey data from teachers, especially given the suggestion in our research that teachers gauge themselves in relation to their peers when using survey response scales. More accurate survey methods for measuring teachers' instruction could enable researchers to discern instructional patterns and associations among instruction, student learning, and other variables over a much larger number of teachers.

Lastly, our work indicates that teachers can have perceptions about the extent of their inquiry-based practices that differ considerably from their actual practices. That gap between perceptions and practice can have implications for the extent to which teachers understand the shortcomings of their instruction and seek to improve it. For this reason, school districts might consider how to monitor gaps between teachers' perceptions and practices carefully and intervene when teachers overestimate the quality of their instruction to a great extent in comparison with observers.

This study specifically points to positive teacher interaction around curriculum implementation as a potential tool for helping teachers better understand their inquiry-based practices and preventing them from overestimating the extent of those practices. Districts seeking to improve the quality of teachers' instruction might consider integrating structures by which teachers with expertise in particular practices consistently interact with other teachers and serve as models for instruction. By cultivating more accurate teacher perceptions, leaders could help teachers to understand and acknowledge any shortcomings in their

knowledge and instruction, which could serve as a motivation for instructional improvement.

Appendix

Table A1 below includes HLM coefficients on survey-observation alignment for Model 7 that was reported in Table 4 of our article, alongside models with the same independent variables where alignment was calculated with a subset of the observation data or survey data. Models 1A and 1B use observation data from the fall versus spring semester. IBP Set 2A and 2B use reverse-coded versus non-reverse-coded survey items. IBP Set 3 uses survey items with the never-always response scale only. We did not include a model on survey-observation alignment with survey items using the agree-disagree response scale since only two items used that scale. We include AIC for each model as a measure of relative fit but should note that the AIC is only comparable for the full model and models 2A, 2B, and 3 that use the same number of cases.

Notes

This work was supported by grants from the National Science Foundation (IERI grant REC-0228343) and the Institute of Education Sciences, U.S. Department of Education (grant R305B1000012). All opinions and conclusions in this article are those of the authors and do not necessarily reflect the views of the funding agencies. The authors wish to thank Marc Chun, Teresa McCaffrey, Rebecca McGraw, Chris Nelson, Laurie Rubel, Marcia Seeley, Jaime Smith, Sarah Spencer, Stephanie Sutherland, Mikyung Wolf, and Bahadir Yanik for help with data collection. We would also like to thank Amy Hillen, LuAnn Malik, Andrea Miller, Marsha Seeley, Jaime Smith, and Stephanie Sutherland for help with data analysis. We are also very grateful to Howard Seltman for his invaluable advice regarding our quantitative analysis and Cynthia Coburn for her input on all aspects of the Scaling Up Mathematics project. Finally, we would like to express our sincere thanks to all the participants in this study for welcoming us into their schools and offices and allowing us to interview them and observe their instruction. Julia Heath Kaufman is a policy researcher at the RAND Corporation in Pittsburgh, PA. Mary Kay Stein is professor of Learning Sciences and Policy, University of Pittsburgh. Brian Junker is professor in the Department of Statistics, Carnegie Mellon University.

1. All authors have no financial interests in either curriculum used in this study. Furthermore, the authors did not have any relationship with either of the two district study sites prior to the present study.

2. Tie strength is a common measure used in social science research to take into account how often a person speaks to others, how many people with whom that person speaks, and how "close" the person feels with those to whom s/he speaks (Burt, 1992; Marsden & Campbell, 1984). The formula used to calculate tie strength is: [frequency of talk] × [closeness] / [number of people with whom s/he talked].

3. Correnti and Rowan (2007) and Rowan and Correnti (2009) are examples of other studies that nest classroom lesson scores within teachers in mixed models.

4. Although the Region Z correlations in Years 1 and 2 appear to contrast sharply, our analysis suggests that a single teacher with a very low survey IBP rating and an average observation rating minimized the Year 2 correlation considerably. If that teacher was removed from the analysis, Spearman's rho would still be nonsignificant at the $p < .05$ level ($r(16) = .45$, $p = 08$) but higher than the correlation in Table 4 ($r(17) = .22$, $p = .39$).

Table A1. Fixed-Effect Estimates (Top) and Variance-Covariance Estimates (Bottom) for Hierarchical Linear Models of the Predictors of Survey-Observation Alignment (with Survey-Observation Alignment Calculated Using Subsets of Observation and Survey Data)

Parameter	Full Model	1A: Obs. Fall Only	1B: Obs. Spr. Only	2A: Survey NoRC Items	2B: Survey RC Items	3: Survey Frequency Items
Fixed effect:						
Intercept	.22 (.16)	.29 (19)	.23 (.25)	.40 (.17)*	.35 (.16)*	.23 (.16)
Greene district	.16 (.18)	.24 (.24)	.06 (.26)	.03 (.19)	.16 (.19)	.17 (.18)
Grades 3–5	.20 (.11)+	.25 (.13)*	.20 (.16)	.19 (.12)	.16 (.12)	.19 (.11)+
MKT score	.05 (.02)*	.03 (.02)	.07 (.03)+	.03 (.02)	.04 (.02)+	.05 (.02)*
Congruence	−.05 (.12)	.12 (.16)	−.25 (.17)	−.06 (.12)	−.07 (.12)	−.06 (.11)
Cognitive demand	−.24 (.03)***	−.23 (.05)***	−.27 (.05)***	−.25 (.04)***	−.25 (.03)***	−.24 (.03)***
Greene × congruence	−.73(.19)***	−.98 (.26)***	−.41 (.28)	−.72 (.20)***	−.79 (.19)***	−.74 (.19)***
Survey IBP[a]	.68 (.06)***	.71 (.07)***	.67 (.09)***	.80 (.06)***	.74 (.06)***	.68 (.06)***
	Full Model	Model 1A	Model 1B	Model 2A	Model 2B	Model 3
Random parameters:						
Intercept for teacher (γ_{00})	.06 (13)	.07 (17)	.14 (26)	.10 (20)	.08 (16)	.06 (13)
Residual	.41 (87)	.34 (83)	.40 (74)	.41 (80)	.41 (84)	.41 (87)
Decline in variance from empty model (%)	53	59	49	57	58	53
AIC	824	402	446	839	831	823
Cases	383	190	193	383	383	383
Teachers	47	47	46	47	47	47

Note.—Standard errors are in parentheses for fixed effects. Percentage of total variance explained is in parentheses for random parameters. The fixed effect for standardized survey IBP for a particular model is what was used in the survey-observation alignment calculation for that model. For example, standardized survey IBP only included the items using the agree-disagree response scale for Model 2A.

+ $p < .10$.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

References

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools* (NCEE 2009-4052), Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting regression-type models. *Psychosomatic Medicine*, **66**, 411–421.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is: Or might be: The role of curriculum materials in teacher learning and instructional reform? *Educational Researcher*, **25**(9), 6–14.

Bates, D., Maechler, M., & Bolker, B. (2011). *Lme4: Linear mixed-effects models using s4 classes*: Available at http://cran.r-project.org/web/packages/lme4/index.html.

Burstein, L., McDonnell, L. M., van Winkle, J., Ormseth, T., Mirocha, J., & Guitton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Coburn, C., & Russell, J. L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis*, **30**(3), 203–235.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, **12**(3), 311–329.

Cohen, D., & Hill, H. (2001). *Learning policy: When state education reform works*. New Haven, CT: Yale University Press.

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal*, **44**(2), 298–338.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, **10**(7), 1–9.

Engle, R. A., & Conant, F. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, **20**(4), 399–483.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, **19**(4), 325–338.

Hamm, J. V., & Perry, M. (2002). Learning mathematics in first-grade classrooms: On whose authority? *Journal of Educational Psychology*, **94**(1), 126–137.

Hightower, A. M., Knapp, M. S., Marsh, J. A., & McLaughlin, M. W. (2002). *School districts and instructional renewal*. New York: Teachers College Press.

Hill, H. C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, **19**(3), 447–475.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, **105**(1), 11–30.

Huntley, M. A. (2012). Using concerns-based adoption model theory to develop tools to examine variations in mathematics textbook implementation. In D. J. Heck, K. B. Chval, I. R. Weiss, & S. W. Ziebarth (Eds.), *Approaches to studying the enacted mathematics curriculum* (pp. 47–66). Charlotte, NC: Information Age.

Kapteyn, A., Smith, J. P., & Van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, **97**(1), 461–473.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, **98**(1), 191–207.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal*, **27**(1), 29–63.

Landis, J. R., & Koch, G. G. (1977). The measurement of agreement for categorical data. *Biometrics*, **33**, 159–174.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Thousand Oaks, CA: Sage.

Marsden, P., & Campbell, K. (1984). Measuring tie strength. *Social Forces*, **63**(2), 482–501.

Marsh, J. A. (2000). *Connecting districts to the policy dialogue: A review of literature on the relationship of districts with states, schools, and communities*. Stanford, CA: University of Washington Center for the Study of Teaching and Policy.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, **21**(1), 29–45.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, **44**(4), 921–958.

Peterson, P. L. (1990). Doing more in the same amount of time: Cathy Swift. *Educational Evaluation and Policy Analysis*, **12**(3), 261–280.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org.

Ross, J. A., McDougall, D., & Hogaboam-Gray, A. (2003). A survey measuring elementary teachers' implementation of standards-based mathematics teaching. *Journal for Research in Mathematics Education*, **34**(4), 344–363.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher*, **38**(2), 120–131.

Ruiz-Primo, M. A., & Li, M. (2002). *Vignettes as an alternative teacher evaluation instrument: An exploratory study*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Schoenfeld, A. S. (1998). Toward a theory of teaching-in-context. *Issues in Education*, **4**(1), 1–95.

Shifter, D. (2001). Learning to see the invisible: What skills and knowledge are needed to engage with students' mathematical ideas? In T. Wood, B. S. Nelson, & J. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 109–134). Mahwah, NJ: Erlbaum.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods and Research*, **22**, 342–363.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies*, **31**(2), 143–175.

Spillane, J., Diamond, J. B., Burch, P., Hallett, T., Jita, L., & Zoltners, J. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy*, **16**(5), 731–762.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis*, **21**(1), 1–27.

Stecher, B., Le, V.-N., Hamilton, L., Ryan, G., Robyn, A., & Lockwood, J. R. (2006). Using structured classroom vignettes to measure instructional practices in mathematics. *Educational Evaluation and Policy Analysis*, **28**(2), 101–130.

Stein, M. K., & Coburn, C. E. (2008). Architectures for learning: A comparative analysis of two urban school districts. *American Journal of Education*, **114**(4), 583–626.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning*, **10**, 313–340.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, **33**(2), 455–488.

Stein, M. K., & Kaufman, J. H. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal*, **47**(3), 663–693.

Stein, M. K., Kaufman, J. H., & Tekkumru Kisa, M. (2013). Mathematics teacher development in the context of district managed curriculum. In Y. P. Li (Ed.), *Mathematics curriculum in school education* (pp. 351–376). New York: Springer.

Stein, M. K., & Kim, G. Y. (2009). The role of mathematics curriculum materials in large-scale urban reform: An analysis of demands and opportunities for teacher learning. In J. T. Remillard, G. Lloyd, & B. Herbel-Eisenmann (Eds.), *Mathematics teachers at work: Connecting curriculum materials and classroom instruction* (pp. 37–55). New York: Routledge.

Stein, M. K., Remillard, J. T., & Smith, M. S. (2007). How curriculum influences student learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 319–369). Greenwich, CT: Information Age.

Strauss, A., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. London: Sage.

Supovitz, J. A. (2006). *The case for district-based reform: Leading, building, and sustaining school improvement*. Cambridge, MA: Harvard Educational Publishing Group.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (6th ed.). New York: Pearson.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Weiss, I. R., Banilower, E. R., McMahon, K. C., & Smith, P. S. (2001). *Report of the 2000 National Survey of Science and Mathematics Education*. Chapel Hill, NC: Horizon Research. Available at: http://2000survey.horizonresearch.com/reports/status/complete.pdf.

Wilson, S. M. (1990). Conflict of interests: The case of Mark Black. *Educational Evaluation and Policy Analysis*, **12**(3), 293–310.