Article

Educational and Psychological
Measurement
2018, Vol. 78(4) 635–652

DOI: 10.1177/0013164417714846
journals.sagepub.com/home/epm
SAGE

What About the "Instruction" in Instructional Sensitivity? Raising a Validity Issue in Research on Instructional Sensitivity

Marsha Ing[1]

Abstract

In instructional sensitivity research, it is important to evaluate the validity argument about the extent to which student performance on the assessment can be used to infer differences in instructional experiences. This study examines whether three different measures of mathematics instruction consistently identify mathematics assessments as being sensitive to instruction. Mixed findings across fourth-grade (n = 8,298) and fifth-grade (n = 9,336) students and their teachers across three school districts raise questions as to whether different ways of measuring instruction provide similar inferences about the instructional sensitivity of assessments. This raises validity concerns about the quality of inferences based on different measures of instruction.

Keywords

instructional sensitivity, mathematics, achievement, instruction

Research on instructional sensitivity focuses on the extent to which assessments are more or less sensitive to the effects of instruction (Burstein, 1983, 1989; D'Agostino, Welsh, & Corson, 2007; Polikoff, 2010; Popham, 2007; Ruiz-Primo et al., 2012; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). Instructional sensitivity refers to the degree to which the item ". . . reflects student knowledge/ability as the consequence of instruction'' (Burstein, 1989, p. 99). Burstein (1983) emphasized that

[1] University of California–Riverside, Riverside, CA, USA

Corresponding Author:
Marsha Ing, University of California–Riverside, 1207 Sproul Hall, Riverside, CA 92521, USA.
Email: marsha.ing@ucr.edu

> An exact explanation of how a student responds to given test items is unanswerable under all but the most trivial circumstances. Nonetheless, it is reasonable to attempt to narrow the range of plausible explanations and to investigate the likelihood that particular instructional experiences activate cognitive processes that account for student responses. (p. 99)

Instructional sensitivity research requires validity evidence to support claims that an assessment is sensitive to detect differences in instructional opportunities. The Standards for Educational and Psychological Testing defines validity as ''the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests'' (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). Validity evidence is an argument-based approach that allows for the credibility of the claims to be examined from different sources (see, e.g., Messick, 1989). While some validity research focuses on the credibility of the interpretations, Kane (2013a) advocates for an interpretation/use argument that ''includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use)'' and ''focuses on a particular use'' or ''range of possible uses'' (p. 2). Kane (2013b) indicates that the interpretation/use arguments will vary depending on the situation and that there is not a particular ''algorithm for validation.'' Others have endorsed Kane's view of validity by pushing for claims to be ''explicitly specified and validated'' (Brennan, 2013, p. 79) and that arguments should involve expertise from different disciplines (Haertel, 2013).

Two potential interpretation/use arguments of instructionally sensitive assessments are (1) students with higher scores received instructional opportunities and (2) this type of assessment can be used for identifying higher quality instructional opportunities. These interpretative arguments can be evaluated using different strategies. The simplest one is comparing student performance between students who have received instruction and students who have not. As evidence, we would expect that students with higher performance received instructional opportunities. However, suppose that students who received instruction performed lower than students who did not receive instruction. Such an unexpected result would suggest that the inference about this assessment as an indicator of instructional opportunities is not appropriate. Another strategy to gather evidence about these interpretations is to measure the quality of instruction. However, since there is no single way to define high-quality instruction, different conclusions could be drawn about the quality of instruction depending on how instruction is measured (Correnti & Martinez, 2012). Thus, whether an assessment is identified as sensitive to instruction might be highly dependent on the way in which instruction is defined and measured. If instruction is defined and measured in a certain way, an assessment might not be identified as instructionally sensitive; but if defined and measured in a different way, the same assessment may be identified as instructionally sensitive (Grossman, Cohen, Ronfeldt, & Brown, 2014). Such conflicting evidence when identifying instructionally sensitive assessments raises concerns about the quality of inferences of student performance.

This study explores issues involved in identifying instructionally sensitive assessments when using different instruments to measure quality of instruction. More specifically, this study explores how three different strategies to measure instruction relate to student performance on two different measures of elementary mathematics achievement. The guiding research question of the study was *Do different measures of instruction lead to similar conclusions about the instructional sensitivity of assessments?* Since measures of instruction can be used as evidence to support claims about the extent to which an assessment is sensitive to detect differences in instructional opportunities, this type of study is critical. The study then contributes to the literature on instructional sensitivity by exploring the importance of the type of measure of instruction in identifying assessments that are sensitive to instruction.

Measuring Instruction in Studies of Instructional Sensitivity

Previous studies of instructional sensitivity include a range of different measures of instruction. Instruction in instructional sensitivity research is defined as the ‘‘content exposure/opportunity *and* the ways in which students have been taught (subject) matter’’ (Burstein, 1989, p. 7). Both *what* is taught and *how* it is taught provide different but related information about the quality of instructional opportunities. For example, teachers may cover the same content standards but have different approaches to implementing the standards. Some teachers might be more effective at implementing the content standards compared with other teachers. Such differences between teachers would not be evident if instruction is measured in terms of *what* was covered. A different but related piece of information is obtained by focusing only on *how* instruction is implemented. Teachers might have similar ways of engaging students, but one teacher might spend a large portion of time covering only a particular content standard while another teacher might cover a range of content standards. With no agreement on the best way to measure instruction, studies of instructional sensitivity have suggested the inclusion of measures of instruction in terms of either *what* or *how* students were taught (Burstein, 1989).

What Students Were Taught. Earlier approaches to measuring instructional sensitivity drew inferences about *what* instruction students received by examining unusual item-response patterns on assessments (e.g., Hanna & Bennett, 1984; Harnisch, 1983; Linn, 1983) or comparing the pre- and posttest performance of groups of students who received instruction with groups of students who did not receive instruction (e.g., Cox & Vargas, 1966; Haladyna & Roid, 1981; Popham, 1971). These earlier approaches did not incorporate information about the instruction students actually received. Instead, unusual response patterns were identified for profiles of responses with the same total score but different responses to particular items. For example, a large value of Sato’s (1975) ‘‘caution index’’ indicated an unusual response pattern that served to caution against the use of the total score as an accurate measure for a particular examinee. An examinee who answered 8 out of 10 items correctly was expected to answer all the ''easy'' items correctly but miss the two most difficult items. However, if the examinee with a score of 8 out of 10 missed the two easy items but correctly answered all the other items, Sato's caution index would be high. Studies used these unusual patterns to flag schools or students who might be randomly responding to items. Researchers suggested that random responses could be because of test anxiety or carelessness but might also indicate that respondents were not instructed on the test material. If respondents were not instructed on the material, their responses might be more random than those of instructed students and might not follow the usual pattern of responses. These early ways of measuring instruction are in fact a possible outcome of ''the elusive indicator of complex student learning'' (Kennedy, 1999, p. 358) but fall short of providing information about instructional opportunities.

Later approaches to measuring instructional sensitivity build on these earlier studies by explicitly including more detailed measures of what students were taught (e.g., Hanson, McMorris, & Bailey, 1986; Mehrens & Phillips, 1986, 1987; Miller & Linn, 1988; Phillips & Mehrens, 1987). Muthén and colleagues (Muthén, 1989, 1994; Muthén et al., 1995; Muthén, Kao, & Burstein, 1991), for example, described the instructional measures in terms of teacher perceptions of student opportunities to learn the content included in the assessment. Teachers were asked two questions regarding each of the items on the mathematics assessment: (1) During this school year did you teach or review the mathematics needed to answer the item correctly? and (2) If during the school year you did not teach or review the mathematics needed to answer this item correctly, was it mainly because (a) it had been taught prior to this school year, (b) it will be taught later, (c) it is not in the school curriculum at all, (d) of other reasons? This method of gathering information relies on the self-reported perceptions of teachers about the instructional opportunities provided but did not directly measure instructional opportunities.

More recent approaches include approximations of instruction that are more closely linked to the instruction students actually received. Ruiz-Primo et al. (2002) studied instructional sensitivity by including multiple learning measures with different proximities to instruction and using student work as a source of information of the opportunities students had to learn the curriculum to estimate the degree of content coverage. These researchers developed different performance assessments that varied in the similarity of the characteristics of the curriculum-based activities conducted in the classroom. Some assessments were very close to these activities (close assessments) and other not as close (distal assessments). Students' science notebooks (considered as a record of the class activities) were collected in each classroom and scored for curriculum implementation, students' learning, and teachers' feedback practices. They administered pretests and posttests for the close and proximal learning measures. After standardizing scores on the pre- and posttest, differences between the standardized values were compared for the different achievement measures. The authors concluded that the assessment more closely related to instruction was more sensitive to instruction than the assessment less closely related to instruction.

How Students Were Taught. Although *what* students were taught provides one source of information about instructional opportunities, another approach focuses on *how* students were taught. Ruiz-Primo et al. (2012), for example, examined the extent to which teachers supported students to transfer what they learned by using a variety of indicators about the curriculum such as the type of knowledge required during particular science lessons. These indicators included classroom videotaping, students' science notebooks, and teacher interviews about how they provided opportunities to learn to capture how students were taught.

Measuring how students were taught requires attention to the underlying theory of how instruction is defined (Shavelson, Webb, & Burstein, 1986). It might seem fairly straightforward to measure instruction, or at least to agree upon the most essential dimensions of instruction; historically, however, this has not been the case. How mathematics is taught can be described in a variety of ways, including focusing on (1) teacher pedagogical content knowledge (Hill et al., 2008; Learning Mathematics for Teaching, 2011), (2) curriculum implementation (Stein & Kaufman, 2010), and (3) time spent on instruction (Weiss, Pasley, Smith, Banilower, & Heck, 2003). It is generally agreed that measuring instruction is a complex process and there is no consensus as to which aspect of instruction should be the focus or how the different aspects of measuring instruction relate to various student outcomes (Schlesinger & Jentsch, 2016).

To build on previous research on instructional sensitivity, this study includes two measures of mathematics achievement and three measures of instruction. Consistent with Burstein's recommendation that measures of instruction include content exposure and pedagogy, this study includes a variable related to *what* students were taught (see, Floden, 2002) and variables related to *how* students were taught (see, Ruiz-Primo et al., 2012). Using data from a unique large-scale project, this study provides evidence of the extent to which inferences about the instructional sensitivity of particular assessments depend on the way instruction is defined and measured. Comparisons across two grade levels and three districts allows for exploration of the consistency of this evidence.

Method

Study Context

Data for this study are from the Gates Measures of Effective Teaching Project (Kane, Kerr, & Pianta, 2014). The purpose of the project was to identify good teaching using a variety of measures (such as measures of student achievement; surveys of students, teachers, and principals; and scores of video-recorded lessons from multiple classroom observation protocols). With cooperation from six school districts throughout the United States, the project gathered data from elementary and secondary classrooms (fourth through ninth grade) across 2 years (academic year 2009-2010 and academic year 2010-2011).

This particular study used only a subsample of the full longitudinal database. Given the focus of this study on elementary mathematics, the subsample includes fourth- and fifth-grade students who completed the student mathematics achievement measures from the first year of data collection (academic year 2009-2010). The decision to include two different grade levels is based on previous work on instructional sensitivity using the same database that suggests differences between grade levels (Ing, 2016; Polikoff, 2016). Since each grade level received a different assessment, grade levels within each state were analyzed separately. Including two different elementary grade levels allows for comparison of how the instructional sensitivity of the assessments might vary for different grade levels.

Per agreements for use of these confidential data, the sample was further reduced to only include districts that had more than 10 teachers in each grade level. Three of the six districts were dropped based on these criteria. In addition, only the first year of data from this project was used because there was a larger sample of teachers and students in the first year compared with the second year. While there are questions about the power of classroom observations to predict student achievement gains (Casabianca, Lockwood, & McCaffrey, 2015; Kane et al., 2014), this study offers a unique opportunity to compare different observational protocols to better understand if there are patterns in the predictive power that differ by grade level and school district.

Participants

First, districts in the Measures of Effective Teaching Project study were recruited through ‘‘opportunity’’ sampling (July-November 2009). After six school districts from across the United States agreed to participate in the study, elementary, middle, and high schools within each district were recruited to participate. Finally, teachers within these schools at targeted grade levels and subject areas volunteered to participate (n = 2,741 teachers).

Overall, the teachers and students who volunteered to participate in the study were similar in terms of demographics to teachers and students in their districts (Kane et al., 2014). The overall sample included teachers who were mostly female (82%) and White (66%). The overall sample also included students who were approximately half female (48%), 31% African American and 36% Latino/Latina. Fifteen percent of the students were designated as English language learners and 11% of the students were designated as Special Education. The fourth- and fifth-grade students and their teachers from the three districts included in this particular study were also similar to the overall sample in terms of demographic characteristics. Specific information about the districts and teachers is not provided for confidentiality purposes.

Measures

Student Achievement.

Student performance on two different measures of mathematics achievement are included: a mathematics assessment specific to each state’s

Table 1. Descriptive Statistics of Student Achievement Scores and Correlations Between the Proximal and Distal Measures.

	Proximal		Distal		
	M	*SD*	*M*	*SD*	Correlation
Grade 4 (*n* = 8,298)	0.02	0.96	−0.01	1.01	.62
State A (*n* = 2,801)	0.01	0.98	−0.58	0.89	.62
State B (*n* = 3,013)	0.05	1.01	0.44	0.95	.76
State C (*n* = 2,484)	0.01	0.89	0.08	0.90	.66
Grade 5 (*n* = 9,336)	0.07	0.96	−0.03	1.00	.64
State A (*n* = 3,294)	0.02	0.95	−0.60	0.82	.62
State B (*n* = 3,194)	0.01	1.01	0.35	0.95	.76
State C (*n* = 2,848)	0.19	0.91	0.22	0.93	.70

Note. All correlations significant at $p < .001$.

curriculum (*proximal*) and a mathematics assessment that was not specific to each state's curriculum (*distal*; Inter-University Consortium for Political and Social Research, 2013). It is hypothesized that the *proximal* measure is more closely linked to the instructional opportunities of a particular classroom whereas the *distal* measure is less closely linked to the instructional opportunities of a particular classroom (Table 1).

Proximal. Students in the same grade from the same district were administered the same mathematics assessment. Each assessment was designed to measure student progress on the particular state's curriculum. Although information about each assessment is not publicly available, the assessments were primarily items with multiple-choice response options that were administered according to the specific state's timeline and procedures (White & Rowan, 2013). A *z*-score (*M* = 0, standard deviation [*SD*] = 1) relative to the grand mean of all students in the state was used as the student outcome on the state assessment.

Distal. All students were administered the Balanced Assessment in Mathematics (BAM). Although BAM was not designed to be aligned to any particular curriculum (referred to as *distal*), there is no evidence to date as to whether the BAM is more or less related to instruction compared with any particular state assessment. The assessment measures higher order reasoning skills and conceptual understanding, defined in terms of the following dimensions: modeling/formatting problems, transforming/manipulating mathematical formalisms, inferring/drawing conclusions, and communicating about mathematics. The assessment took approximately 50 to 60 minutes to complete and included multiple forms of four to five open-ended tasks. The assessments were scored on the same 4-point scale for each dimension assessed (which ranged from *attribute not present* to *attribute predominantly present*). Similar to the *proximal* measure, a *z*-score (*M* = 0, *SD* = 1) relative to all students based on the total score on all dimensions was used as the student outcome on the *distal* measure.

Table 2. Descriptive Statistics of Classroom Observation Protocol Scores.

	Subject Specific		General 1		General 2	
	M	*SD*	*M*	*SD*	*M*	*SD*
Grade 4 (*n* = 30)	1.33	0.17	2.67	0.20	4.64	0.36
District A (*n* = 10)	1.25	0.11	2.76	0.14	4.75	0.17
District B (*n* = 10)	1.32	0.16	2.65	0.22	4.63	0.45
District C (*n* = 10)	1.41	0.21	2.59	0.23	4.54	0.40
Grade 5 (*n* = 30)	1.36	0.19	2.74	0.18	4.68	0.35
District A (*n* = 10)	1.27	0.14	2.65	0.18	4.50	0.29
District B (*n* = 10)	1.40	0.22	2.84	0.14	4.91	0.33
District C (*n* = 10)	1.41	0.17	2.74	0.19	4.61	0.33

Instruction. Instruction was measured using data from classroom observations. Although the theoretical and methodological challenges of measuring instruction are widely acknowledged (see, e.g., Schlesinger & Jentsch, 2016), classroom observations are assumed to provide information about ‘‘the kind of intellectual work that teachers are asking of their students’’ as a ‘‘better indicator of the kind of work students are actually learning to do’’ (Kennedy, 1999, p. 346). Videotapes of the classrooms were coded by trained raters (Kane et al., 2014). The videos of mathematics instruction were scored using three different observational protocols: Framework for Teaching (Danielson, 2011), the Classroom Assessment Scoring System (Pianta, Hamre, Hayes, Mintz, & LaParo, 2008), and the Mathematical Quality of Instruction (Hill et al., 2008). The Mathematical Quality of Instruction is an observational protocol of subject-specific aspects of mathematics classroom instruction (referred to in this study as *Subject Specific*). In contrast, the Framework for Teaching (referred to in this study as *General 1*) and Classroom Assessment Scoring System (referred to in this study as *General 2*) are observational protocols of general aspects of classroom instruction (Table 2).

There were multiple checks and balances throughout the coding process including independent quality-control checks by external coders (Kane et al., 2014). The reported reliability coefficients for scoring these measures using trained raters ranged from .31 to .37 and increased to .6 to .7 when increasing the number of observations (Ho & Kane, 2013).

Measures of how students were taught.[1] One way to measure instructional quality is to consider generic or general aspects such as classroom management or cognitive demand (Pianta & Hamre, 2009). These are aspects of instruction applicable to any subject-area or grade level. In this particular study, two general measures of classroom instruction were included to describe how students were taught. Some example items from the general measures of instruction are classroom organization (behavior management, productivity, instructional learning formats) and classroom environment (creating an environment of respect and rapport, establishing a culture of learning).

Table 3. Correlations Between Classroom Observation Protocol Scores.

	1.	2.	3.
Subject-specific	—	.15	.22
General 1	.39*	—	.88**
General 2	.53**	.82***	—

Note. Correlations above diagonal for fourth grade (n = 30) and correlations below diagonal for fifth grade (n = 30).
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Another way to measure instructional quality is to consider the subject-specific aspects of mathematics instruction such as the mathematical accuracy of classroom discussions. These are aspects of instruction not applicable to all subject areas. In this particular study, one subject-specific measure of mathematics classroom instruction was included to describe how students were taught. Some example items from the subject-specific measure of instruction are the richness of the mathematics and responding to students' mathematical ideas. A single score averaging across the segments and items was created for each classroom on each of the three measures. A *z*-score (M = 0, SD = 1) relative to all classrooms based on the total score on all items were used as the classroom-level indicators of how students were taught. The correlation between these three measures of how students were taught was highest for the two general measures of instruction (Table 3). The correlation was lowest for fourth grade between the subject-specific measure of instruction and the two general measures of instruction. A similar pattern of relationships between these measures of instruction was consistent when looking at correlations of the measures within different school districts.

Analysis

The data used in this study were the typical structure of large-scale educational research with students nested in classrooms. An intercepts-as-outcomes multilevel regression analysis (Raudenbush & Bryk, 2002) was used to relate student performance to instructional opportunities:

$$Y_{ij} = \beta_{0j} + r_{ij}, \tag{1}$$

where β_{0j} is the mean class performance for students in class j and r_{ij} is the student-level residuals, assumed to be normally distributed with a mean 0 and variance σ^2.

In Equation (2), the goal is to determine whether the classroom-level measure of instruction variable (*Subject-Specific*) predicts student performance (Y_{ij} is the items student i in class j answered correctly):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Subject\text{–}Specific)_j + \mu_{0j}, \tag{2}$$

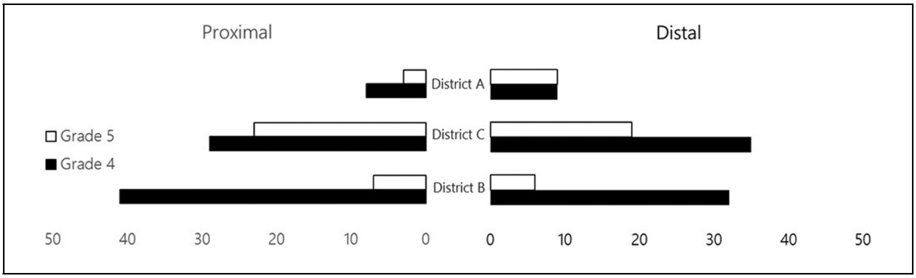

Figure 1. Percentage of variation between teachers on the proximal and distal assessments for fourth and fifth grades.

where γ_{00} is the average class mean achievement for population of classes; γ_{01} the main effect of *Subject-Specific* measure, that is, the expected change in class mean performance when *Subject-Specific* measure increases 1 unit; and u_{0j} the error between classes (classroom-level residuals), which is assumed to be normally distributed with a mean of 0 and variance, τ_{00}.

This model was replicated by substituting *General 1* and *General 2* as classroom-level predictors. Significant results for any of the instruction variables suggest that the particular way of defining and measuring instruction predicts student performance. Differences by district and grade level were also explored using the same approach.

Results

The intraclass correlations on the *proximal* assessment were typically higher than the correlations for the *distal* assessment (Figure 1). The intraclass correlations were also typically higher on the *proximal* measure compared with the *distal* measure (with the exception of District C fourth-grade student performance on the *distal* assessment). District B had a range of intraclass correlation with the highest correlation for fourth-grade student performance *proximal* assessment (41%) and lowest correlation for fifth-grade student performance on the *distal* assessment (6%). These correlations suggest there are differences between the *proximal* and *distal* measure depending on the grade level and district.

Figure 2 provides an example of how the same classrooms might be characterized in different ways using the three different measures of instruction. Classrooms that scored positively based on one measure of instruction might not be viewed the same way using another measure of instruction. While there are differences within each classroom in terms of different characterizations of the same observation, there were no differences between districts on *General 1,* $F(2, 57) = 0.90, p = .41$) or *General 2,* $F(2, 57) = 1.71, p = .19$. However, none of the three measures of instruction explained variation in student achievement. There was no additional variance explained after

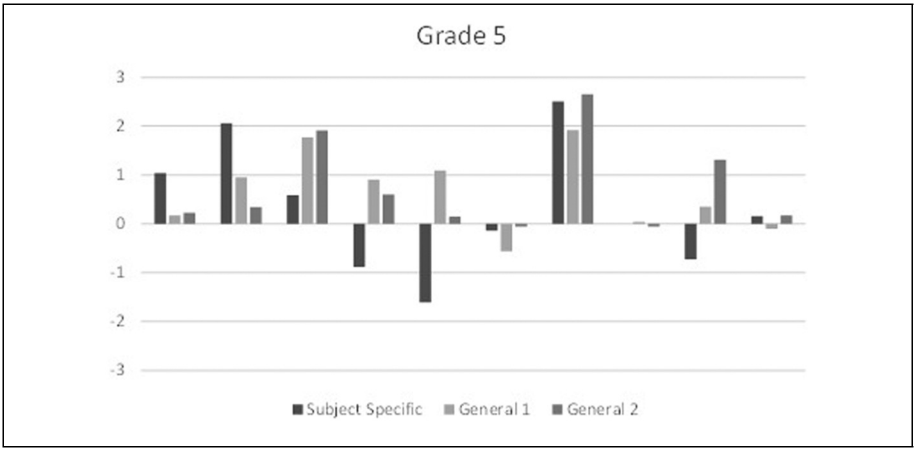

Figure 2. Example of variation in classroom observation *z*-scores for 10 fifth-grade classrooms.

Table 4. Summary of Multilevel Regression Coefficients.

	Grade 4 (*n* = 8,298)	Grade 5 (*n* = 9,336)
Proximal		
Subject Specific	−0.16 (0.60)	0.24 (0.36)
General 1	1.53 (0.34)***	1.01 (0.30)**
General 2	0.92 (0.21)***	0.38 (0.14)**
Distal		
Subject Specific	0.60 (0.61)	1.00 (0.59)
General 1	0.65 (0.61)	1.49 (0.47)**
General 2	0.51 (0.34)	0.81 (0.24)**

Note. Values are coefficients with standard errors in parentheses.
**$p < .01$. ***$p < .001$.

adding each of the measures, which suggests that none of these measures predict student achievement.

There was some variation of the predictive power of these different measures of instruction depending on grade level and type of assessment (Table 4). While both general measures were related to fourth- and fifth-grade student performance on the *proximal* measure, the subject-specific measure of instruction was not related to performance on the *proximal* or *distal* measure for either grade level. The consistency of the predictive relationship for fifth grade on the *proximal* and *distal* measures suggests that both the proximal and distal assessments might be sensitive to instruction but only if instruction is measured in terms of one of the two general measures of instruction. A different relationship between the variables was observed for fourth

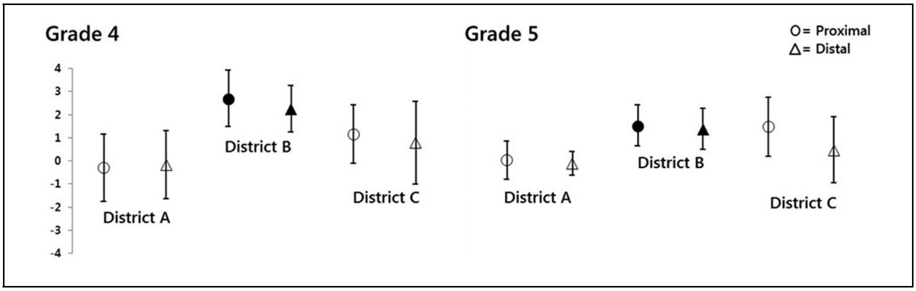

Figure 3. Coefficients by grade level and district for General 1 predicting student performance on the proximal and distal assessments.
Note. Shapes that are filled indicate a significant coefficient. Shapes that are not filled indicate an insignificant coefficient.

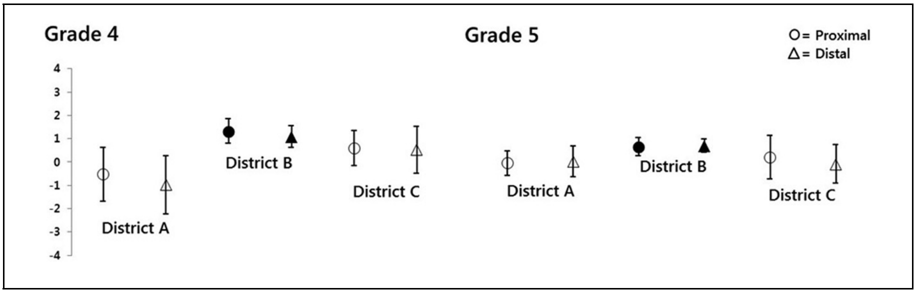

Figure 4. Coefficients by grade level and district for General 2 predicting student performance on the proximal and distal assessments.
Note. Shapes that are filled indicate a significant coefficient. Shapes that are not filled indicate an insignificant coefficient.

grade. For fourth graders, the lack of consistency suggests that the *proximal* measure is sensitive to instruction if instruction is measured using either of the two general measures but the *distal* measure is not sensitive to instruction. However, for either grade level, if instructional sensitivity was measured using the *Subject-Specific* measure, neither assessment would be identified as sensitive to instruction.

The relationship between the general measures of instruction and student performance was only observed for District B but not the other two districts (Figures 3 and 4). The S*ubject-Specific* measure of instruction was not correlated with student performance for any of the districts or grade levels. Neither of the two general measures of instruction significantly predicted student achievement in fourth and fifth grades for Districts A and C. For District B, there was consistency across fourth and fifth grades in terms of the relationship between both of the two general measures of instruction and student performance. This suggests that conclusions about which measure of instruction used to identify assessments that are sensitive to instruction varies by district.

Discussion

Popham (2007) referred to instructional sensitivity as ‘‘accountability’s dire drawback’’ given the emphasis on large-scale accountability tests. If interpretive arguments about the use of tests as indicators of the quality of instruction continue, evidence to support these claims are needed (Kane, 2013a, 2013b). The purpose of this study was to provide evidence about the role that different measures of instruction may potentially play in evaluating the instructional sensitivity of assessments. Results suggest that different measures of instruction lead to different conclusions about the instructional sensitivity of assessments.

The inconclusive results may be because of the difficulty in using classroom observations to measure mathematics instruction (Schoenfeld, 2013). While there is greater access to large-scale data from classroom observations, researchers continue to raise methodological issues about the use of observational data (Derry et al., 2010). In particular, researchers have questioned how well instruction can be measured at scale in ways that reflect differences in instructional opportunities (Correnti & Martinez, 2012). However, despite such claims, the use of these measures are widespread and are increasingly used for decisions about instruction.

Thus, even though the measures included in this study were close approximations of instruction (as compared with teacher or student survey responses), other ways of characterizing instruction (Ing & Webb, 2012) or additional observations may be required to better represent the variation in instructional opportunities (Ho & Kane, 2013). This study also did not consider the content of the instructional opportunities in relation to how students were taught. With more detailed data and drawing research related to teaching particular mathematical content, future research in this area could more closely link instruction and student performance. Another issue to consider is that these classroom-level approximations were not intentionally designed to capture individual student differences within a particular classroom. This allows for inferences based on classroom-level data rather than student-level data. However, given that most of the variation in student outcomes is at the student level, measures that capture instructional opportunities at a student level should be considered.

A broader issue to consider is whether or not these observational measures should be designed in ways that empirically relate to student outcomes. The *Subject-Specific* measure of instruction, for example, was not related to student performance in this sample of data. This lack of relationship is consistent with other analyses using the same data (see, e.g., Cantrell, 2012; Kane et al., 2014; Polikoff, 2016). However, it is possible that additional observations or different ways of aggregating or summarizing the segments are needed to better measure the instructional opportunities for all students (Hill & Grossman, 2013). It is also possible that the measures of instruction included in this study did not capture important differences in instructional opportunities in ways that relate to student outcomes. More general measures of instruction allow for scaling-up efforts and analyzing data across grade levels, but questions about the predictive ability of these measures remain. In contrast, more subject-specific measures of instruction, such as the one included in this study, might not be as easy to scale-up and generalize and relate to student outcomes, but might provide actionable information on instructional improvement indirectly linked to student outcomes. How these issues of scale and detail relate to inferences about the instructional sensitivity of different assessments is a topic for future directions of research in this area. In particular, future research could include different levels of approximation along the dimensions of *what* and *how* instruction is measured to allow for greater confidence around inferences of instructional quality as measured by student performance on different measures (Ruiz-Primo et al., 2012). Future studies that constrain the content of what is measured might better capture nuanced differences in instruction. In addition, a tighter link between instruction and the assessment items (rather than a summary score across all items) by including information about the content of the assessment items in relation to the content of the instruction is recommended. This sort of analysis was not possible with this particular data set but is a direction for future research in this area.

The distance of the assessments from instruction could be another source of inconsistency. In this study, the distance from instruction was assumed but not explicitly detailed. Ruiz-Primo et al. (2002, 2012) recommend being able to identify how closely linked an assessment is to instruction. In other words, how closely related to instruction is an assessment or to particular items on the assessment? This study did not consider the relationship between instruction and specific items on the assessment. Lacking this sort of information has implications for the methodology used to link instruction and student performance on the different assessments or different items from the same assessment (Naumann, Hochweber, & Klieme, 2016). This is a particularly important issue in this study of instructional sensitivity because it might be the case that the proximal measure is more of a distal measure if teachers did not implement the instruction as the professional development program intended. Measures of instruction that do not accurately capture differences in instructional opportunities might not allow for accurate inferences about differences in instructional opportunities based on student performance. Perhaps more nuanced measures of instruction that can be more directly tied to specific items within a particular assessment might lead to different conclusions about the overall instructional sensitivity of an assessment.

In taking these steps to disentangle the relationship between instructional opportunities and assessment, future research could rule out rival hypotheses that limitations in the measures of instruction obstruct evidence of validity. In doing so, future research can promote more in-depth studies of the extent to which student performance on an assessment can be used to infer differences in instructional experiences.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Measures of what was taught were considered but not included in this study. Teachers indicated the topic of the classroom observation but for this particular subsample, more than 41% of the segment topics were ‘‘random’’ topics. The next highest rated topic was multidigit multiplication and division (21%) and adding and subtracting fractions (11%). The remaining 27% of the topics were categorized into nine other areas. In five cases, only one segment was categorized into a particular topic (e.g., one segment was categorized as creating and analyzing graphs and tables and one segment was categorized as operations on rational numbers). For the other cases, segments were categorized into multiple topics. Several approaches to identifying the focal topic were considered (such as collapsing these focal topics into fewer categories and creating profiles of the content) but not used in the final analyses.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (2013). Commentary on ‘‘Validating the Interpretations and Uses of Test Scores.’’ *Journal of Educational Measurement*, *50*, 74-83.

Burstein, L. (1983). A word about this issue [Editor’s note]. *Journal of Educational Measurement*, *20*, 99-102.

Burstein, L. (1989). *Conceptual considerations in instructionally sensitive assessment* (CSE Technical Report 333). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Cantrell, S. M. (2012). The Measures of Effective Teaching Project: An experiment to build evidence and trust. *Education Finance and Policy*, *7*, 203-218. doi:10.1162/EDFP_a_00062

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*, 311-337.

Correnti, R., & Martinez, J. F. (2012). Conceptual, methodological, and policy issues in the study of teaching: Implications for improving instructional practice at scale. *Educational Assessment*, *17*, 51-61.

Cox, R. C., & Vargas, J. S. (1966). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

D’Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state’s standards-based assessment. *Educational Assessment*, *12*, 1-22.

Danielson, C. (2011). *The framework for teaching evaluation instrument*. Princeton, NJ: Danielson Group.

Derry, S. J., Pea, R. J., Barron, B., Engle, R. A., Erickson, F., Goldman, R., … Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences*, *19*, 3-53.

Floden, R. E. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231-267). Washington, DC: National Research Council.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*, 293-303.

Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement*, *50*, 84-90.

Haladyna, T. M., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement*, *18*, 39-53.

Hanna, G. S., & Bennett, J. A. (1984). Instructional sensitivity expanded. *Educational and Psychological Measurement*, *44*, 583-596.

Hanson, R. A., McMorris, R. F., & Bailey, J. D. (1986). Differences in instructional sensitivity between item formats and between achievement test items. *Journal of Educational Measurement*, *23*, 1-12.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement*, *20*, 191-206.

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*, 430-511.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*, 371-394.

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

Ing, M. (2016). Initial considerations when applying an instructional sensitivity framework: Partitioning the variation between and within classrooms for two mathematics assessments. *Applied Measurement in Education*, *29*, 122-131.

Ing, M., & Webb, N. M. (2012). Characterizing mathematics classroom practice: Impact of observation and coding choices. *Educational Measurement: Issues and Practice*, *31*, 14-26.

Inter-University Consortium for Political and Social Research. (2013). *Year 1 section-level analytical file 4th-8th grade codebook (ICPSR 34309-0001)*. Ann Arbor, MI: Author.

Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1-73.

Kane, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*, 115-122.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project*. San Francisco, CA: Jossey-Bass.

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, *21*, 345-363.

Learning Mathematics for Teaching. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25-47.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement*, *20*, 179-189.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement*, *23*, 185-196.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement*, *24*, 357-370.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education/Macmillan.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, *25*, 205-219.

Muthén, B. O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika*, *54*, 385-396.

Muthén, B. O. (1994). Instructionally sensitive psychometrics: Applications to the Second International Mathematics Study. In I. Westbury, C. A. Ethington, L. A. Sosniak, & D. P. Baker (Eds.), *In search of more effective mathematics instruction* (pp. 293-324). Norwood, NJ: Ablex.

Muthén, B. O., Huang, L. C., Khoo, S. K., Goff, G. H., Novak, J. R., & Shin, J. C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis*, *17*, 371-403.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*, 1-22.

Naumann, A., Hochweber, J., & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. *Educational Assessment*, *21*, 89-101.

Phillips, S. E., & Mehrens, W. A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. *Journal of Educational Measurement*, *24*, 1-16.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109-119.

Pianta, R. C., Hamre, B., Hayes, N., Mintz, S., & LaParo, K. M. (2008). *Classroom Assessment Scoring System–Secondary (CLASS-S)*. Charlottesville: University of Virginia.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issue and Practice*, *29*(4), 3-14.

Polikoff, M. S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment*, *21*, 102-119.

Popham, W. J. (1971). Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction* (pp. 79-98). Englewood Cliffs, NJ: Educational Technology.

Popham, W. J. (2007). Instructional sensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*, 146-150.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching*, *49*, 691-712.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. S., & Klein, S. (2002). On the evaluation of systematic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369-393.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meiji Tosho.

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, *48*, 29-40.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM Mathematics Education*, *45*, 607-621.

Shavelson, R. J., Webb, N. M., & Burstein, L, (1986). The measurement of teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50-91). New York, NY: Macmillan.

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

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K-12 mathematics and science education in the United States*. Chapel Hill, NC: Horizon Research.

White, M., & Rowan, B. (2013). *User guide to measure of effective teaching longitudinal database*. Ann Arbor: Inter-University Consortium for Political and Social Research, University of Michigan.