Title

Mathematical Classroom Discussion as an Equitable Practice: Effects on Elementary English Learners’ Performance

Journal
Journal of Language Identity & Education, 17(6)

ISSN
1532-7701

Authors
Banes, Leslie C
Ambrose, Rebecca C
Bayley, Robert
et al.

Publication Date
2018-11-02

DOI
10.1080/15348458.2018.1500464

Peer reviewed

Relating Performance on Written Assessments to Features of Mathematics Discussion

Abstract

Many researchers have illustrated the multi-faceted nature of classroom mathematics discussions, but few have demonstrated the effect of discussion on students' assessment performance. We developed and employed a discussion observation instrument in 20 third and fourth-grade classrooms in an economically-disadvantaged, linguistically-diverse school district and used hierarchical linear modeling (HLM) to determine whether between-class variation in word-problem-test scores can be explained by levels of class discussion. Results suggest overall class discussion scores, as well as two specific discussion features, variety of approaches and opportunities to speak, are significantly related to test performance. These results suggest classroom instruction including high-level math discussion may improve students' performance on written measures of achievement.

Keywords: hierarchical linear modeling, discussion, math education, observation instrument

Educators have long advocated discussion to enhance students' mathematics learning (e.g., NCTM, 1989, 2000, 2014), yet discussion has remained rare in United States classrooms, especially in high-poverty schools with many English Learners, or Emergent Bilinguals (EBs)[1] (Gallimore, Hiebert, & Ermeling, 2014). The recent adoption of the Common Core Standards has shifted professional rhetoric about instruction and may be motivating teachers to include more discussion. However, many educators remain skeptical about the benefits of discussion because research demonstrating the relationship between discussion and performance on assessments is still in an early stage (Stein, Correnti, Moore, Russell, & Kelly, 2017). Moreover, orchestrating productive discussions is demanding, requiring teachers to attend to several facets of classroom activity (Walshaw & Anthony, 2008). In the present reform era with renewed emphasis on discussion, it remains to be seen which aspects of discussion teachers implement, and whether implementing some aspects of productive discussion affects performance on assessments.

This study analyzes data from 20 third and fourth grade classrooms in an economically disadvantaged community where administrators encouraged teachers to incorporate discussion in their classrooms. We investigate which aspects of discussion teachers implemented and their relationship to achievement. We address the following questions:

1. What aspects of mathematics discussions were observed in a poor linguistically-diverse elementary school district?
2. What aspects of mathematics classroom discussions were associated with students' performance on a mathematics achievement measure?

Theoretical Frames

Discussion motivates students, supports teachers' assessment of students' understanding,

[1] We use the term emerging bilingual (EB) instead of English learner to refer to students who are in the dynamic process of developing bilingual competencies (Escamilla et al., 2014) and to emphasize the value of bilingualism (García, 2009).

and helps students understand the nature of mathematical activity (Cirillo, 2013). While we value these benefits, here we focus on the benefits associated with improved performance on achievement measures, which policy makers and administrators emphasize. In explaining the paucity of research connecting instructional practices to mathematics achievement, Hiebert and Grouws (2007) noted that contextual factors must be accounted for when analyzing instruction. These contextual factors explain why few researchers have embraced generic observational measures of mathematics instruction; rather, researchers have developed frameworks to describe specific aspects of instruction present in particular settings. In some cases, researchers developed frameworks to explain observations of productive activities in a single or a few classrooms (Cobo & Fortuny, 2000), while other researchers designed measures to evaluate teaching in a larger number of classrooms (Sawada et al., 2002). These instruments and frameworks include analyses of discussion among other aspects, including cognitive demand of tasks and teacher moves.

Facets of Discussion

Subscribing to the belief that "effective teaching requires the skillful coordination of multiple practices" (Halpin & Kieffer, 2015, p. 263), we began to develop an observation rubric with the assumption that each facet was distinct while remaining open to the possibility that some facets might co-vary. In the following sections we present the facets of discussion appropriate to our study. We also explain how participating in these facets of discussion, in theory, should affect student performance on tests. We close by discussing how observational measures for mathematics instruction in other studies account for similar facets.

Opportunities to speak. Cognitive scientists found when students explain ideas to themselves they learn more than when they work in silence, and when they explain ideas to another, they learn even more (Rittle-Johnson, Saylor, & Swygert, 2008). Moreover, students learn more quickly and retain more when they articulate a strategy than when they spend the same amount of time practicing a procedure (McEldoon, Durkin, & Rittle-Johnson, 2013). When students put an idea into their own words, they can encounter otherwise unnoticed holes in their thinking. To realize the benefit of class discussion, student contributions need to be substantive so they have the opportunity to articulate and revise their mental models (Chi, 2000).

Equitable participation. While a classroom may have a high proportion of student talk, sometimes a few students dominate the discussion while those positioned as having low status are left out (Herbel-Eisenmann, Steele, & Cirillo, 2013). Equitable participation implies that the majority of students are making their thinking public, which, in addition to verbal contributions, could include students' showing hand signals or holding up whiteboards to share their work and their answers publicly (Forman, McCormick, & Donato,1998). We include nonverbal displays of thinking to broaden what counts as participation for EBs, who may not be able to share ideas verbally in English, but can communicate their thinking visually (Moschkovich, 2013). Others have found that silent students gain from discussion when they are actively listening so verbal, as well as non-verbal, participation should enhance students' performance on achievement measures (O'Connor, Michaels, Chapin, & Harbaugh, 2017).

Variety of approaches and resources. One could imagine discussions where many students speak, all students make their thinking visible in some way, and the talk is about a *single* procedure or strategy. Heinze, Marschick, and Lipowsky (2009), however, have shown students benefit by having access to *multiple strategies and resources* for solving problems. Strategies are approaches for solving a problem and can be recognized as different when they include different operations, numbers of steps, or quantities that are decomposed differently or altered. Discussions that reveal an array of strategies provide students with access to strategies at their own, as well as more advanced levels (Murata, 2012). Access to multiple strategies can allow students to appropriate strategies when they are ready to make sense of them and help students become more adaptive by selecting the strategy most appropriate for the task at hand (Siegler, 1996).

Concrete and visual representations provide greater access, especially for EBs who may have difficulty comprehending verbal explanations (Moschkovich, 2013). In addition to making mathematics more accessible, offering a variety of resources promotes engagement and autonomy by offering choice (Assor, 2012). Students might choose different representations but use them in a similar fashion. Based on research that shows students learn more when they solve a problem in multiple ways than when they solve only one problem using one way for each (Evans & Swan, 2014; Silver et al., 2005), we consider a discussion of *multiple* approaches for solving a problem a higher-level discussion than one that includes multiple resources to represent the *same* approach. While both multiple approaches and multiple resources may support students in discussion by offering multiple entry points and invoke different conceptual understandings (Richland et al., 2017), we position multiple approaches as the heart of a rich math discussion. When a problem is solved more than one way, opportunities are created for explanation, justification, and making connections as students work to understand and critique others' ideas (Kazemi & Hintz, 2014).

Conceptual explanations. When students discuss their mathematics work, sometimes they focus on procedures and sometimes on underlying principles. When students explain "why" they did what they did, they learn more than when they simply explain "how" they solved a problem (Matthews & Rittle-Johnson, 2009). When students understand concepts, they tend to transfer that knowledge to new problems (Hiebert & Wearne, 1993). Conceptual explanations, which include reasoning and justification, activate reflection on the situations for which a procedure is effective so when students have a variety of problems to solve on an achievement test they are better positioned to choose an appropriate approach.

Connections between ideas. While hearing multiple approaches can provide students with one that matches their developmental level, considering the connections between various approaches can deepen their mathematical understanding of underlying concepts (Richland et al., 2017; Durkin, Star, & Rittle-Johnson, 2017). Hiebert and Carpenter (1992) argue that connecting various representations and strategies leads to the robust understanding that allows students to solve a range of problems, while Jones, Swan, & Pollit (2015) suggest effective discussions enable students to connect new concepts to prior knowledge. Connections may arise when students compare different strategies or recognize the limitations of a particular tool. For example, instead of a discussion that simply includes a "show and tell" of various approaches for adding ¼ and ¼, a rich discussion might include comparing and contrasting approaches with the goal of illuminating why we do not add denominators.

Because learners have to construct their individual understandings by building on existing knowledge (Mathematics Learning Study Committee, 2001), teachers cannot provide students with connections; rather students make connections by reflecting on one another's contributions and sharing their conclusions. The robust understanding that grows from making connections among strategies and across problem types should enable students to transfer their knowledge to the unfamiliar problems that might be provided on an achievement test.

Discussion Facets on Observation Measures in Other Studies

During the past 20 years, researchers have developed observation instruments to capture instructional practices specific to mathematics instruction. These instruments vary in purpose and emphasize different aspects of classroom activity (Boston, Bostic, Lesseig, & Sherman, 2015). Most instruments include attention to aspects of discussion. Here, we describe how the facets highlighted above have been included in four instruments employed in other studies: 1) the

Classroom Observation Instrument (COI) used in the QUASAR studies (Stein & lane, 1996); 2) the Reformed Teaching Observation Protocol (RTOP) (Sawada et al., 2002); 3) The Mathematics Scan (M-Scan) (Walkowiak, Berry, Meyer Rimm-Kaufman & Ottmar, 2014); 4) the Instructional Quality Assessment Classroom Observation Tool (IQA) (Boston, 2012).

The COI requires observers to respond to a series of questions about the lessons they observed, including the nature of the mathematics task, classroom discourse, and the intellectual environment (Stein, Grover & Henningsen, 1996). The RTOP and subsequent adaptations (e.g., Akkus & Hand, 2010), designed for both science and mathematics classrooms, have been widely used to assess teachers (Boston et al., 2015). The M-Scan and the IQA are more recent instruments designed to be used in a large number of classrooms to determine general mathematics instructional quality. Both include rubrics that describe activities in classrooms at different levels.

While "opportunities to speak" was not treated as a separate facet in any of the instruments, the other four facets of discussion included in our rubric were present in most. Specifically, the RTOP (Sawada et al., 2002) and M-Scan (Walkowiak, et al., 2014) treated these four facets as distinct, while the other two instruments combined two or more of our facets. The IQA (Boston, 2012) combined conceptual explanations, variety of approaches and connections into a single rubric category. All of the instruments identified variety of approaches/resources and conceptual explanations as facets of quality instructions. "Connections" was defined in slightly different ways across instruments. The RTOP and M-Scan included references to connections made to the "real world" and the M-Scan alluded to connections between strategies by mentioning translations between representations. All except the COI included equitable participation.

While researchers generally agree about the various facets, projects vary in how they combine facets and their prominence in the overall scale. All of these instruments measure other aspects of instruction in addition to these facets of discussion and aggregate ratings to determine a final score. The goal of developing our own observation rubric was for a more fine-grained analysis that would allow us to determine which facets of discussion were most common and the ways in which the facets were associated with student performance on our achievement measure.

Role of curriculum in shaping discussion. While our facets correspond to facets others have measured, we did *not* include a facet that does appear on all of the other measures, specifically, the nature of the mathematical tasks that Stein, Grover, and Henningsen (1996) identified as a major determinant in students' opportunities to learn in discussions. Instead of analyzing the task as a separate component, each category on our observation rubric captures how mathematical ideas were treated within the discussion. For example, a low score in explanations implies students simply stated a brief answer. Conversely, high-level discussions are more likely to include open-ended tasks that allow students to compare various solution strategies and explain their reasoning (Crespo & Harper, in press).

Although we did not score the mathematical tasks themselves, the features on our instrument implicitly capture some task dimensions, the types of questions teachers asked, and the classroom expectations for conceptual understanding and justification of ideas. We did not include a specific component assessing the task in our observation instrument because teachers in our study had limited discretion over the tasks they would use. Rather, teachers were required to use the published curriculum and follow the pacing guide provided by the district administrators (Diamond, 2007). Fortunately, curriculum employed by the district supported discussion to an extent. Moreover, since teachers were free to use the text/tasks in any way they chose, we found it appropriate to assess how students engaged with the tasks through class discussion, rather than the tasks themselves. We provide more details about the text and tasks in the methods section.

Previous Research

Walshaw and Anthony (2008) pointed out the "enormous complexity" (p. 543) of facilitating discussion with all of these facets. Given this complexity, as teachers learn to integrate discussion into their practice, they may implement some features before others. In this section we discuss studies that documented the extent to which these various facets appeared in elementary school class discussion. We organize this section into two parts: studies that discuss the participation aspects of discussion, i.e. opportunities to speak and equitable participation, and studies that discuss the content aspects, including multiple approaches/resources, explanations, and connections. We then discuss studies that explore the relationship between mathematics instruction and student achievement.

Empirical Studies Measuring Participation

Some researchers have looked for relationships between student level of participation in discussion and performance on mathematics achievement tests. Kosko and Miyazaki (2012) found mixed results when they analyzed the extent of students' participation. They surveyed 3,632 teachers about how often particular students discussed the solutions to math problems with other children using a dichotomous variable, "at least once a week" or "less than once a week." They compared this with students' scores on a standardized mathematics test and found that in some schools participation in discussion was positively associated with students' math scores and in other schools discussion was negatively associated with students' math scores. Kosko and Miyazaki claimed that social factors, such as student effort and teacher training, contribute more to the success of discussion than the frequency. This suggests that opportunities to speak in math discussion may be insufficient to ensure strong performance on achievement measures.

In their study of five elementary classrooms where teachers engaged students in discussion, Baxter, Woodward and Olson (2001) found that low achieving students seldom spoke in whole group discussion and often seemed distracted. They attributed this, in part, to lack of opportunities to speak, and noted: "low achievers seemed to disappear during whole-class discussion" (p. 545). Similarly, Planas & Gorgorió (2004) found that English-only students spoke more during mathematical discussions than their multilingual counterparts. These findings highlight the importance of attending to both opportunities to speak and to equitable participation.

Empirical Studies Measuring Content of Discussion

In a report of the QUASAR project, Stein, Grover, and Henningsen (1996) analyzed 144 mathematics lessons in diverse urban schools. They found that 60% of the lessons involved multiple representations (symbols and non-symbols) and half included multiple solution strategies and mathematical explanations. Thirty-three percent engaged students in making connections in some way. Like some of the teachers in our study, the QUASAR teachers were involved in professional development, and they were working in economically disadvantaged schools with large numbers of students of color.

In their study of four 4th grade classrooms, Kazemi and Stipek (2001) found that students were sharing solution strategies in all the classrooms. These findings suggest that opportunities to speak and variety of approaches may be the most common features in classroom discussions. Kazemi and Stipek observed only one classroom where the teacher pressed for justification that elevated students' explanations to focus on underlying concepts, suggesting that it may be more difficult for teachers to elicit conceptual explanations from students than to simply to get them talking, and may depend on other discussion features being in place.

Spillane and Zeuli (1999) observed instruction in 25 4th and 7th grade classrooms with teachers who were well versed in reform ideas. In four classrooms, children explained their reasoning, justified their approaches, and defended their solutions. In two of these classrooms, a few students dominated discussions. In 10 classrooms, students talked about their approaches to solving problems and "teachers seemed to recognize the importance of having students publicly explain and support their ideas" (p. 15), but answers rather than concepts were the focus. In the remaining 11 classrooms, discussion focused on learning a single procedure to generate the answer, and students rarely participated. In this study, opportunities to speak about multiple strategies were more common than conceptual explanations and connections between them. Moreover, even when teachers have some knowledge of the importance of class discussion 44% did not provide opportunities for students to speak.

Forman, McCormick, and Donato (1998), who analyzed a 6th grade lesson in which students presented different strategies for one problem, illustrate how rare the connections aspect of discussion is. Students' contributions were substantive and lengthy. The teacher pressed the students for complete explanations. She also encouraged students to generalize their strategies, elevating them to a more conceptual level. Because the teacher privileged one approach to the problem above others, she did not provide students an opportunity to look for similarities and differences between approaches. In this case, students had opportunities to speak, the class discussed a variety of approaches, and the explanations were conceptual. Nevertheless, connections were absent.

In considering the various kinds of discussions between students, Mercer (1996) differentiated between "cumulative talk" and "exploratory talk". In "cumulative talk," students have opportunities to speak and make their ideas public, and a variety of approaches are available for students to consider. In "exploratory talk," approaches are open to scrutiny and challenged, so that students have to explain and justify their approaches. In this kind of talk, students are more likely to make connections because they are more actively engaging in one another's thinking. The set of studies reviewed above suggests that "cumulative talk" is more likely to be present than "exploratory talk" in our findings. Specifically, among the features of classroom discussion that we identified, we are most likely to find multiple approaches and opportunities to speak. Corkin, Coleman and Ekmekci's work (2019) in urban, high-poverty classrooms indicates that teachers have great difficulty implementing student-centered mathematics instruction in these settings, so it's possible that in many classrooms, students may not have opportunities to say more than one or two words. Since mathematics discussions in high-poverty schools continue to be rare, we believe it's important to establish the relationship between discussion and student performance on achievement tests. We turn to this relationship in the next section.

Relations Between Mathematics Instruction and Achievement

Like Hiebert and Grouws (2007), we found few studies that correlated mathematics classroom discussion practices with students' achievement. Specifically, Hiebert and Wearne (1993) developed pre and post written assessments to capture skills that were emphasized in both traditional and reform second-grade classrooms. Students who spent time asking questions and explained problems using multiple solution strategies scored higher on achievement tests than those who practiced rote procedures. Similarly, Mercer and Sams (2006) used pre and post Student Attainment Tasks (SATs) to analyze the mathematical development of elementary students. They found that students in discussion-based classrooms achieved greater gains on the SATs than students in classrooms without discussion. In Stein and Lane's (1996) study, student achievement gains were greater in schools where students had lessons that tended to include multiple solution strategies and explanations than in schools where these aspects of discussion were less frequent. Merritt, Rimm-Kaufman, Berry, Walkowiak, and Larsen (2011) found that mathematical instructional quality accounted for 8% of classroom level variance in third graders' math achievement. Merritt et al. treated instructional quality as a single variable, including multiple representations, use of mathematical tools, cognitive depth, mathematical discourse community, explanation and justification, problem solving, and connections and applications.

The studies just mentioned included the classroom or school as the unit of analysis, categorizing groups of students according to the level of discussion in their environment. Other researchers analyzed individual students' engagement in discussion and associated that with achievement. They have tried, with mixed results, to determine if students who are more active in discussion perform better than less active students. Webb, et al. (2014) found that students who explained their ideas in detail and engaged with the details of classmates' ideas by asking questions or disagreeing with their approaches performed better on a mathematics achievement measure than their less engaged peers. This study shows the significant effect that active participation in all facets of discussion can have on student performance. Kosko and Miyazaki (2012), however, who used teachers' ratings of students' participation in discussion and correlated these with students' performance on a standardized mathematics assessment, did not find a correlation between participation in discussion and achievement. We note that Webb et al.'s (2014) study took place in a single elementary school where mathematics discussions usually included the facets we have outlined, whereas Kosko and Miyazaki's study included students in classrooms across the country where presumably the teachers were using a range of instructional strategies. Together these studies demonstrate the importance of attending to various facets of mathematics classroom discussion to better understand what aspects contribute to achievement.

As Hiebert and Grouws (2007) pointed out, research connecting classroom discussions and student learning is just beginning to emerge. In addition, they cautioned that teaching takes place in a complex system where a variety of forces interact to affect what takes place. They indicated a continued need for studies that look for correlations between teaching practices and learning outcomes. The next section provides information about the context of our study.

Methods and Data Sources

Setting

Our study took place in a small elementary school district in California with five Kindergarten through 6th grade schools. As Table 1 shows, the district serves an ethnically diverse, low-income urban community with a substantial EB population and a veteran teaching force. With the exception of School M, which was smaller, the schools were of similar size and their student bodies were similar to one another. Table 1 displays demographic information of the teachers and students in participating schools, with low income defined by percentage of students qualifying for free or reduced-priced lunch, and percent at or above grade level proficiency in math measured by the California Standards Test (CST).

Table 1. Demographics of Participating Schools

	# students	% low income	% EB	% students proficient in math	% Hispanic	% Asian	% African American	Teacher mean # years teaching
School M	270	88.5	30.0	56.2	42.6	15.2	16.3	19
School T	453	90.3	48.3	45.2	44.6	21.9	16.6	24
School B	425	99.1	44.9	58.0	46.2	17.5	16.3	16
School R	501	91.4	43.3	64.0	41.3	18.6	6.6	14
School G	467	89.5	49.5	42.4	59.5	10.3	20.8	17

Some of the teachers in the district volunteered to participate in mathematics professional development (PD), which included monthly two-hour meetings and one week of summer workshops to discuss children's mathematical thinking. In the workshops, teachers engaged in mathematical discussions and observed videos of children explaining their strategies for problem solving.

All teachers were required to use the district's mathematics textbook, *enVision Math* (Charles et al., 2009) and adhere to a pacing guide, although teachers varied substantially in how and to what extent they used the textbook to guide their class discussions. In this series, each daily lesson included an *Interactive* section, which invited students to contemplate the mathematics concept of the day by solving a problem. Sometimes the problems could be solved in more than one way, and while the teachers' manual did not identify the variety of approaches that might emerge in the classroom, this part of the curriculum could support a discussion of strategies that students figured out for themselves. After the *Interactive* section, every lesson included a *Visual Learning Bridge*, a computer animation that guided students through a procedural explanation with discussion prompts. Teachers were expected to engage students in extended talk about these prompts. The remaining sections provided practice problems and assessments. Because each lesson included more material than a teacher could use, teachers had some discretion about which parts they included, but most felt obligated to use the text in some way every day. We did not delve into teachers' fidelity of curriculum use, because we observed that there were strong classroom discussions that adhered closely to the curriculum and others that did not. Due to curriculum including discussion prompts and teacher participation in our discussion-focused PD, and a district initiative focused on Common Core math instruction, we assume teachers and students in every classroom had been learning at least something about math discussion prior to this study, though levels of discussion varied from class to class, as indicated by our measure of discussion described below.

Classroom Discussion Instrument Development

Because we did not find an instrument that matched our needs, we created our own, which consisted of the five facets of discussion described above (see supplemental materials for full rubric). Each facet was scored on a five-point scale (0-4). To determine how well the facets and associated rating scales captured the quality of whole class discussions, we piloted it with three live classroom discussions and two classroom discussion videos. We then revised the instrument to clarify the criteria for each level and decided to combine multiple approaches and representations to align with what we saw in classrooms. We also used the term "resources" instead of representations to include the tools students used, such as multiplication tables and rulers, in addition to representations such as drawings, graphs, and symbolic notation.

Measure of Discussion

Two researchers visited all 20 of the third and fourth grade classrooms in the district's five schools in January and February 2013. Teachers were asked to teach a typical lesson, and the observers took field notes during the lesson to document various facets of discussion. The researchers did not know the teachers or have preconceived notions about the classrooms.

The researchers scored the lesson independently immediately after observations using our project-developed rubric and then discussed their scores to resolve scoring differences. Consensus estimates, one approach to inter-rater reliability, measure the degree to which a group of raters give the same score to the same performance and are calculated as a percentage of total or adjacent agreement (Jonsson & Svingby, 2007). Thus, total scores within 10% (2 points) of each other were considered to be consistent. Our final consensus estimate for individual scores between the two raters was 75%. Stemler (2004) suggests consensus estimates of 70% or higher are generally accepted as an indication of inter-rater reliability for rubric scores. Moreover, our process of debriefing after scoring individually mitigates potential issues with inter-rater reliability. Because both scorers scored every discussion and discussed any discrepancies until reaching agreement, all scoring differences were resolved in the final scores used for analysis.

Measures of Achievement

To measure mathematics achievement, we created a linguistically modified math assessment (LMMA) by modifying word problems from past Trends in Math and Science Study

(TIMSS) and National Assessment of Educational Progress (NAEP) assessments, following recommendations for assessing EBs (Abedi & Lord, 2001). Strategic modification of word problems is one way of increasing comprehensibility and reducing linguistic bias for EBs (Abedi & Lord, 2001). Our research team modified the language of items while keeping the mathematics and content-related vocabulary the same. Differential item functioning demonstrated that no single item was more difficult for EBs than for their peers, showing that the linguistic modifications may have reduced the linguistic complexity enough to eliminate any substantial linguistic bias for EBs. For more information on the development of the LMMA and linguistic modification of the items, see Banes, Ambrose, Bayley, Restani, and Martin (2018).

The third-grade version of the test contained 11 items. The fourth-grade version had an additional 6 items. The LMMA included eight open-response items; of these, two items required students to write explanations of their solution. It also included several non-canonical and multistep word problems aligned to the Common Core math assessments.

To measure the validity of the LMMA we correlated students' scores with their performance on the Iowa Test of Basic Skills using Z scores and found it to be .752, indicating that the LMMA has convergent validity and that the construct it measures (conceptual mathematical understanding) overlaps substantially with the ITBS. However, we believe the correlation between the two tests is not higher due to linguistic modification of the word problem test to reduce bias for EBs and absence from the LMMA of items that are strictly computational while the ITBS includes a large computation section. Cronbach's alpha for the 3rd grade test was .69 and .81 for the 4th grade test, which indicates "good" internal reliability (Kline, 2000). The higher reliability on the 4th grade test is likely due to the addition of 6 items on that test. Given the amount of time it takes younger students to complete multi-step word problems, adding word problems to the third-grade test would have been inappropriate. For the full assessment and its psychometric properties, see: https://www.researchgate.net/publication/317238876_Linguistically_Modified_Mathematics_Assessment.

A trained research assistant scored the tests. Students' California Standards Test (CST) scores from the previous school year were used as a control variable to account for students' baseline mathematics achievement. The CST was used instead of the ITBS as a measure of prior mathematics achievement because it was administered before students entered their current classrooms, while the ITBS was given at roughly the same time as the LMMA.

Analysis

We began our analysis with classroom scores on each of the elements in our observation rubric. We used factor analysis to remove the redundancy of correlated variables and to uncover any latent variables within our five observed discussion features. We ran an exploratory factor analysis with STATA using each of our five discussion facets and computed Eigenvalues. As discussed below, we determined there were two latent factors and then used an orthogonal rotation to determine which of our discussion elements loaded on each of the two factors.

After determining that our discussion facets could be aggregated into two factors, we employed hierarchical linear modeling (HLM) to determine how much of the between-class variation in performance could be explained by class discussion overall and by each of the two factors. HLM offers a way to uncover the relationship between class-level and student-level variables while ensuring more credible statistical results than traditional regression modeling (Raudenbush & Bryk, 2002). Although discussion scores captured by our observation rubric are ordinal, we follow Long and Freese (2006) and Pasta (2009) who make a strong case for treating ordinal variables as continuous, when the assumption of linearity is confirmed, to reveal statistical relationships that might otherwise be overlooked. Ordinal variables with five or more categories can often be used as continuous without harm to the analysis (Norman, 2010; Sullivan & Artino, 2013). Thus, discussion scores in our analysis represent an ordinal approximation of a continuous variable.

A set of models was investigated first using the total discussion score as the primary class-level predictor and later with the score of each factor of class discussion to assess the significance of their effects. HLM models were specified beginning with limited models and adding parameters sequentially, keeping only those that proved significant (Hox, 2010). Model 1 is the null model, which includes only the intercept, student-level error, and class-level error. It was used to determine the amount of variation in LMMA scores between classes. Model 2 includes the student-level explanatory variable CST as a fixed effect, meaning its effect was not allowed to vary across classes. In model 3, the class-level explanatory variable Discussion is included and fixed. In model 4, the effect of CST is allowed to vary randomly across classes. Model 5 includes a two-way interaction between Discussion and CST and was used to determine whether the relationship between discussion scores and LMMA scores varies by students' previous mathematics performance. Finally, models 6 through 10 investigate the effects of each facet of discussion on LMMA performance, while models 11 and 12 investigate these facets of discussion grouped into two separate factors. Following Hox (2010), we compared deviance statistics of reduced and full models to determine which parameters produce the best fitting models.

Findings

What Aspects of Mathematics Classroom Discussion Were Observed?

Table 2 illustrates the means for each discussion facet. Data indicate that a variety of approaches was available in most classrooms for at least part of the observed lesson. Usually, children shared their strategies in sustained verbal contributions, but the four scores of 1 and 2 in "opportunities to speak" indicate that the teacher asked short answer questions about strategies rather than have students explain independently. Most classrooms received scores of 3 or 4 on "equitable participation," meaning children were actively engaged and had some means of communicating their thinking, including speaking, hand signals, or showing work on whiteboards. Connections between ideas and conceptual explanations were much less frequent than the other facets.

Table 2. Discussion Means Across Classrooms

Discussion Feature	Mean Score	Sd	Range
Conceptual Explanations	2.3	.86	1-4
Connection between ideas	2.55	.84	1-4
Opportunities to speak	3.05	.77	1-4
Equitable participation	3.6	.56	2-4
Variety of approaches	3.6	.64	2-4
Total Score	**15.1**	**2.69**	**10-20**

Table 3 illustrates the results of the factor analysis used to determine which discussion facets tended to be related to one another. The two-factor solution had Eigenvalues greater than one and accounted for an additional 27% of the variance over the one factor solution.

Table 3. Principal Component Analysis

Factor	Eigenvalue	Proportion of variance accounted for	Cumulative variance accounted for
Factor 1	2.62	.52	.52
Factor 2	1.34	.27	.79
Factor 3	.69	.14	.93
Factor 4	.26	.05	.98
Factor 5	.09	.02	1.0

We performed a rotation using the orthogonal varimax approach to determine the loadings of each element on the two factors (see Table 4). We found that the two facets, connections and conceptual explanations, constituted Factor 1 with loadings greater than .9. The two facets, opportunities to speak and variety of approaches, constituted Factor 2 with loadings greater than .85. The facet equitable participation was distributed fairly evenly across the two factors, but since its uniqueness level was not high (.42), it did not stand alone as a third factor.

Table 4. Factor Loadings

Discussion Elements	Loading on Factor 1	Loading on Factor 2
Connections	.97	.09
Conceptual Explanations	.91	.08
Equitable Participation	.58	.49
Opportunities to Speak	.09	.85
Variety of Approaches	.10	.92

Factor 1 (connections & explanations) and Factor 2 (opportunities to speak & variety of approaches) represent conceptually different aspects of discussion. While Factor 1 is largely related to the *quality* of classroom talk, we see Factor 2 as more closed aligned with the *quantity* of classroom talk because it includes the amount of student talk that is elicited and the number of approaches available for students to talk about. Factor analysis eliminated the collinearity between elements of the discussion rubric and reduced our observation data to allow us to look for relationships between classroom discussion and students' performance on our achievement test.

What Aspects of Mathematics Classroom Discussion Affected Students' Performance?

The intraclass correlation (ICC) indicates that 12% of the variation in LMMA scores lies between classes and thus, HLM is an ideal method for analyzing this data, the results of which are presented in Table 5. Comparison of model deviance statistics shows that the slopes of CST vary across classes, and accounting for this variation increases the precision of the model. The coefficient on Total Discussion is significant, indicating that overall class discussion scores have a significant relationship with students' LMMA test performance. The significantly lower deviance between a reduced model without discussion scores and Model 4 with discussion scores indicates that classroom discussion contributes to explaining the between-class variation in LMMA scores above and beyond that explained by prior mathematics achievement.

Table 5. *HLM Predictions of LMMA Performance (Z-Scores)*

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Fixed Effects												
CST		0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Total Discussion			0.03	0.04*	0.15*							
			(0.02)	(0.02)	(0.08)							
CST* Discussion					-0.00							
					(0.00)							
Variety of Approaches						0.22**						
						(0.08)						
Explanations							0.05					
							(0.08)					
Opportunity to Speak								0.10				
								(0.08)				
Equitable Participation									0.21*			
									(0.10)			
Connection Between Ideas										0.08		
										(0.07)		
Factor 1											0.04	
											(0.06)	
Factor 2												0.11*
												(0.05)
Constant	0.00	-3.06***	-3.44***	-3.80***	-5.77***	-3.90***	-3.22***	-3.40***	-3.89***	-3.32***	-3.11***	-3.09***
	(0.09)	(0.20)	(0.39)	(0.39)	(1.19)	(0.35)	(0.28)	(0.32)	(0.43)	(0.29)	(0.22)	(0.21)
Random Effects												
Var(Residual)	0.88	0.50	0.50	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
Var(Con)	0.12	0.063	0.058	0.25	.17	0.23	0.26	0.26	0.30	0.29	0.27	0.20
Var (CST)				.000003	.000002	.000003	.000004	.000004	.000004	.000004	0.00000	0.00000
Cov (CST,Con)				-.0009	-.0007	-.0009	-.0009	-.0009	-.0009	-.001	-0.0010	-0.0008
Model Fit												
AIC	1317.9	913.3	914.1	907.9	907.0	904.6	911.5	910.2	907.8	910.7	911.5	907.6
BIC	1330.4	929.4	934.2	936.0	931.1	932.7	939.6	938.4	935.9	938.8	939.6	935.7
Deviance	1311.9	905.3	904.1	893.9	895.0	890.6	897.5	896.2	893.8	896.7	897.5	893.6

Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, ***

The absence of a statistically significant interaction effect demonstrated by model 5 indicates that class discussion impacts students exhibiting low prior achievement and high prior achievement similarly. To further illustrate this, we graphed each student on a scatter plot by locating each student's CST score in 2012 and LMMA score and plotted the line of best fit for the group of students in classrooms that received discussion scores below 15, and another line for the group of students in classrooms receiving discussion scores of 15 and above. The slope of the lines in Figure 1 demonstrate the predicted correlation between students' LMMA scores and their CST scores, showing that students who did well on one test tended to do well on the other. The dashed line shows that students in the 10 classrooms receiving discussions scores above 15 tended to have higher scores on the LMMA than students in the other classes, when controlling for students' prior mathematical achievement.

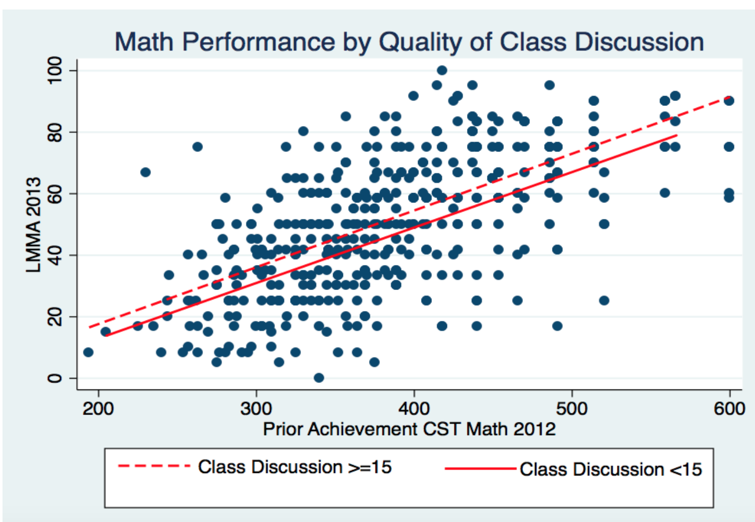

Figure 1. Lines of best fit relating prior achievement (2012 CST) with LMMA score by class discussion

In continuing the HLM analysis we found the deviance of model 11 is significantly lower than the deviance of the same model without Factor 1 (χ^2 (1*df*) = 73.5, $p < .001$). Therefore, including Factor 1 (connections & explanations) in the model produces a significantly better fitting model even though the coefficient is not statistically significant. The coefficient for Factor

2 (opportunities to speak & variety of approaches) is statistically significant at the $p < .01$ level. Factor 2 explains approximately 18.6% of the between class variation in LMMA scores and together with CST scores explains approximately 50% of the between class variation. The effect size is .11, which is small by Cohen's (1992) standards, but not trivial. Therefore, in general, we would expect an increase in LMMA score of .11 standard deviations (approximately 2.5 percentage points) for every one-point change in Factor 2 (opportunities to speak & variety of approaches) for a student in the average class. However, this should be interpreted with caution because we cannot assume that each point increase on the discussion rubric represents the same "distance." However, our data support a general trend that higher rubric discussion scores are associated with higher LMMA scores.

Discussion

What Aspects of Mathematics Classroom Discussion were Observed?

The observational data indicated high mean scores for "opportunities to speak" and "equitable participation." This demonstrated that students in these 20 3rd and 4th grade classrooms in high poverty schools had opportunities to speak, and most students publicly displayed their knowledge in some way. Furthermore, our observations indicated that students tended to be in well-managed classrooms where most students were on task, actively engaged in classroom activity. In 16 of the classrooms, students spoke frequently, and in 6 of these, student contributions were consistently substantive, comprising several sentences, while in the other 10 classrooms, some contributions were substantive. Only one observation involved students providing one- to two-word answers to teachers' prompts with no other opportunities to speak. This result was encouraging because, historically, the majority of classrooms in high poverty schools have tended to be either "dysfunctional" or "orderly and restrictive," with students rarely having opportunities to voice their thoughts (Knapp, Shields, & Turnbull, 1995). When teachers offer students the opportunity to speak, they are placing faith in the students to respond appropriately and to make valuable contributions. We found it promising that the classes were more interactive than didactic, in contrast to many high poverty schools (Diamond, 2007).

The high mean for multiple approaches was also promising. In Stein et al. (1996), only 50% of the observed lessons included multiple strategies. In the present study, 90% of the lessons included multiple strategies. Seeing several ways to solve a problem provided students with different access points, and some strategies were more concrete than others, so students at various points along the developmental trajectory could make sense of at least one of the strategies they saw (Murata, 2012). Our factor analysis indicated that opportunities to speak and multiple strategies tended to co-vary indicating that the teachers had adopted a "cumulative talk" approach (Mercer & Sams, 2006) in which students talk about different approaches to solving problems.

Most of the lessons observed entailed equitable participation with more than half of the lessons receiving a top score in this category. This result indicates that the students in this district were being asked to publicly display their thinking either through various hand signals, holding up an answer, agreeing or disagreeing), using whiteboards to show their answers and work, or verbally sharing. We thought it appropriate in a district with a large number of EBs to measure student participation in this way. While we would hope that the majority of students would contribute verbally to group discussion, we did not require this to take place for lessons to achieve the highest score because lower proficiency EBs may be hesitant to speak to a large audience.

The discussion elements "connections" and "conceptual explanations" occurred less frequently than the other three discussion elements, and the factor analysis indicated that scores on these two elements co-varied. We therefore combined them into a single facet, "exploratory talk." The rarity of exploratory talk is consistent with findings from the research cited above (Kazemi & Stipek, 2001; Spillane & Zeuli, 1999; Stein & Lane, 1996). In their review of the literature on classroom discussion, Walshaw and Anthony (2008) note that several studies have shown that teachers who successfully elicit students' thinking are not always sure how to lend structure to it. Fraivillig, Murphy, and Fuson (1999) found that extending children's mathematical thinking in the form of "analyzing, comparing and generalizing" (p. 167) was rare in the first-grade classes that they studied, because the teachers were not ready to "relinquish intellectual authority" and lacked a robust image of teaching that would involve more conceptual explanations and connections. They suggested that this challenging form of teaching took a great deal of time to develop. In a more focused study of three classrooms, Otten and Soria (2014) found that teachers had difficulty engaging their middle-school students in a discussion of connections among approaches and that students' explanations tended to stay at the level of procedural descriptions. Given the challenges of promoting conceptual explanations and helping students to make connections, we found it hopeful that we saw at least some glimmers of this activity in the classrooms we observed.

What Aspects of Mathematics Classroom Discussion Affected Students' Performance?

We used HLM modeling to determine the relationship between classroom discussion and performance on our achievement measure because it allowed us to analyze the effect of a whole-class variable on students' individual performance and control for students' prior achievement, which tends to be highly predictive of students' performance on subsequent tests of achievement (Hemmings, Grootenboer, & Kay, 2011). The analysis indicated that students in classrooms receiving higher scores on our discussion measure tended to perform better on the LMMA than students in classrooms with lower discussion scores. As noted above, two of the discussion facets, opportunities to speak and variety of approaches, on our measure co-varied and together explained the variation in achievement related to class discussion. Although these are not the only important facets of discussion, our data indicate that being in a classroom in which math discussion includes multiple approaches to solving problems and frequent, sustained opportunities to speak may offer benefits on measures of written achievement.

The few studies that have linked classroom activity to student achievement have demonstrated similar results. Stein and Lane (1996), for example, found that students at schools with access to multiple strategies performed better on an achievement test than students at a school where students were expected to learn specific procedures. Stein and Lane's research involved a population similar to the population in this study, including large numbers of students of color living in poverty. Their students were in 6th to 8th grade. Our findings extend theirs by showing that urban elementary students can engage in discussions of multiple strategies and that higher mathematics achievement is associated with this practice. Finally, while the effect sizes in this study were modest, they were close to the mean effect sizes that Seidel and Shavelson (2007) found in their meta-analysis of correlational research on the effects of teaching on student learning in mathematics.

Limitations

Our study, like Stein and Lane's (1996), used a test of our own design. This was necessary because we wanted to measure students' abilities to make sense of and solve open-ended problems, and could not find a measure that had eliminated the unnecessary linguistic complexity that increases measurement error for EBs (Abedi & Lord, 2001). Our work to simplify the linguistic demands of our test made it more accessible to EBs and the teachers who participated in the study reported to us that they considered our test to be a better measure of their students' abilities than the state test in use at the time of our study. Had we employed a more widely-used test, we would have been able to compare our students' performance with that of other students outside the district, but it might not have been sensitive to the learning we were trying to measure.

We also acknowledge that our results might not generalize to other similar districts. Three aspects of our context may be atypical. First, some of the teachers had been involved in a professional development effort, which supported them in using discussion in their classrooms. Districts without such a professional development effort might not have the same level of discussion we observed. Second, the *enVision* text adopted in the district might have supported discussion more than other texts. Each lesson had components that encouraged discussion, such as open-ended questions. Other texts might lack similar components. Classrooms where teachers have less autonomy over what and how they teach may face different challenges implementing class discussion.

In addition, in contrast to many schools serving students in poverty, the district had a stable veteran teaching force. Given their experience, teachers in our study might be more successful managing their classrooms, thus facilitating discussion. Finally, we do not claim that discussion caused students' improved performance on our measure. Some other aspect of the classrooms could explain our findings. For example, classrooms receiving higher discussion scores could have teachers who have stronger social bonds with their students. The HLM model we used did control for students' prior achievement but other aspects of the classroom population or the teachers' work with the students could be responsible for our results.

We acknowledge that our observation instrument does not include a specific category evaluating the cognitive demand of the tasks. We are aware that the openness of a mathematics problem allows for opportunities to discuss and connect students' solution strategies (Featherstone et al., 2011). Although the cognitive demand of the task is embedded into the components of our instrument, we cannot report on the specific level of the tasks teachers presented to students.

Lastly, a limitation to our study is that we only observed a snapshot of the classroom on a particular day. If the teachers presented a different task on a different day, the quality of the discussions could have been different. The extent to which this is true is constrained by the classroom norms that were established and maintained throughout the year. The responsibility of publicly sharing mathematical ideas requires a slow release of responsibilities from the teacher to the students over time (Hufferd-Ackles et al., 2004). Therefore, we assumed classrooms with regular discussions would use talk moves (Kazemi & Hintz, 2013) and argue about correctness of strategies (Lampert, 1990) more readily than classrooms that prioritized teacher-talk. Likewise, in classrooms where discussions are rare, it would be extremely unlikely for the teacher to be able to orchestrate a high-level discussion, eliciting thorough explanations and connections from students, if they are not accustomed to doing so.

Conclusion

Despite the limitations noted above, we found it promising that discussion scores were associated with higher performance on our achievement measure. While few of the classrooms we observed had **all** of the aspects of discussion mathematics education researchers have been calling for, we were encouraged to observe students discussing a variety of approaches to problems. We suggest that this is a first step in achieving the reform vision that has remained elusive for so long. We hope that this study can become part of a collection of studies showing similar results so that policy makers and administrators might be persuaded that children in classrooms with an emphasis on discussion perform better on written measures of mathematics achievement. Results such as those presented here provide support for reform efforts such as the Common Core Standards that emphasize the importance of students' sharing their ideas with one another.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234.

Akkus, R., & Hand, B. (2010). Examining teacher's struggles as they attempt to implement dialogical interaction as part of promoting mathematical reasoning within their classrooms. *International Journal of Science and Mathematics Education*, 9, 975-998.

Banes, L. C., Ambrose, R. C., Bayley, R., Restani, R. M., & Martin, H. A. (2018). Mathematical Classroom Discussion as an Equitable Practice: Effects on Elementary English Learners' Performance. *Journal of Language, Identity & Education*, *17*(6), 416-433.

Baxter, J. A., Woodward, J., & Olson, D. (2001). Effects of reform-based mathematics instruction on low achievers in five third-grade classrooms. *The Elementary School Journal*, 529-547.

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal*, *113*(1), 76-104.

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A Comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator*, *3*(2), 154-175.

Charles, R. I., Fennell, R., Caldwell, J., Cavanagh, M., Sammons, K., Schlielack, J. *et al.* (2009). *enVision Math.* Glenview, IL: Pearson Scott Foresman.

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models*. Advances in instructional psychology, 5, 161-238.*

Cirillo, M. (2013). What does research say the benefits of discussion in mathematics class are? *Research Brief*. National Council of Teachers of Mathematics.

Cobb P., Stephan, M., McClain, K., & Gravemeijer, K. (2001). Participating in classroom mathematical practices. *Journal of the Learning Sciences, 10*(1&2), 113-163.

Cobo, P. & Fortuny, J. (2000). Social interactions and cognitive effects in contexts of area-comparison problem solving. *Educational Studies in Mathematics*, *42*(2), 115-140.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Corkin, D. S., Coleman, S. L., & Ekmekci, A. (2019). Navigating the Challenges of Student-Centered Mathematics Teaching in an Urban Context. *The Urban Review*, *51*(3), 370-403.

Crespo, S., & Harper, F. K. (in press). Learning to pose collaborative mathematics problems with secondary prospective teachers. *International Journal of Educational Research*. https://doi.org/10.1016/j.ijer.2019.05.003

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

Evans, S, & Swan, M. (2014). Developing students' strategies for problem solving in mathematics: The role of pre-designed "Sample Student Work". *Educational Designer*, *2*(7), 1–31.

Featherstone, H., Crespo, S., Jilk, L. M., Oslund, J., Parks, A., & Wood, M. (2011). Smarter together! Collaboration and equity in the elementary math classroom. Reston, VA: National Council of Teachers of Mathematics.

Forman, E. A., McCormick, D., & Donato, R. (1998). Learning what counts as a mathematical explanation. *Linguistics and Education, 9*(4), 313–339.

Fraivillig, J. L., Murphy, L. A., & Fuson, K. C. (1999). Advancing children's mathematical thinking in everyday mathematics classrooms. *Journal for Research in Mathematics Education*, *30*(2), 148-170.

Gallimore, R., Hiebert, J. & Ermeling, B. (2104). Rich classroom discussion: One way to get rich learning. *Teachers College Record.* http://www.tcrecord.org ID Number: 17714.

Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice A new approach to analyzing classroom observation data. *Educational Researcher*, *44*(5), 263–277.

Heinze, A, Marschick, F., Lipowsky, F. (2009). Addition and subtraction of three-digit numbers: Adaptive strategy use and the influence of instruction in German third grade. *ZPM Mathematics Education, 41*(5)*,* 591–604.

Hemmings, B., Grootenboer, P., & Kay, R. (2011). Predicting mathematics achievement: The influence of prior achievement and attitudes. *International Journal of Science and Mathematics Education, 9*(3), 691-705.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, *28*(5), 524–549.

Herbel-Eisenmann, B. A., Steele, M.D., & Cirillo, M. (2013). (Developing) teacher discourse moves: a framework for professional development. Mathematics Teacher Educator, 1(2), 181-196. https://www.jstor.org/stable/10.5951/mathteaceduc.1.2.0181

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65-97). New York: Macmillan.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning*, *1*, (pp. 371–404). Charlotte, NC: Information Age.

Hiebert, J., & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal*, *30*(2), 393–425.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Hufferd-Ackles, K., Fuson, K. C., & Sherin, M. G. (2004). Describing levels and components of a math-talk learning community. *Journal for Research in Mathematics Education, 35*(2), 81-116.

Hunter, R. (2008). Facilitating communities of mathematical inquiry. In *Navigating Currents and Charting Directions* (Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia), edited by M. Goos, R. Brown, and K. Makar, Vol. 1, 31–39. Brisbane: MERGA.

Jones, I, Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgment. *International Journal of Science and Mathematics Education*, 13(1), 151–177.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Kazemi, E., & Hintz, A. (2014). *Intentional talk: How to structure and lead productive mathematical discussions*. Portland: Stenhouse Publishers.

Kazemi, E., & Stipek, D. (2001). Promoting conceptual thinking in four upper-elementary mathematics classrooms. *Elementary School Journal, 102*(1), 59-80.

Kline, P. (2000) *Handbook of Psychological Testing* (2nd ed.) London: Routledge.

Knapp, M., Shields, P. M., & Turnbull, B. J. (1995). Academic challenge in high poverty classrooms. *Phi Delta Kappan, 76*, 770–776.

Kosko, K., & Miyazaki, Y. (2012). The effect of student discussion frequency on fifth-grade students' mathematics achievement in U.S. schools. *Journal of Experimental Education, 80*(2), 173–195. doi: 10.1080/00220973.2011.566588

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

Mathematics Learning Study Committee. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology*, *104*(1), 1-21.

McEldoon, K. L., Durkin, K. L., & Rittle-Johnson, B. (2013). Is self-explanation worth the time? A comparison to additional practice. *British Journal of Educational Psychology*, *83*(4), 615–632.

Mercer, N. (1996). The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction*, *6*(4), 359-377.

Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education*, *20*(6), 507-528.

Merritt, E. G., Rimm-Kaufman, S. E., Berry, R. Q., Walkowiak, T. A., & Larsen, R. A. (2011). The contribution of mathematics instructional quality and class size to student achievement for third grade students from low income families. Paper presented at Fall Conference of *Society for Research on Educational Effectiveness, Washington, D.C.*.

Moschkovich, J. (2013). Principles and guidelines for equitable mathematics teaching practices and materials for English language learners. *Journal of Urban Mathematics Education*, *6*(1), 45-57.

Murata, A. (2012). Diversity and high academic expectations without tracking: Inclusively responsive instruction. *Journal of Learning Sciences, 22*(2), 312–335.

National Council of Teachers of Mathematics (2000). Principles and Standards for School Mathematics. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation *standards for school mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2014). Principles to actions: Ensuring mathematical success for all. Reston, VA: Author.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education*, *15*(5), 625-632.

O'Connor, C., Michaels, S., Chapin, S., & Harbaugh, A. G. (2017). The silent and the vocal: Participation and learning in whole-class discussion. *Learning and Instruction*, *48*, 5-13.

Otten, S., & Soria, V. M. (2014). Relationships between students’ learning and their participation during enactment of middle school algebra tasks. *ZDM*, *46*(5), 815-827.

Pasta, D. J. (2009). Learning When to Be Discrete: Continuous vs. Categorical Predictors. SAS Global Forum 2009. ICON Clinical Research, San Francisco, CA.

Planas, N., & Gorgorió, N. (2004). Are different students expected to learn norms differently in the mathematics classroom?. *Mathematics Education Research Journal*, *16*(1), 19-40.

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

Richland, L. E., Begolli, K. N., Simms, N., Frausel, R. R., & Lyons, E. A. (2017). Supporting mathematical discussions: the roles of comparison and cognitive load. *Educational Psychology Review, 29*(1), 41-53.

Rittle-Johnson, B., Saylor, M., & Swygert, K. E. (2008). Learning from explaining: Does it matter if mom is listening? *Journal of Experimental Child Psychology*, *100*(3), 215–224.

Sawada, D., Piburn, Judson, E., M., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics Journal*, 102, 245–253.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*:], 454-499. *doi:10.3102/0034654307310317*

Siegler, R.S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Silver, E. A., Ghousseini, H., Gosen, D., Charalambou, C., & Font Strawhun, B. T. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions to problems in the mathematics classroom. *Mathematics and Behavior*, *24*, 287–301.

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.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, *2*(1), 50-80.

Stein, M. K., Correnti, R., Moore, D., Russell, J. L., & Kelly, K. (2017). Using theory and measurement to sharpen conceptualizations of mathematics teaching in the common core era. *AERA Open*, *3*(1), 2332858416680566.

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.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1–19.

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education*, *5*(4), 541-542.

Walkowiak, T., Berry, R.Q., Meyer, J.P., Rimm-Kaufman, S. E. & Ottmar, E.R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics, 85*, 109128.

Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research*, *78*(3), 516-551.

Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research*, *63*, 79–93.