3-2020

Developing an Understanding Procedures Observation Rubric for Mathematics Intervention Teachers

Angela R. Crawford
Boise State University

Evelyn S. Johnson
Boise State University

Yuzhu Z. Zheng
Boise State University

Laura A. Moylan
Boise State University

Developing an Understanding Procedures Observation Rubric for Mathematics Intervention Teachers

Angela R. Crawford*
Boise State University
Angelacrawford1@u.boisestate.edu

Evelyn S. Johnson
Boise State University

Yuzhu Z. Zheng
Boise State University

Laura A. Moylan
Boise State University

This research was funded by IES Award No. R324A150152.

Abstract

This study describes the initial psychometric evaluation of the Understanding Procedures observation rubric for use as an instrument for feedback to teachers working in mathematics intervention settings. The rubric translates the research base from mathematics education and special education into practice in the form of specific items and descriptors of performance levels. A sample of 16 intervention teachers across three states provided three videos each of their instruction of students in mathematics intervention classes. Ten external raters evaluated the videos. We analyze the ratings using many-facet Rasch measurement. Analyses of the teacher, item, rater, and lesson facets show good psychometric quality for the instrument. Implications for research and professional development are discussed.

The Every Student Succeeds Act of 2015 and the recent position paper of the National Joint Committee on Learning Disabilities (Gartland & Strosnider, 2017) reaffirm the commitment to ensuring that all students receive high-quality education that enables them to develop higher-order skills. Yet studies have shown a persistent mathematics achievement gap between students with or at-risk for mathematics difficulty (i.e., students who experience difficulty with mathematics as a result of disabilities or other factors, SMD) and their peers (NCES, 2015; Schulte & Stevens, 2015), and this gap increases as SMD progress through secondary grades (Geary, Hoard, Nugent, & Bailey, 2012; Wei, Lenz, & Blackorby, 2013). This persistent, widening achievement gap suggests SMD are not receiving the instruction they need to achieve high-quality education standards related to mathematics, including developing understanding, fluency, reasoning, and problem solving.

A body of research has identified specific mathematics instructional practices that are effective in supporting SMD in achieving these goals (e.g., Fuchs, Malone, Schumacher, Namkung, & Wang, 2016; Gersten, Chard, et al., 2009). These findings have been shared with practitioners through numerous practice guides (e.g., Gersten, Beckmann, et al., 2009; VanDerHeyden & Allsop, 2014). Yet, a synthesis of observation studies of mathematics instruction found these practices are rarely used in classrooms (McKenna, Shin & Ciullo, 2015).

One way to increase implementation of research-based practices may be the use of teacher observation instruments designed for intervention settings (i.e., settings in which SMD are receiving additional instruction to remediate specific areas of difficulty or support development of proficiency). Such instruments should provide reliable and accurate feedback to teachers that is aligned with content-specific practices found to improve achievement for SMD (Johnson & Semmelroth, 2014). Although there are observation instruments that focus on mathematics instruction in general

education settings (e.g., MQI, Learning Mathematics for Teaching Project, [2011]; RTOP, Sawada, et al., [2002]), these instruments are not fully aligned with practices identified as most effective for SMD in interventions. There is a need for observation instruments focused on effective practices for mathematics instruction in intervention settings.

To meet this need we have developed four mathematics observation rubrics (Crawford, Johnson, Moylan, & Zheng, 2018): Understanding Concepts, Understanding Procedures, Practice, and Problem Solving. Our goal has been to translate research to actionable practice with a balance of specificity and flexibility. Items describe specific teacher actions to support instruction for mathematical understanding, fluency, and processes. Yet, the rubrics have flexibility to be relevant for SMD across contexts, content, and curricula.

Identifying Instructional Practices for Mathematics Intervention

While there remain considerable tensions between the fields of mathematics education and special education (Munter, Stein, & Smith, 2015), observation instruments for mathematics interventions, if they are to be useful, should comprise practices found to be effective by both research communities. To identify commonalities, we began with a review of research syntheses, meta-analyses, and practice recommendations for mathematics instruction and students with learning disabilities (e.g., Gersten, Beckmann, et al., 2009; Siegler et al., 2010; VanDerHeyden & Allsop, 2014; Woodward et al., 2012). We reviewed the empirical studies that were the bases for these recommendations and also conducted searches of more recent literature through databases using key terms identified from these studies. (For additional information about rubric development and a complete list of references, see Crawford et al., 2018; Johnson, Crawford, Moylan, & Zheng, 2018)

From these sources, we identified the features of instruction that were indicated as effective for supporting mathematics achievement for SMD. We developed matrices for several mathematical domains and broad goals, such as algebra, fractions, conceptual understanding, and fluency. Through a process of coding, clustering similar practices, and noting overlap across clusters, we identified the following inter-connected instructional practices with strong support from both fields of research that could be translated into items on rubrics.

Conceptual and Procedural Understanding

We define conceptual understanding as rich knowledge of mathematical properties and relations and procedural understanding as comprehension of, flexibility with, and using judgment to apply procedures (Star, 2005). Depth of understanding is related to the degree of connectedness among concepts and procedures (Baroody, Feil, & Johnson, 2007). Mathematics interventions should be focused on developing understanding of concepts and procedures, and interrelations between them should be made clear for students (e.g., Gersten, Chard, et al., 2009; Rittle-Johnson, Siegler, & Alibali, 2001). Depending upon the needs of the students, this can be accomplished through methods such as connecting to prior knowledge, using contexts and visual representations, explicit discussions, think-alouds, and calling attention to links that are otherwise implicit in student work (e.g., Fuchs et al., 2016; Fuson, 1990; Selling, 2016).

Systematic Instruction

In mathematics, systematic instruction is purposeful and carefully sequenced. Systematic instruction provides scaffolds and a range of examples that is responsive to the needs of the students (e.g, Kalyuga 2007). An example of a systematic approach with strong research support in both special education and mathematics education is a concrete-representational-abstract progression (e.g., Fennema et al., 1996; Misquitta, 2011).

Student Engagement

A high level of student engagement is an important component of mathematics instruction for SMD. Student engagement takes many forms: verbalizing understanding, modeling, demonstrating connections between ideas, providing choral or individual responses, frequent opportunities for practice, etc. (e.g., Doabler et al., 2015; Fuchs et al., 2016).

Visual Representations

Visual representations support students' abilities to connect mathematical concepts with procedures and provide scaffolds for problem solving (e.g., Fuson, 1990; Jitendra et al., 2015). Visual representations include number lines, manipulatives, informal drawings, graphs, etc. They should align with the structure of the concept, and connections within and between representations should be clear to students (e.g., van Garderen & Montague, 2003).

Accurate, Precise, and Meaningful Language

Teachers should use accurate and precise mathematical language that is aligned to students' receptive vocabulary (e.g., Riccomini, Smith, Hughes, & Fries, 2015). Academic language should be used consistently to support students' abilities to understand and use the terminology appropriately.

In addition to these instructional practices, SMD benefit from structures that support their ability to complete complex tasks. These structures, which include mnemonics, organizers, and guiding questions, support students with completing multi-step problems, organizing information, monitoring thinking when problem solving, or recalling facts (e.g., Jitendra et al., 2015; Montague, Enders, & Dietz, 2011). Therefore, we identified this as a sixth instructional practice that is integrated into the rubrics as a means to support students' independence when using mathematics.

Development of Items and Rubrics

Through analysis of ways these practices were enacted in research and described in practice guides, we translated the practices into specific teacher actions. This was an iterative process of returning to the source documents to determine how the practice was instantiated, drafting an item, comparing language of the item to the source documents, testing with video, eliciting subject matter expert input, and revision. As a result, we found that the items clustered more coherently around the goals of developing conceptual understanding, developing understanding of procedures, providing for effective student practice, and word problem solving than they did around domains such as algebra or fractions. These items can compose one large rubric that applies to instruction over multiple days or can be separated into rubrics focusing on each of those goals, with several items appearing in more than one rubric.

Table 1 shows how the items focused on understanding of procedures align with the instructional practices described previously. Eleven of the 17 items are directly related to connecting conceptual and procedural understanding. Five items are related to student engagement, and five are related to language use. Four items are related to systematic instruction. Two items are focused on visual representations, and one item is related to use of heuristics and cognitive strategies. [Insert Table 1].

Psychometric Issues for Observation Instruments for Instructional Practices

Observation instruments that support teachers' abilities to implement effective practices should consist of items that accurately and reliably distinguish levels of performance and provide concrete guidance for improvement (Hill & Grossman, 2013). This suggests the need for high-inference instruments that include quality indicators rather than instruments that only provide frequency counts. Previous studies of high-inference observation instruments have indicated that many factors contribute to variance in scores, suggesting that multiple facets of observation systems (items, teachers, lessons, and raters) should be investigated (Hill, Charalambous, & Kraft, 2013). Variance associated with items should be related to difficulty rather than construct-irrelevant factors. Variance associated with teachers is desirable to distinguish levels of performance, while variance associated with raters and lessons (whether due to lesson content or time of occurrence) would restrict the reliability of the instrument.

However, it has been reported that the instructional dimensions of observation instruments are the most challenging for raters to score reliably, and issues persist despite efforts to improve reliability, such as increased training and calibration requirements (Casabianca, Lockwood, & McCaffrey, 2015; Cash, Hamre, Pianta, & Myers, 2012; Mantzicopoulos, French, Patrick, Watson, & Ahn, 2018). Acknowledging that raters will differ in their severity but can be trained to be consistent in their own scoring may be a more attainable reality (Eckes, 2011). To investigate this issue with high-inference observation instruments, variance associated with raters can be separated from teacher performance scores and item and lesson difficulty scores with many-facet Rasch measurement (MFRM) (Eckes, 2011; Linacre, 1994).

This study reports on an initial psychometric analysis of the items that compose the Understanding Procedures rubric (see Crawford et al., 2018). This rubric was selected for initial testing because the greatest number of videos in our video library map to a goal of developing procedural understanding. The Understanding Procedures rubric contains 17 items that operationalize instructional practices on a 3-point scale, where 3 is proficient implementation, 2 is partial implementation, and 1 is ineffective or no implementation. (Raters can further divide partial implementation into three levels: 2+, 2, or 2-.) To conduct the analysis, we use MFRM, a method that explores the multiple sources of variance, or facets in teacher observations (teachers, lessons, items, and raters). Each facet is characterized by parameters, "ability" in the case of teachers, "difficulty" in the cases of lessons and items, and "severity" in the case of raters. MFRM calculates reliability within each of these parameter estimates and measures of model fit for each element within the facet. These data provide information about the rubric's ability to reliably distinguish performance across teachers and the functioning of items as intended.

Purpose of the Current Study

Observation rubrics need accompanying evidence for the validity of their intended use. The purpose of this study is to use MFRM analyses to examine the psychometric quality of the Understanding Procedures rubric for use as a measure of implementation of research-based instruction in mathematics for SMD. Therefore, the research questions are:

1) To what extent does the Understanding Procedures rubric reliably distinguish performance across teachers?
2) To what extent does the rubric reliably evaluate performance across multiple lessons?
3) Which items, if any, are influenced by construct-irrelevant variance?
4) What are the levels of interrater agreement and intra-rater consistency?

Method

Participants

Two groups, mathematics intervention teachers and raters, participated in the study. We obtained institutional review board approval and enforced protections for both groups.

Mathematics Intervention Teachers. Sixteen teachers from three states provided three video-recorded lessons for a total of 48 videos. These teachers were part of a larger data collection effort for the (Recognizing Effective Special Education Teachers) RESET project. Inclusionary criteria included providing mathematics instruction in an intervention setting, either small group or one-to-one. One teacher declined to give personal demographic information. Fifteen teachers were female, teaching from 2nd to 8th grade levels. Two of the teachers were Asian, and 13 teachers were White. Their years' experience ranged from 3 to 36. Seven teachers had bachelor's degrees, and eight teachers had master's degrees. All teachers held degrees in education, and three participants held additional endorsements specific to special education. Nine of the teachers reported designing lesson plans using Common Core standards or district frameworks. Three teachers reported using a curricular program based on explicit instruction that emphasizes conceptual understanding. Four teachers reported using a curricular program that employs direct instruction with spiraling content.

Raters. Ten raters, one male and nine females, from six states participated in this study. We used a purposive sampling technique to recruit raters with strong knowledge of mathematics instruction, special education instruction, and teacher observation. One rater was Asian, one was African, and eight were White. All raters were education professionals with between 3-20 years of working experience. Two raters held bachelor's degrees, six held master's degrees, and one held a doctoral degree. At the time of the study, four raters worked as classroom teachers, five were in doctoral degree programs, and one worked as special education faculty at a university.

Procedures

Video Collection. Over the course of a school year, teachers provided 15-20 video-recorded lessons, ranging in length from 18-50 minutes, from a consistent instructional period using the Swivl ® capture system with a single camera tracking the teacher and a microphone carried by the teacher. From this video bank, we selected three videos from each teacher with content applicable to the Understanding Procedures rubric and adequate sound and video quality. Videos were assigned an ID number and randomized to control for order effects.

Rater Training. Rater training involved four, three-hour sessions conducted by project staff plus additional video viewing and scoring time. We gave an overview of project goals and described how the mathematics instruction rubrics were developed. We also provided raters with a training manual that included in-depth descriptions of the items, definitions of quality descriptors, and exemplars of performance levels. We explained each item in the Understanding Procedures rubric using a model video and clarified any questions raters had about the items.

Then, over the course of the training, raters watched and scored three videos of mathematics instruction that presented a range of teaching quality. Raters were asked to include time-stamped evidence and explanations for their ratings. We shared master scored rubrics and led discussions of how the scoring criteria apply to those videos. We reviewed and discussed any disagreements, clarifying the intention of the items and rationales for master scores. Though we monitored agreement, we did not establish minimum rater performance standards. Other studies have found that, despite establishing minimum performance standards, raters still account for large portions of variance, and issues, such as drift, persist (Cash et al., 2012; Jones, 2019; Kane & Staiger, 2012.) Therefore, we focused on supporting raters in establishing understanding and consistency demonstrated through rationales supported by evidence. MFRM allows us to investigate the internal consistency of the raters and adjusts parameters of items, teachers, and lessons for discrepancies in severity of raters.

Upon completion of the training, raters were assigned a randomly ordered list of videos. We created a rating scheme that required each rater to score 21 of the 48 total videos but allowed for connection of ratings across raters and teachers (Eckes, 2011). We asked raters to submit scores and evidence for each item electronically via Qualtrics ®. We reminded raters to consult the manual as they completed their observations and allowed six weeks to complete ratings.

Data Analysis

Data were analyzed using the three-point scale. Rasch models transform ordinal data to an interval scale. A rating-scale model was used for the MFRM analyses in this study, given by:

$$\ln\left(\frac{P_{nijok}}{P_{nijo(k-1)}}\right) = B_n - D_i - C_j - T_o - F_k$$

where P_{nijok} is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*. $P_{nijo(k-1)}$is the probability of teacher *n*, when rated on item *i* by judge (rater) *j* on occasion (lesson) *o*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_j is the severity of judge *j*, T_o is the stringency of the occasion *o*, and F_k is the difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011).

The MFRM analyses were conducted using the program FACETS, version 3.67.1 (Linacre, 2017). FACETS produces logit measures that place facets on the same scale of measurement, allowing for comparisons among them. Also, the analyses produce infit and outfit statistics for each facet, indicating the degree to which the observed score matches the expected score produced by the model. An outfit mean square value is the average of the standardized residuals between observed and expected scores, while infit is calculated so that scores at ends of the scale are weighted less heavily. Fit values greater than one show more variation than expected, and values less than one show less variation than expected. Ranges in fit statistics from +/- .5 to 1.5 are considered acceptable (Eckes, 2011; Linacre, 2014).

In addition, FACETS provides separation statistics for each facet. The separation ratio compares the spread of scores to their precision and is used to mathematically determine the number of distinguishable strata in the data. The reliability of separation is computed as the ratio of "true" variance to the observed variance and indicates the

reproducibility of the facet measure if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007; Eckes, 2011). A chi-square statistic tests the null hypothesis that there is no statistically significant separation in the population (Eckes, 2011).

Results

MFRM analyses place facets on the same scale, and these results are displayed in Figure 1. The scale along the left, titled "Measr," is the logit measure for the elements within each facet. The scale, estimated from the pattern of the data, ranged from -2 to +2. A score of 0 indicates the mean level of performance or difficulty. Teachers with a logit measure higher than an item's logit measure had a greater probability of receiving a high score for that item than teachers with a lower logit measure. Category statistics show that 19% of scores were 3-implemented, 57% were 2-partially implemented, and 24% were 1-not implemented. [Insert Fig 1 here]

Teachers

The first research question asked to what extent the rubric reliably distinguished performance across teachers. MFRM provides three methods for exploring this question through logit measures, fit, and separation information (see Table 2). First, comparisons of teachers' observed and fair logit measures (scores adjusted to account for rater severity) indicate whether teachers are rank ordered reliably. There were minimal differences between the observed and fair scores with no difference in the rank ordering of teachers. [Insert Table 2]

Second, the fit statistics measure the extent to which each teacher's pattern of performance matches that predicted by the model, and therefore can be used to identify teachers whose performance has not been evaluated consistently. Table 2 shows that all fit statistics were within the acceptable range, suggesting that evaluation with the rubric was consistent in determining teachers' levels of implementation of instruction focused on understanding procedures.

Third, the strata statistic indicated nine distinguishable levels of performance with a statistically significant reliability of .98, χ^2 (15) = 744.9, $p < .001$. This reliability is comparable to coefficient alpha (Engelhard & Wind, 2018). This indicates that these teachers differed in the extent of their implementation of the practices beyond what can be attributed to error.

Lessons

The second research question asked to what extent the Understanding Procedures rubric reliably evaluated performance across multiple lessons. For each teacher, the lesson number indicates the order in which the lesson was recorded. Figure 1 shows that each of the three lessons were approximately the same difficulty, ranging from -0.7 to 0.8 logits. Fit statistics for the lesson facet were all close to 1, indicating good model fit. Separation statistics indicated two levels of difficulty with reliability of .62, χ^2 (2) =5.6, p=.06. Thus, the "difficulty" level across the three lessons was not statistically significant.

Items

The third research question asked which items, if any, were influenced by construct-irrelevant variance. MFRM provides two methods for exploring this question. First is the logical ordering of the item difficulties (Smith, 2001). The second column in Figure 1 orders the items by difficulty. Items on which teachers tended to receive low scores are considered more difficult than items on which teachers tended to receive high scores. We tallied scores for the most and least difficult items. Items 5 and 17 were the most difficult. Item 5 addresses reviewing or teaching of key vocabulary and mathematical symbols. Few teachers fully implemented this, with 36% of possible responses scored as 1-not implemented and 55% scored as 2-partially implemented. Item 17 examines the mathematical quality of feedback and was frequently rated as not present or not linked adequately to mathematical reasoning or concepts, with 35% scored as 1-not implemented and 54% scored as 2-partially implemented.

Items 1 and 2 were the least difficult. Item 1 examines whether the content of the lesson is focused on understanding of critical procedures. It is logical that this item would result in only 7% scored as 1-not implemented because videos were selected to test items related to understanding of procedures. Item 2 is related to use of visual representations that align to the mathematical structure of the content. For this item, 38% of scores were rated as 3-implemented while only 16% were scored as 1-not implemented.

Another method for examining construct-related variance is to evaluate the fit statistics. Item fit statistics (see Table 3) indicate whether raters have applied items in a manner consistent with the model. Items outside the acceptable range for fit may need to be removed or revised. The fit statistics for all items were within the acceptable range. Separation statistics indicated almost six strata with statistically significant reliability of .95, χ^2 (16) = 288.8, $p < .001$, demonstrating that the 17 items were separated along a continuum of difficulty. [Insert Table 3]

Raters

The last research question addresses rater agreement and consistency. The rater column in Figure 1 ranks the raters from most severe (Rater F) to most lenient (Rater G). The figure shows that four of the 10 raters (Raters C, D, E, and H) had logit measures that were very similar and could be considered interchangeable. Overall exact rater agreement was 51%. Table 4 provides exact agreement for each rater. The percent of exact agreement across raters ranged from 46.7%-54.1%. Separation statistics indicated four strata with a significant reliability at .92, χ^2 (9) = 115.5, $p < .001$, demonstrating that raters differed in their severity. [Insert Table 4]

The fit statistics reported in Table 4 indicate the degree to which the observed ratings matched those predicted by the model and can be used to identify raters whose scores do not fit the overall pattern. Raters who fit the model are demonstrating intra-rater consistency. All raters are expected to demonstrate some degree of error, but too much will threaten the validity of the measurement process (Bond & Fox, 2007; Linacre, 1994). The fit statistics for raters were within the acceptable range, with one rater approaching the edge of the range with 1.48.

Discussion

This analysis of the Understanding Procedures rubric was conducted to provide psychometric evidence for its use as a measure of implementation of research-based practices for mathematics intervention instruction. The results indicate that the Understanding Procedures rubric can provide reliable assessments of implementation and useful guidance to teachers.

Implications of the Analyses for Assessing Reliability

Separation of teachers into nine strata with good reliability indicates that the rubric can distinguish between levels of performance. While there are few items to precisely measure the "ability" level of the teachers at the upper and lower ends of the scale, this is not necessarily problematic. The rubric is intended to measure implementation of research-based practices rather than precisely define teacher "ability". The separation of items into almost six levels of difficulty indicates the rubric provides the ability to deliver feedback to teachers on items that range from easier to more difficult to implement. A lack of separation for lessons suggests that the difficulty did not vary systematically based on the order in which they occurred.

Fit statistics for all facets indicate that the observed ratings match the expected ratings produced by the model and suggest that the analysis was not unduly influenced by construct-irrelevant factors (Eckes, 2011). The item difficulty ranking provides information to assess the degree of construct validity of an instrument through the evaluation of the logic of the order of the items (Smith, 2001). The rank order of items on this rubric is logical; there is a general pattern of more difficult items requiring more content knowledge of the teacher to be implemented fully and effectively. The "easier" items are related to the lesson topic and visual representations and may demand less mathematical content knowledge of the teachers because many standards and curricula provide these features. However, this pattern may not hold for the most difficult item, effectively reviewing or teaching key vocabulary or symbols. Over a third of possible scores for this item were at the "not implemented" level. Although content knowledge may affect one's ability to effectively implement this item, the lack of implementation may be due to other factors such as skipping review to save time.

An important consideration with observation instruments is consistency among the raters. Our results found statistically significant differences in severity. Other studies have also reported variance associated with raters despite more rigorous training and calibration requirements (Casabianca et al., 2015; Cash et al., 2012; Jones, 2019; Kane & Staiger, 2012) than those we employed. Also, our levels of exact agreement, around 50%, are consistent with those reported across other studies (Casabianca et al., 2015; Cash et al., 2012; Kane & Staiger, 2012). Differences in raters' knowledge or beliefs about mathematics instruction may account for these differences in severity. For example, item 3 is related to the conceptual background knowledge and skills that support understanding of procedures. Variations

in severity in this item may be related to raters' depth of knowledge about mathematics or beliefs about what is important to understand in order to perform a procedure successfully. It will be important to consider the role of content knowledge and beliefs in future studies of raters' application of the items.

Despite this, fit statistics are within acceptable ranges, indicating that overall raters' applications of items to teachers' performances were consistent. These data, coupled with similarities in teachers' observed and fair scores, indicate that while some raters were more lenient than others, they were rank ordering teachers similarly.

Taken together, the results of the analyses of facets indicate that the Understanding Procedures rubric can serve as a reliable tool for characterizing a teacher's implementation of research-based mathematics instructional practices for SMD.

Implications for Translating Research into Practice

The Understanding Procedures rubric is an instrument that translates research from mathematics education and special education into practice in the form of specific teacher actions and performance level descriptors. Our study finds that implementation of research-based practices was most often some degree of partial implementation (57% of scores) or ineffectively or not implemented (24% of scores). Just over two-thirds of the sample were rated with an "ability" level at or below the mid-point of the logit scale. This indicates that those teachers had high probability of receiving a low score on half of the items in the rubric. Also, three teachers had a high probability being rated as ineffectively or not implementing most items in the rubric. With only 19% of the total scores given a 3-implemented rating, our sample does not include many examples of implementation of practices to support understanding of procedures. These results are consistent with other observation studies which have documented the research-to-practice gap (Mantzicopoulos et al., 2018; McKenna et al., 2015).

Addressing this gap will require accurate assessment and specific feedback to support implementation. Because the rubric includes levels of performance, rather than a simple checklist stating whether or not practices are present in a lesson, it can provide teachers with more specific guidance, i.e., what they are doing well, what needs improvement, and what is missing from their instruction. With the ability to distinguish levels of performance and good reliability and model fit, this rubric can provide accurate and useful information with which to characterize levels of implementation. Also, good model fit suggests we were successful in creating an observation instrument that maintained specificity with flexibility. The specificity of the rubric comes from detailing effective instruction as specific teacher actions characterized at three levels of implementation. The flexibility is reflected in the application of the rubric to videos with varied content, settings and, curricula.

Limitations and Future Research

Although the results are promising, the differences in severity across raters merits further investigation. This is especially true considering that each item was scored on a three-point scale, and over half of the scores were rated at the level of 2-partially implemented. Reducing rater error will be critical in achieving the goal of an instrument that can provide highly reliable and accurate ratings across samples of teachers. Further investigations may focus on the extent to which raters used criteria for assigning scores that were consistent with the scoring protocol. By further studying teachers' performances and raters' rationales for scores, it may also be possible to refine the scale and provide more fine-grained distinctions of partial implementation.

Though results show strong reliability, the sample size of teachers who provided video is small, warranting caution in generalization. Exploratory work such as this with small samples can be performed with Rasch measurement, though recommendations for stable estimates are typically 30 per parameter (Wright & Stone, 1999). As we continue to develop a video bank, we can conduct studies with larger samples to verify the results reported here.

The goal of the Understanding Procedures rubric is to bridge the research-to-practice gap in instruction of mathematics procedures when working with SMD. Future research plans with the rubric include examining its impact as a formative assessment to guide improvements in implementation of research-based practices. Following baseline evaluation and throughout the schoolyear, teachers can set goals for implementation and receive feedback with the rubric. Also, we plan to investigate the use of the four mathematics observation rubrics with a complete unit of instruction.

Another goal for the observation rubrics is to connect teacher performance to student growth and to examine the relative contributions of each of the instructional practices at the item level. This would allow for professional development efforts to focus on those elements of instruction which have the most impact on the mathematics achievement of SMD. This can also contribute to an emerging body of research on the relative effectiveness of specific components of instruction (e.g., Doabler et al., 2015).

For decades, the mathematics achievement of SMD has remained behind that of their peers. One potential explanation for the continued lower achievement of SMD is that research-based practices are either not implemented or are not implemented effectively to realize the positive effects reported in the literature (McKenna et al., 2015). This is consistent with what we have observed while developing the RESET observation system. Although most teachers are doing their best to serve SMD well, there is a disconnect between instructional practices as implemented in the classroom with what is described in the research base. If we are to bridge this research-to-practice gap and improve mathematics outcomes for SMD, there is a critical need for observational systems that align targets for high quality mathematics instruction with observations of teachers who deliver these practices.

References

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (2018). Research-based practices in mathematics instruction for students with disabilities rubrics. Recognizing Effective Special Education Teachers (RESET). Boise State University: Boise, ID. Accessible at: https://education.boisestate.edu/reset/rubrics/.

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education*, 115-131.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Chicago: Institute for Objective Measurement.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542.

Doabler, C. T., Baker, S. K., Kosty, D. B., Smolkowski, K., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *The Elementary School Journal, 115*(3), 304-333.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement.* Frankfurt: Peter Lang.

Engelhard, J. G., & Wind, S. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. London: Taylor and Francis.

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education*, *27*(4), 403-434.

Fuchs, L. S., Malone, A., Schumacher, R., Namkung, J. & Wang, A. (2016). Fraction intervention for students at-risk for mathematics learning disabilities: Lessons learned from five randomized control trials. *Journal of Learning Disabilities. 50*(6), 631-639.

Fuson, K. C. (1990). Conceptual structures for multiunit numbers: Implications for learning and teaching multidigit addition, subtraction, and place value. *Cognition and Instruction*, *7*(4), 343-403.

Gartland, D. & Strosnider, R. (2017). Learning disabilities and achieving high-quality education standards. *Learning Disability Quarterly, 40*(3), 152-154.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology, 104*(1), 206-223.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools* (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U. S. Department of Education. Retrieved from the NCEE website: https://ies.ed.gov/ncee/wwc/PracticeGuides.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*(3), 1202-1242.

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

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371-384.

Jitendra, A. K., Peterson-Brown, S., Lein, A. E., Zaslofsky, A. F., Kunkel, A. K., Jung, P.-G., & Egan. A. M. (2015). Teaching mathematical word problem solving: The quality of evidence for strategy instruction priming the problem structure. *Journal of Learning Disabilities*, *48*(1), 51-72.

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2018). Using Evidence-Centered Design to create a special educator observation system. Educational Measurement: Issues and Practice. 37(2), 35-44.

Johnson, E., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it matters, what makes it challenging, and how to address these challenges. *Assessment for Effective Intervention*, *39*(2), 71-82.

Jones, N. (2019, February). Observing special education teachers in high-stakes teacher evaluation systems. Presentation given at the Pacific Coast Research Conference, Coronado, CA.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509-539.

Kane, T. J., & Staiger, D.O. (2012). Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. (Report). Retrieved from: http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-2/.

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

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago: University of Chicago Press.

Linacre, J. M. (2014). *A user guide to Facets.* Chicago: Winsteps.com.

Linacre, J. M. (2017). *Facets 3.76.1* [Computer software].

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the Framework For Teaching and the Classroom Assessment Scoring System, *Educational Assessment, 23(*1), 24-46.

McKenna, J. W., Shin, M., & Ciullo, S. (2015). Evaluating reading and mathematics instruction for students with learning disabilities: A synthesis of observation research. *Learning Disability Quarterly*, *38*(4), 195-207.

Misquitta, R. (2011). A review of the literature: Fraction instruction for struggling learners in mathematics. *Learning Disabilities Research & Practice*, *26*(2), 109-119.

Montague, M., Enders, C., & Dietz, S. (2011). Effects of cognitive strategy instruction on math problem solving of middle school students with learning disabilities. *Learning Disability Quarterly*, *34*(4), 262-272.

Munter, C., Stein, M. K., & Smith, M. A. (2015). Dialogic and Direct Instruction: Two Distinct Models of Mathematics Instruction and the Debate (s) Surrounding Them. *Teachers College Record*, *117*(11), 1-32.

National Center for Education Statistics (NCES). (2015). *The condition of education 2015* (NCES 2015-144). Washington, DC: U.S. Department of Education, Authors.

Riccomini, P. J., Smith, G. W., Hughes, E. M., & Fries, K. M. (2015). The language of mathematics: The importance of teaching and learning mathematical vocabulary. *Reading & Writing Quarterly*, *31*(3), 235-252.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346-362.

Sawada, D., Piburn, M. D., Judson, E., 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*, *102*(6), 245-253.

Schulte, A. C., & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children*, *81*(3), 370-387.

Selling, S. K. (2016). Making mathematical practices explicit in urban middle and high school mathematics classrooms. *Journal for Research in Mathematics Education 47*(5), 505-551.

Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., & Wray, J. (2010). *Developing effective fractions instruction for kindergarten through 8th grade: A practice guide* (NCEE #2010-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U. S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/PracticeGuides.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, *2*(3), 281-311.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 404-411.

van Garderen, D., & Montague, M. (2003). Visual-spatial representation, mathematical problem solving, and students of varying abilities. *Learning Disabilities Research & Practice, 18*(4), 246.

VanDerHeyden, A., & Allsopp, D. (2014). *Innovation configuration for mathematics* (Document No. IC-6). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/tools/innovation-configuration/

Wei, X., Lenz, K. B., & Blackorby, J. (2013). Math growth trajectories of students with disabilities: Disability category, gender, racial, and socioeconomic status differences from ages 7 to 17. *Remedial and Special Education*, *34*(3), 154-165.

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., Koedinger, K. R., & Ogbuehi, P. (2012). *Improving mathematical problem solving in grades 4 through 8: A practice guide* (NCEE 2012-4055). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U. S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/PracticeGuides.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials.* Wilmington, DE: Wide Range.

Table 1

Alignment between instructional practices identified in review of research and items focused on developing understanding of procedures.

Item	Combining conceptual and procedural understanding	Implementing systematic instruction	Encouraging high levels of student engagement	Using visual representations	Language that is accurate, precise, and meaningful to students	Using heuristics and cognitive strategies
1	x					
2	x			x		
3	x	x				
4	x					
5		x				
6		x				
7	x					
8	x				x	
9	x		x	x		
10		x			x	
11					x	
12	x				x	
13	x		x			
14			x		x	
15	x		x			
16			x			x
17	x					

Table 2

Teacher Measure Report from Many-Facet Rasch Measurement Analysis

Teacher Number	Difficulty (Logits)	Model *SE*	Infit MNSQ	Outfit MNSQ	Observed Average	Fair Average
1	1.37	.10	1.09	1.06	2.4	2.45
15	.85	.12	1.10	1.11	2.3	2.28
5	.58	.12	1.01	1.01	2.2	2.19
11	.56	.12	.82	.82	2.2	2.19
2	.40	.10	.89	.89	2.2	2.13
6	.02	.12	1.00	1.00	2.0	2.01
9	-.02	.12	1.09	1.09	2.0	1.99
4	-.05	.12	1.42	1.42	2.0	1.98
8	-.20	.12	1.05	1.04	1.9	1.94
12	-.26	.12	.95	.95	1.9	1.91
13	-.34	.12	.86	.86	1.9	1.89
14	-.44	.12	.97	.97	1.9	1.85
10	-1.10	.13	1.08	1.08	1.6	1.64
3	-1.13	.10	.89	.89	1.6	1.63
7	-1.20	.13	.92	.92	1.6	1.60
16	-1.73	.13	.92	.94	1.5	1.44
Mean	-.17	.12	1.00	1.00	1.9	1.95
SD	.83	.01	.14	.14	.3	.27

Note. Root mean square error (model) = .12; adjusted *SD* = .82; separation =6.83, strata =9.44, reliability = .98; fixed $\chi^2 = 744.9$; $df = 15$; $p < .001$.

Table 3

Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Difficulty (Logits)	Model *SE*	Infit *MNSQ*	Outfit *MNSQ*
5	.72	.12	1.19	1.20
17	.62	.12	1.04	1.03
3	.45	.12	1.25	1.24
13	.42	.12	.89	.88
14	.37	.12	1.14	1.14
12	.37	.12	.91	.90
9	.37	.12	.79	.79
7	.31	.12	.97	.97
15	.03	.12	1.05	1.05
16	-.13	.12	.92	.93
10	-.14	.12	1.07	1.07
8	-.19	.12	.82	.82
11	-.29	.12	.92	.92
6	-.36	.12	.84	.84
4	-.62	.12	1.26	1.25
2	-.83	.12	1.09	1.08
1	-1.12	.12	.87	.86
Mean	.00	.12	1.00	1.00
SD	.52	.00	.15	.15

Note. Root mean square error (model) = .12; adjusted *SD* = .51; separation = 4.17; strata = 5.89; reliability = .95; fixed $\chi^2 = 288.8$; $df = 16$; $p < .001$.

Table 4

Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater	Severity (Logits)	Model *SE*	Infit *MNSQ*	Outfit *MNSQ*	Exact Obs Agreement %
F	.47	.09	1.04	1.05	51.2
B	.43	.09	.76	.77	54.1
J	.18	.09	.97	.96	52.4
A	.07	.09	1.48	1.48	48.1
C	-.01	.09	.61	.62	55.2
E	-.02	.09	1.15	1.15	52.0
D	-.04	.09	.98	.98	53.4
H	-.04	.09	1.28	1.26	47.5
I	-.41	.09	1.00	1.00	46.7
G	-.64	.09	.72	.72	51.6
Mean	.00	.09	1.00	1.00	
SD	.34	.00	.26	.26	

Note. Root mean square error (model) = .09; adjusted *SD* = .32; separation = 6.13; strata = 4.93; reliability = .92; fixed $\chi^2 = 115.5$; $df = 9$; $p < .001$.

Measr	-Items	+Teacher	-Rater	-Lesson	Scale
2					(3)
					———
		Teacher 1			
1					
		Teacher 15			
	5				
	17	Teacher 11, Teacher 5			
			Rater F		
	3, 9, 12, 13, 14	Teacher 2	Rater B		
	7				
			Rater J		
			Rater A	Lesson 1	
0	15	Teacher 6, Teacher 9	Rater C, Rater D, Rater E, Rater H	Lesson 3	2
	10, 16	Teacher 4		Lesson 2	
	8	Teacher 8			
	11	Teacher 12, Teacher 13	Rater I		
	6	Teacher 14			
			Rater G		
	4				
	2				
-1					
	1	Teacher 10, Teacher 3			———
		Teacher 7			
					———
		Teacher 16			
-2					(1)

Figure 1. Variable map of Understanding Procedures rubric facets: items, teachers, raters, and lessons.