3-2020

Examining Rater Accuracy and Consistency with a Special Education Observation Protocol

Evelyn S. Johnson
Boise State University

Yuzhu Zheng
Boise State University

Angela R. Crawford
Boise State University

Laura A. Moylan
Boise State University

Publication Information

Johnson, Evelyn S.; Zheng, Yuzhu; Crawford, Angela R.; and Moylan, Laura A. (2020). "Examining Rater Accuracy and Consistency with a Special Education Observation Protocol". *Studies in Educational Evaluation, 64*, 100827-1 - 100827-9. https://doi.org/10.1016/j.stueduc.2019.100827

Examining Rater Accuracy and Consistency with a Special Education Observation Protocol

Evelyn S. Johnson*
Department of Early and Special Education
Boise State University

Yuzhu Zheng
Project RESET
Boise State University

Angela R. Crawford
Project RESET
Boise State University

Laura A. Moylan
Project RESET
Boise State University

Author Note

Evelyn S. Johnson, Department of Early and Special Education, Boise State University; Yuzhu Zheng, Project RESET; Boise State University; Angela Crawford, Project RESET, Boise State University, Laura A. Moylan, Project RESET, Boise State University.

This research was supported the Institute of Education Sciences, award number R324A150152 to Boise State University. The opinions expressed are solely those of the authors. Correspondence regarding this manuscript should be addressed to: Dr. Evelyn Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

Abstract

Research indicates that instructional aspects of teacher performance are the most difficult to reach consensus on, significantly limiting teacher observation as a way to systematically improve instructional practice. Understanding the rationales that raters provide as they evaluate teacher performance with an observation protocol offers one way to better understand the training efforts required to improve rater accuracy. The purpose of this study was to examine the accuracy of raters evaluating special education teachers' implementation of evidence-based math instruction. A mixed-methods approach was used to investigate: 1) the consistency of the raters' application of the scoring criteria to evaluate teachers' lessons, 2) raters' accuracy on two lessons with those given by expert-raters, and 3) the raters' understanding and application of the scoring criteria through a think-aloud process. The results show that raters had difficulty understanding some of the high inference items in the rubric and applying them accurately and consistently across the lessons. Implications for rater training are discussed.

Keywords: rater accuracy; teacher observation; rater consistency; feedback; special education

Many students with disabilities (SWD) perform significantly below their peers in math on national and state level assessments (National Center for Education Statistics (NCES), 2015; Schulte & Stevens, 2015). These large achievement gaps tend to worsen as students move through later grades (Geary, Hoard, Nugent & Bailey, 2012; Wei, Lenz & Blackorby, 2013), suggesting that SWD are not receiving the math instruction they need to be successful. For nearly two decades there have been calls for *all* students to receive mathematics instruction that leads to strong conceptual understanding, procedural fluency, reasoning, and problem-solving abilities (Kilpatrick, Swafford, & Findell, 2001), and there is an emerging body of research that identifies mathematics instructional practices that effectively support SWD 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); however, a synthesis of observation studies reports that these practices are rarely used in classrooms (McKenna, Shin & Ciullo, 2015).

One way to increase the implementation of research-based practices is through the use of teacher observation instruments. In recent studies, teacher observation systems have been linked to improved practice and increased student achievement when evaluators use the observation system to provide teachers with clear and specific feedback (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Biancarosa, Bryk, & Dexter, 2010; Taylor & Tyler, 2012). Such instruments should be aligned with content-specific practices found to improve achievement for SWD so that teachers receive accurate and consistent feedback about how to improve their instructional practice (Johnson & Semmelroth, 2014). One example of an observation system designed specifically for special education teachers is the Recognizing Effective Special Education Teachers (RESET) observation system (Johnson, Crawford, Moylan & Zheng, 2018).

RESET Observation System

The RESET observation system consists of a set of observation protocols that are aligned with evidence-based instructional practices (EBPs) for SWD (Johnson, Crawford, Moylan & Zheng, 2018). RESET was developed using the principles of Evidence-Centered Design (ECD; Mislevy, Steinberg & Almond, 2003) and consists of 21 observation protocols that detail EBPs organized in three categories: a) instructional methods, b) content organization and delivery, and c) individualization. A series of studies to investigate the reliability and validity of the RESET observation system have been conducted and summarized in Johnson, Crawford, Moylan and Zheng (2019). Although there are observation instruments that focus on mathematics instruction in general education settings (e.g., Mathematical Quality of Instruction (MQI), Learning Mathematics for Teaching Project, [2011]; Reformed Teaching Observation Protocol (RTOP), Sawada, et al., [2002]), these instruments are not fully aligned with practices identified as most effective for SWD in interventions (Crawford, Johnson, Zheng & Moylan, 2019). There are four RESET observation protocols that depict EBPs in math instruction, including *Developing Conceptual Understanding, Developing Understanding of Procedures, Practicing Computation (Fluency),* and *Word Problem Solving* protocols. To date we have been able to examine the psychometric evidence for the *Developing Understanding of Procedures* protocol (see Appendix A), which has been demonstrated to have strong reliability (Crawford, et al., 2019).

The primary intended use of the RESET system is to provide teachers with feedback to improve their instruction. In the area of math instruction, the *Developing Procedural Understanding* protocol focuses on the knowledge of action sequences for solving problems (Rittle-Johnson & Alibabali, 1999). Students with greater conceptual understanding of math concepts tend to develop greater procedural skill (Cauley, 1988; Baroody & Gannon, 1984; Cowan, Dowker, Christakis, & Bailey, 1996; Byrnes & Wasik, 1991), and it has also been shown that students who develop strong procedural understanding can improve their conceptual understanding (Rittle-Johnson, Siegler, & Alibali, 2001).

An extensive research base identifies instructional practices that are effective for SWD who struggle to develop a strong understanding of procedures, including: 1) explicit instruction of the mathematical concept and procedures to support student understanding (Doabler, Baker, Kosty, Smokowski, Clarke, Miller & Fien, 2015; Gersten et al., 2009), 2) visual representations to support students' conceptual understanding and their ability to connect the math concept to the math procedures (Gersten et al., 2009; Witzel, Mercer & Miller, 2003), 3) cognitive strategy instruction that teaches students to identify and solve a variety of problem types (Griffin & Jitendra, 2009), 4) explicit inquiry routines to support students' strategic competence and ability to solve complex equations (Impecoven-Lind & Foegen, 2010; Scheuermann, Deshler & Schumaker, 2009), and 5) teaching students to verbalize their math reasoning and using mathematical vocabulary (Gersten, Chard, Jayanthi, Baker, Morphy & Flojo, 2009).

Many special education teachers however, lack a strong understanding of specific math pedagogical strategies for teaching SWD (DeSimone & Parmar, 2006; Faulkner & Cain, 2013; Gerretson & McHatton, 2009). Teachers of SWD are rarely taught how to represent and formulate math instruction in ways that support SWD's conceptual and procedural understanding, their strategic competence, and their ability to reason mathematically and make connections. As a result, many teachers of SWD have neither developed pedagogical content knowledge, nor the understanding of the strategies most likely to be successful in supporting the math learning needs of their students (Shulman, 1986). The theory of change that underlies the RESET observation system posits that through ongoing feedback aligned with the elements of EBPs in math instruction, teachers' practice will improve, and ultimately SWD will achieve improved outcomes. Given this intended use of RESET, the data informing the feedback must be robust; it must be both accurate and consistent with the intent of the elements of the EBPs detailed in the observation rubric (Mantzicopoulos, French, Patrick, Watson & Ahn, 2018).

Rater Mediated Assessments

RESET, like most teacher observation systems, relies on trained raters to apply the protocol's scoring procedures and criteria to record their judgement about the quality of a teacher's practice (Bell et al., 2014; Myford & Wolfe, 2003). In this way, the evaluation and feedback that teachers receive on RESET can be significantly mediated by rater effects. The major concern with rater-mediated assessments is the quality of ratings provided by the raters (Hamp-Lyons, 2007; Johnson, Penny, & Gordon, 2009; Wang, Engelhard, & Wolfe, 2016; Wesolowski & Wind, 2017; Wind & Peterson, 2018). Ratings are often associated with characteristics of raters and not necessarily with the performances themselves (Engelhard, 2002; Graham, Milanowski, & Miller, 2012). The rater-mediated problem is well-documented in the teacher observation research (McCaffrey, Yuan, Savitsky, Lockwood & Edelen, 2015; Rockoff & Speroni, 2009).

A number of teacher observation studies suggest that the instructional dimensions of observation protocols are the most challenging for raters to score reliably (Bell et al., 2014, Bill and Melinda Gates Foundation, 2011; Gitomer et al., 2014), with raters accounting for between 25 - 70% of the variance in scores assigned to the same lesson (Casabianca, Lockwood & McCaffrey, 2015). In fact, across studies of teacher observation using a variety of protocols, low levels of exact agreement (ranging from 47-52%) on the instructional dimensions of observation protocols have been reported (Bell et al., 2015; Cash, Hamre & Pianta, 2012; Jones, 2019; Kane & Staiger, 2012). In the field of special education, emerging evidence also suggests that raters with special education expertise tend to be better at rating special education instruction than those without (Lawson & Cruz, 2018), presumably because they are more familiar with the instructional practices being observed.

Two reasons that the rater issue is exacerbated when evaluating instruction include: 1) many observation protocols are *high inference* tools that place a significant demand on raters for processing performance information across a variety of settings and contexts, and 2) raters have their own ideas about high quality teaching and learning based on their own experiences (Gitomer et al., 2014). This is problematic because the potential for an observation system like RESET to be used effectively in the evaluation and improvement of special education teachers' math instruction depends not only on the validity of the observation protocol but also on the accuracy and consistency of the raters.

Rater Accuracy and Consistency

Rater accuracy can be defined as the match between the scores and rationales provided by raters with those provided by the expert raters (Englehard, Wang & Wind, 2018). Rater consistency is defined as the rater's ability to apply the scoring criteria and procedures in the same way across teachers and across lessons (Eckes, 2011). Accuracy matters because inaccurate scoring compromises the diagnostic function of observations and can prevent teachers from receiving the feedback they need to improve instruction (Hill & Grossman, 2013). Consistency is also critical, because teacher evaluation and feedback should be based on the consistent application of scoring criteria rather than halo effects, rater by teacher or rater by lesson interactions.

A number of approaches have been investigated to improve rater accuracy and consistency in observation systems. Methods to improve rater reliability and consistency such as increased training and calibration requirements have been investigated, but issues persist even as raters gain experience and with ongoing calibration efforts (Casabianca et al., 2015). Researchers have used different models to investigate the quality of raters, including models based on generalizability theory (Brennan, 1992; Crawford, Johnson, Moylan, & Zheng, 2019; Hill, Charalambous & Kraft, 2012), the rater bundle model (Wilson & Hoskens, 2001), hierarchical rater models (Patz, Junker, Johnson, & Mariano, 2002), Rasch measurement models (Myford & Wolfe, 2003, 2004; Engelhard, 1996, 2013; Johnson et al., 2018; Razynski, Engelhard, Cohen, & Lu, 2015), a signal-detection rater model (DeCarlo, Kim, & Johnson, 2011), and a latent trait model (Wolfe & McVay, 2012). Different indices are used to indicate rating quality such as rater agreement, rater error (inconsistency), and rater accuracy (Johnson, et al., 2009; Wind & Engelhard, 2013). Although these models and indices differ from a statistical perspective and can be used to address different research questions, all of them attempt to identify systematic patterns that may influence the rating quality from a quantitative approach (Wesolowski & Wind, 2017).

Psychometric approaches that focus on the raters' consistent application of the scoring criteria have been used to address issues such as halo effects and central tendency use (Wang, Engelhard, Raczynski, Song & Wolfe, 2017). Many-facet Rasch measurement (MFRM) modeling for example, is one approach that establishes a statistical framework that allows for the study of the influences of any number of rater biases on rating behavior (Engelhard, Wang & Wind, 2018; Myford & Wolfe, 2003). These approaches to improving rater behavior are helpful when considering the general set of raters and the need is to improve the consistency of scores across a set of evaluations.

Psychometric approaches are useful for understanding and "correcting" the ratings provided on teacher observation instruments designed to serve as an overall evaluation of teacher performance. However, when teacher observations are used to provide teachers with timely, formative feedback on their practice, it is generally not feasible to have multiple raters score the same lessons in order to statistically adjust for variation in rater effects (Hill & Grossman, 2013). Instead, raters need to be trained to use the observation protocol accurately so that the nature of a teachers' feedback is consistent with the instructional improvements they need to make to implement evidence-based instructional practices with fidelity. Without common frames of reference for raters to understand how specific instructional behaviors map to specific levels of performance on specific dimensions of the protocol, it remains challenging to improve rater accuracy and to realize the potential of teacher observation systems on a large scale.

Purpose of the Study

A teacher's scores on a RESET observation protocol are intended to represent how well the teacher can implement the evidence-based practice being observed, and to guide the feedback she/he needs to improve practice. In research conducted to date to examine the reliability of various RESET protocols including the *Developing Understanding of Procedures* protocol, exact levels of rater agreement have ranged from 50-52% (Crawford et al., 2019; Johnson et al., 2018; Johnson, Zheng, Crawford, & Moylan, 2018; Johnson, Moylan, Crawford, & Zheng, 2019). Although these ranges are consistent with other large-scale teacher observation studies, and can be "corrected" through statistical procedures for evaluation purposes, the low-level of agreement raises concern about the extent to which special education teachers will receive consistent and accurate feedback about how to improve their implementation of evidenced-based math instruction.

Whereas *more* rater training offers one approach to developing common understanding and application of the scoring criteria, in practice, extensive training to calibrate raters (typically school administrators), is not feasible given the resource constraints that most schools face. In practice, raters are generally not well-calibrated, and a growing research base demonstrates that school administrators tend to rate teachers more leniently (Kraft & Gilmour, 2017; Weisberg, Sexton, Mulhern & Keeling, 2009), even when rating teachers they do not know or supervise (Lawson & Cruz, 2018). Another approach to improving training efforts is to investigate the ways in which raters interpret and apply scoring criteria to inform specific adjustments that could be effective in improving rater accuracy and consistency (Joe, Tocci, Holtzman & Williams, 2013). Think-aloud exercises are one commonly used method to study rating behaviors (Suto, 2012). For example, think-aloud studies of rater reasoning have indicated that raters use various strategies other than using the observation protocol to rate a lesson depending on the context (Bell et al., 2014; Qi, Bell, Jones, Lewis, Weatherspoon & Redash, 2018). These strategies included things such as reasoning from something remembered from training or calibration videos, or using their own internal criteria (Bell et al., 2014; Qi et al., 2018).

Therefore, the purpose of this study was to investigate: 1) the extent to which raters are able to consistently represent the scoring criteria in the *Developing Understanding of Procedures in Math* protocol, 2) the raters' accuracy in scoring two math instruction lessons using the protocol, and 3) the raters' understanding and application of the scoring criteria through a think-aloud process.

Methods

The study reported in this manuscript was part of a larger effort to examine the functioning of the various facets of the *Developing Understanding of Procedures in Math* rubric using MFRM analysis (Crawford et al., 2019). Drawing on data collected to examine the psychometric properties of this rubric, here we report specifically on the rater facet to provide a measure of raters' internal consistency. We also present the additional data collected to further understand rater accuracy and the raters' application of the scoring criteria to evaluate special education teachers' instructional practice.

Participants

Special Education Teachers. Sixteen special education teachers from three states were recruited to provide video recorded lessons of their math instruction to SWD. One teacher declined to give personal demographic information. The remaining 15 participants were female, teaching in second to eighth grade level classrooms in a small-group intervention setting. Thirteen of 15 teachers were White, and two teachers were Asian. All teachers were certified special education teachers, and their teaching experience ranged from three to 36 years (*M* = 12.03, *SD* = 10.02). Each participating teacher provided three video recorded lessons of their math instruction which was verified by research project staff to align with the *Developing Understanding for Procedures* protocol, resulting in a total of 48 lessons ranging in length from 20 - 50 minutes. Lessons were video-recorded during the 2015-16 and 2016-17 school year using the Swivl ® capture and upload system. From this set of 48 lessons, we randomly selected two lessons to investigate rater accuracy.

Raters. Ten raters, one male and nine female, from six states participated in this study. Raters were recruited through a purposive sampling technique focused on selecting individuals with experience providing math instruction to SWD and experience with teacher observation. Eight raters were white, one was Asian-American, and one was African-American. All raters were special education professionals with between three to twenty years of experience. Two raters had a Bachelor's degree, six had a Master's Degree, and one had a Doctoral Degree. At the time of the study, four raters worked as special education classroom teachers, five were in special education doctoral degree programs with an emphasis on math instruction for SWD, and one worked as a special education faculty at a university where they teach graduate and undergraduate coursework in effective math instruction for SWD.

Expert Raters. Two female researchers from the RESET project team evaluated the two videos that were assigned to every rater. The expert raters were the primary developers of the RESET rubrics, which required synthesizing the research on the instructional practices included on the observation protocols. One expert rater had a doctoral degree in Curriculum and Instruction with a concentration in mathematics instruction, and eight years of experience teaching math to students with or at-risk for math related learning disabilities. The other expert rater was in her final year of a doctoral program in special education with a focus on evidence-based instructional practice and teacher observation, and had ten years of experience as an elementary teacher in an inclusive classroom, two years of experience as an instructional coach for special education teachers, three years of experience as an elementary school principal and two years of experience as a federal programs director.

Materials

Instructional Lessons for Accuracy. As noted, two lessons from the set of 48 were randomly selected to be scored by each participating rater. The first lesson was 26 minutes long and included a special education teacher working with four third grade students to teach them how to subtract with regrouping. In the lesson the teacher is using manipulatives with students to help them understand when and why regrouping is necessary and why the procedure works to solve problems. The second lesson was 32 minutes long and included a special education teacher working with eight, seventh grade students on prime factorization. In the lesson, the teacher used a white board to guide students through the procedures used to identify the prime factors of a number. Each rater was asked to score and conduct a think-aloud of their scoring process for each video. Video order was counterbalanced across raters to control for order effects.

Developing Understanding of Procedures Observation Protocol

The *Developing Understanding of Procedures in Mathematics* protocol (see Appendix A) was used in this study. This observation protocol is one of the content specific protocols of RESET, and consists of 17 items that detail the elements of developing students' understanding of the conceptual basis for a mathematical procedure and the reasons for the steps in the procedures. Each item is rated on a three-point scale (1 = Not implemented, 2 = Partially implemented, and 3 = Implemented) to evaluate a teacher's level of proficiency in implementing that specific item. The protocol is designed to be used by raters with high levels of expertise in math intervention. A recent study examining the protocol's psychometric quality through many-faceted Rasch measurement (MFRM) reported strong item, lesson, teacher, and rater fit statistics (Crawford, et al., 2019).

Procedures

Rater Training. All raters participated in a four-day rater training provided by the research team. Raters were first provided with an overview of the RESET project goals and a description of how the *Understanding Procedures* protocol was developed. Then, research project staff explained each item of the protocol using a model video and clarified any questions the raters had about the items. Raters were also provided with a training manual that included detailed descriptions of each item, along with examples for each item across each level of performance. Then, raters independently watched and scored another three videos that had been scored by project staff. The scores were reconciled with the master scored protocol for each video. Any disagreements in scores were reviewed and discussed. Rater agreement with the master scored training lessons averaged 55% across all items. Rater training did not include a certification threshold, as sufficient evidence has demonstrated that raters do not sustain high levels of agreement even with certification and calibration efforts (Casabianca, Lockwood & McCaffrey, 2015), and because in practice, most districts do not have the resources to invest in training raters to these standards.

Data Collection. A concurrent triangulation approach (Creswell, 2009) was used in this study, which involved collecting both quantitative and qualitative data to gain additional insight of raters' accuracy and consistency of using the observation protocol. The quantitative data was collected from the raters who were assigned a randomly ordered list of 21 videos (including the two common videos used to examine rater accuracy) and asked to evaluate the videos following the assigned order, to score each item, to provide time stamped evidence that they used as a basis for the score, and to provide a brief explanation of the rationale for their score. To maintain a feasible video observation load for each rater, we developed a rating scheme that would allow us to link scores across raters and videos without requiring each rater to score each video (Eckes, 2011). Raters were also reminded to consult the training manual as they completed their observations and were given a timeframe of six weeks to complete their ratings. Completed evaluations were submitted using an electronic version of the protocol developed in the Qualtrics ® survey system. The scores from this survey system were used for the quantitative analysis.

At the same time, qualitative data were collected from each rater through the think-aloud approach to investigate the accuracy and consistency of raters' understanding of the observation protocol. The think-aloud approach is defined as a "verbal report in which participants state their thoughts and behaviors" (Block, 1986, p.463), and has been widely used as a research tool to measure people's thinking processes. Each rater was asked to evaluate the two videos and to say aloud and record everything they thought while watching the video and scoring each item. Raters were asked to provide a rationale for their given score for each item in the think-aloud task, and to include evidence from the video recording.

Expert Rating the Two Videos. The two expert raters first independently scored the two videos. Inter-rater agreement between the two expert raters was 87% for the first lesson, and 81% for the second lesson. After independently scoring the videos, expert raters worked together to review the evidence and rationales for the assigned scores and to come to consensus on any items about which they disagreed. The consensus scores and rationales provided by the expert raters were used to evaluate accuracy.

Data Analysis

Rater Consistency. The scores assigned to the recorded lessons were analyzed through MFRM analyses. The model used for the MFRM analysis in this study is given by:

$$ln\left(\frac{P_{nirlk}}{P_{nirl(k-1)}}\right)= B_n - D_i - C_r - T_l - F_k$$

where P_{nirlk} is the probability of teacher *n*, when rated on item *i* by rater *r* on lesson *l*, being awarded a rating of *k*. $P_{nirl(k-1)}$ is the probability of teacher *n*, when rated on item *i* by rater *r* in lesson *l*, being awarded a rating of *k*-1, B_n is the ability of teacher *n*, D_i is the difficulty of item *i*, C_r is the severity of rater *r*, T_l is the stringency of lesson *l*, and F_k is the difficulty overcome in being observed at the rating *k* relative to the rating *k*-1 (Eckes, 2011).

The MFRM analysis was conducted using the computer program FACETS version 3.80 (Linacre, 2017). MFRM analysis produces infit and outfit statistics for each facet, two quality control statistics that indicate whether the measures have been confounded by construct-irrelevant factors (Eckes, 2011). Ranges in fit statistics from .5 to 1.5

are considered acceptable (Eckes, 2011; Engelhard, 1992). In addition to measures of fit, FACETS also provides reliability and separation indices. The reliability index indicates the reproducibility of the measures if the test were to be administered to another randomly selected sample from the same population (Bond & Fox, 2007). Separation indicates the number of statistically distinguishable strata in the data. In this analysis, we report on these results for the rater facet only. A full analysis of teacher, lesson and item analyses is reported elsewhere (Crawford et al., 2019).

Rater Accuracy. To compute rater accuracy, we compared each rater's score with the expert ratings on the two lessons that every rater scored. For each item, we indicated whether the rater agreed with the expert scored rubric, and then computed the overall percentage of items for which there was perfect agreement. We computed an accuracy score for each rater for each lesson.

Think-Aloud Analysis. The think-aloud data were first transcribed. Then, the score for each item given by each rater for each lesson and its corresponding rationale were sorted. First, we created five main categories in which each score and corresponding rationale was placed. These five categories included: 1) score and rationale were the same as the expert rater, 2) score was the same but the rationale was different from the expert rater, 3) score was different but the rationale was the same as the expert rater, 4) score and rationale were both different from the expert rater, and 5) no rationale for the score was provided.

Key words, short phrases, or sentences from the rater's explanation for the score for each item from the think-aloud transcripts were identified as the unit of analysis. Selective coding of these units was conducted to create a set of codes that described the ways in which the rationales provided by the raters differed from those provided by the expert raters. Finally, the data were organized by category and code and summarized both by rater and by lesson.

Results

To investigate the consistency of raters in their application of the scoring criteria and procedures, we first examined the statistics for the rater facet from the MFRM analysis (see Table 1). The rater severity ranges from -.64 logits (*SE* = .09) for Rater 7 who is the most lenient to .47 logits (*SE* = .09) for Rater 6 who is the most severe. The rater fixed chi-square value tests the assumption that all raters share the same severity measure, after accounting for measurement error. A significant value means that the severity measures of at least two raters included in the analysis are significantly different (Eckes, 2011). The raters differed in severity when evaluating the teachers (χ^2 =115.5, df = 9, p < .001). The rater separation ratio measures the spread of the rater severity measures relative to the precision of those measures. The closer the separation ratio is to zero, the closer the raters are in terms of their severity (Eckes, 2011). The rater separation ratio of 6.13 and the separation reliability of .92 further confirmed the variability in rater severity. As can be seen in Table 1, the fit statistics for the rater facet fell between .72 to 1.48, which is within acceptable range and indicates the consistency of each rater. However, the exact agreement across all assigned scores was 51%. In other words, raters differed in severity when compared to other raters, but each individual rater was able to apply scoring criteria and procedures in a consistent manner.

Rater Accuracy

The overall exact agreement between the raters and the expert raters across the two lessons selected for the think-aloud was 54%. Table 2 shows that overall, Rater 9 had the highest accuracy level at 74%, and Rater 1 had the lowest accuracy level at 38%. Raters' accuracy levels differed across lessons, with some raters scoring Lesson 1 more accurately and some scoring Lesson 2 more accurately. For example, Rater 9 was 100% accurate with the first lesson, but only 47% accurate with the second lesson. Rater 8 was 53% accurate with the first lesson and 35% accurate with the second lesson. Rater 1 was only 29% accurate with Lesson 1, and 47% accurate with Lesson 2. Table 2 also shows that the raters' accuracy by item by lesson varied considerably. The most extreme example is Item 1, *The lesson is **consistently** focused on understanding of critical mathematics procedures*, which all raters scored accurately for Lesson 1, but no raters scored accurately for Lesson 2. Item 8, *The teacher engages students in making connections between representations, meanings of operations, and procedures,* presented similar difficulties, with seven raters who scored this item accurately for Lesson 1, but no raters who scored it accurately for Lesson 2. No clear pattern across the two lessons emerged by item, although Item 6, *There is an explicit, systematic progression within and/or across lessons that supports understanding,* and Item 7, *The teacher **clearly** explains and **sufficiently** emphasizes the conceptual meaning of the procedure and/or operation,* appeared to be two of the most challenging items to score. The variability in item level accuracy suggests that rater accuracy varied depending on the specific lesson and specific element of instructional practice they evaluated.

Differences in Rater and Expert Rater Rationales

After examining the accuracy of the scores assigned by the raters, we analyzed their think-aloud data to compare the rationales they provided with those of the expert raters. The results are presented for each rater in Table 3. Approximately 46% of the scores and rationales were similar. For example, in Lesson 1, one rater gave the teacher a score of 3 on Item 2, *The teacher uses visual representations that support understanding of the procedure* because "Students were creating numbers using blocks. They were exchanging ten blocks for 10 ones in order to take away the correct number of ones. This is appropriate and helping students understand the quantities that are linked to numbers as well as place value, which supports the given score based on the rubric." The expert raters gave the same teacher a score of 3 for the same reasons.

An additional 8% of the items were scored accurately, but raters used rationales that differed from the expert raters to support their scores. For example, in Lesson 2, one rater and the expert raters each gave the teacher a score of 2 on Item 17, *Feedback is consistently linked to mathematical reasoning and concepts*, but gave different rationales for their scores. The rater gave a score of 2 because "The teacher didn't give the students enough feedback on how they were doing in class," whereas the expert raters gave the same score because "The teacher's feedback was not consistently linked to math reasoning."

In 4% of the responses, the raters used rationales that were similar to the expert raters but applied different scores. For example, in Lesson 2, one rater gave a score of 2 on Item 8, *The teacher engages students in making connections between representations, meanings of operations, and procedures*, because "The teacher did not make any connection." For the same reason, the expert raters gave a score of 1. In these cases, it is difficult to determine whether the scoring difference was a data input error, or a rater error in applying the scoring criteria to the item.

Approximately 40% of items had both scores and rationales that differed from the expert raters. For example, in Lesson 1, one rater gave a score of 3 on Item 9, *The teacher provides clear explanations for all of the mathematical reasons for the steps in the procedure*, because "The teacher provided clear explanations for all math reasons for the steps and the procedure." The expert raters gave a score of 2 because "The teacher explained the reasons, but the explanation is not always clear and direct." These discrepancies, in both scores and the way in which evidence is interpreted to arrive at the score, serve as the most useful starting point to guide future training efforts. Finally, approximately 1% of the scores did not include rationales for the scores applied. Although each individual rater had a different pattern of findings, the most common explanations (33% of coded rationales) for having scores and rationales inconsistent with experts were related to differences in how the actions in a lesson were mapped on to the items.

Given that the accuracy ratings for several items differed substantially across lessons, we aggregated the results by lesson to better investigate how raters' scores differed from expert raters depending on the context of the observed lesson (see Table 4). As can be seen in Table 4, there were more instances in Lesson 2 (13 counts) of raters citing the same evidence as expert raters but then applying a different score than there were for Lesson 1 (1 count). Additionally, in cases where both the score and rationale differed from the expert rater, differences about the quality of implementation were more prevalent for Lesson 2 (34 counts) than Lesson 1 (13 counts). Finally, raters had a more difficult time interpreting terms like "adequate" in the same way as the expert raters for Lesson 1 (51 counts) than in Lesson 2 (15 counts).

Discussion

As is the case with most teacher observation systems, the RESET observation protocols rely on raters whose subjectivity is unavoidably introduced into the evaluation process, which threatens the accuracy of the assigned ratings (Suto, 2012). Rater accuracy is not only the accuracy of performance ratings, but also the accuracy of rationales that raters provide to support their scores (Sulsky & Balzer, 1988). The results of the MFRM analysis suggest that although the raters differed in severity when evaluating teachers, each rater was internally consistent throughout the evaluation in this study. Some researchers have argued that rater variability is inevitable in complex performance assessments and raters typically cannot function interchangeably as expected even after extensive trainings (Eckes, 2011). Therefore, accounting for rater severity could be accomplished through statistical analyses such as MFRM, which allows for adjustments in assigned scores and computes a "fair average score" (Linacre, 2017).

It is highly improbable that a state or district however, has the capacity and resources to employ such methods in the context of teacher observation. Correcting scores for rater severity may be theoretically possible when observation protocols are being used for evaluation, but it is not feasible when scored protocols are used as formative assessments to support teachers in improving their practice. Therefore, rater consistency seems to be a necessary but insufficient quality if an intended use of observation protocols is to provide teachers with accurate feedback about how to improve their instruction.

It is unlikely that districts will have the capacity to invest in training observers to meet high levels of agreement with expert raters. Some researchers argue that rater accuracy is less important under low-stakes conditions when the observation-feedback is for improving instruction than it is in high-stakes conditions, when observation is evaluative (i.e., scores may result in consequences such as promotion or dismissal). We disagree. Inaccurate information about teacher performance can lead to ineffective or inappropriate feedback to teachers and prevent observations from promoting effective instructional changes. Accurate ratings of instruction are foundational to realizing the promise of observation systems to improve instructional practice at scale.

Achieving rater accuracy in both research and practice however, has remained an elusive goal. In the current study, raters had expertise in the content area, yet varied widely in their application of the scoring criteria. In practice, it is highly likely that special education teachers will be observed and provided with feedback by raters with significantly less expertise. Understanding how raters use observation protocols, apply scoring criteria and interpret evidence may improve training efforts that lead to more consistent understanding of the depicted instructional elements (Qi, et al., 2018). For example, short video clips used as exemplars across performance level descriptors for more complex items could be used for training and reference for raters.

Our results show that whereas raters were consistent within their own application of the scoring criteria, the exact agreement across all rater scores was only 51%. This finding was not unexpected given the low levels of agreement we were able to achieve during training (55%). Although many large-scale observation systems require raters to achieve a high level of agreement to be ‘certified’ to use the instrument, studies have indicated that certification and calibration efforts do not result in sustained high levels of agreement (Casabianca et al., 2015), and that raters tend to rely less on written scoring criteria and more on their initial perception of lesson quality over time as they score (Qi et al., 2018). Exact agreement with the expert raters across the two lessons was 54%. Examining the raters' rationales with those provided by the expert raters' show that the agreement on rationales for the exact scores was even lower, at 46%. This suggests that even when the raters gave the same scores as the expert raters, their understanding of the items and the teacher's performance differed from the expert raters. Therefore, even when raters agree on scores, the feedback provided to a teacher may vary if raters do not share the same understanding.

The results also show that it was difficult for raters to map different performances to some items consistently across lessons. The comparison of rationales indicated that the alignment for some items was strong in one lesson, but then compromised in the other lesson (e.g., Item 1). The variability in the accuracy levels for Item 1 (*The lesson is consistently focused on understanding of critical mathematics procedures*, see Table 2) across the two lessons highlights the challenges of rater accuracy. In Lesson 1 the teacher was focused on helping her students understand the process of regrouping for subtraction. Throughout the lesson, the teacher discussed with her students the importance of understanding why they were "trading tens for ones" and what it means to “trade”. The activities and the language the teacher used were aligned with "focused on understanding of critical math procedures". The expert raters and all ten raters agreed that this item, in this lesson, was proficiently implemented.

In contrast, in Lesson 2, the teacher was working with her students on prime factorization. The teacher used factor trees to show students how to go through the procedure. The teacher's instructions and questions were consistently related to "what is the next step in the process". There was no discussion of what prime factorization is and no use of models that conceptually show how a number can be composed of prime numbers, such as an area model. The lesson was a step-by-step explanation of using a factor tree (a procedure) to do prime factorization. While there is nothing wrong with using a factor tree to facilitate completing the procedure, the teacher did not support the link between conceptual understanding of the concept and the reason for the procedures used to solve a problem. The expert raters scored this item as a 1, not implemented, whereas seven raters scored it as fully implemented. This teacher could receive very different feedback about what to do differently to support her students' understanding of the procedure depending on who observes her teaching.

After examining raters' rationales for this item across the two lessons, one potential explanation for the difference in accuracy levels is that the content and representations that were used in Lesson 1 were much more familiar to raters and the teacher was clearly focused on developing conceptual understanding of the procedure. However, in Lesson 2, fewer raters may even be aware of the conceptual meaning of prime factorization or what representations might communicate the conceptual meaning. Instead, raters may have mistaken the teacher's focus on procedural facility for conceptual understanding of the procedure, concentrating on the fact that the lesson was focused, but less attuned to the criteria of teaching for understanding. Given the intent of the rubric, expert raters communicated that proficient implementation would have included the teacher using an area model to first conceptually explain prime factorization, and then connecting the area model to the factor tree for students. The larger point is that raters need to have a deep understanding of each item so that they can consistently apply it across a wide variety of contexts and lessons. Using examples like the one provided in Lesson 2 during training could be helpful in developing this deep understanding.

The *Developing Understanding of Procedures* protocol includes elements that are quite specific, but the variability with which raters interpreted a number of items and the differences in the degree to which they relied on evidence that was consistent with the expert raters' rationales is disconcerting. Although several teacher observation researchers have commented on the lack of shared understandings of quality teaching (Bell et al., 2014; Gitomer et al., 2014; Hill & Grossman, 2013), our study suggests that even when the elements of an instructional practice are highly detailed and grounded in a strong evidence-base, interpreting items across a variety of teachers and lessons, and consistently mapping performances to a set of scoring criteria remains a challenge. Analyses such as the one conducted for this study may help rater trainers create training tools and approaches that help clearly discriminate across levels of performance and across different contexts.

Finally, although it may not be new information that raters tend to view and evaluate instructional practice quite differently, our findings have important implications for the effective use of observation protocols like the *Developing Understanding of Procedures* protocol. It will take a significant amount of time and resources to create a cadre of observers who have common understandings of instructional practice. A growing research base indicates that a focus on improving knowledge for content specific pedagogy results in positive gains in teacher knowledge and instructional practice with the potential to impact student achievement (Brownell et al., 2017; Garet, Heppen, Walters, Smith & Yang, 2016; Griffin et al., 2018). Therefore, the investment in developing shared understandings of instructional practice through the use of observational protocols like RESET could be regarded as a professional development opportunity, focused on improving instructional practice to improve outcomes for SWD.

Limitations and Future Directions

In addition to the small sample of raters and teachers, which reduces the generalizability of our findings, there are two important limitations of this study. First, the coding system used to analyze raters' rationales and think-aloud processes was grounded in just two observations of a teacher's instructional practice. The general nature of the coding categories and their emphasis on the alignment of rater evidence with expert raters however, limit this concern. Second, we did not conduct more in-depth interviews with raters to gain more clarity and insight into their scoring procedures. There were times when the rationales provided by a rater were simply a restatement or paraphrased version of the performance level descriptor. This limited our ability to analyze potential reasons for their application of that specific scoring criteria. Follow up or cognitive interviewing may allow for stronger insight into a rater's thought process, but it is difficult to know if that insight would generalize across a larger group of raters.

Nevertheless, this study adds to the sparse literature examining rater behavior within teacher observation systems in important ways. Our analyses suggest multiple ways in which rater accuracy might be improved. First, clearer definitions and exemplars for items that are particularly challenging to rate should be provided. Over time, we have developed a more detailed training manual and are working to collect video examples that depict items that are problematic. This of course, is a time-consuming process, but will likely be needed as a way to develop common understandings of the instructional practices being observed. Second, short videos that demonstrate the difference in performance level descriptors of specific items could provide a helpful alternative to more descriptive text and examples for raters to distinguish the difference between "proficient implementation" and "partial implementation". Finally, rater training might require more individualized training to address the unique biases a rater brings to the evaluation process. Specific training could either focus on items that appear problematic for that rater, or it could focus on rater practices, such as routinely consulting the manual, or ensuring that evidence is directly connected to the performance level descriptors provided. All of these approaches unfortunately, are not quick fixes and will require a substantial amount of resources to implement. Rather than invest resources in raters, it could prove more effective and efficient to simply invest in improving teachers' understanding of EBPs, and to support teachers' ability to use the observation protocols aligned with EBPs to set goals, self-evaluate, and make progress towards those goals.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. S*cience, 333*(6045), 1034–1037

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., Gitomer, D. H., McCaffrey, D. F., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project.* San Francisco, CA: Jossey-Bass.

Biancarosa, G., Bryk, A. S., & Dexter, E. (2010). Assessing the Value-Added Effects of Literacy Collaborative Professional Development on Student Learning. *Elementary School Journal, 111*(1), 7-34.

Bill and Melinda Gates Foundation. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Author. Retrieved from https://files.eric.ed.gov/fulltext/ED540960.pdf

Block, E. (1986). The comprehension strategies of second language readers. *TESOL Quarterly, 20*(3), 463–494.

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

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Brownell, M., Kiely, M. T., Haager, D., Boardman, A., Corbett, N., Algina, J., … Urbach, J. (2017). Literacy learning cohorts: Content-focused approach to improving special education teachers' reading instruction. *Exceptional Children, 83*(2), 143-164.

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

Crawford, A., Johnson, E. S., Moylan, L. A, & Zheng, Y. Z. (2018). Variance and Reliability in Special Educator Observation Rubrics. *Assessment for Effective Intervention*, 1-11.

Crawford, A., Johnson, E. S., Moylan, L. A, & Zheng, Y. Z. (in press, 2019). Developing an understanding procedures observation rubric for mathematics intervention teachers. *School Science and Mathematics*.

Creswell, J.W. (2009). Research design: Qualitative, quantitative, and mixed approaches (3rd, ed.). Thousand Oaks, CA: Sage.

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333-356.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.

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

Engelhard Jr, G. (1996). Evaluating rater accuracy in performance assessment. *Journal of Educational Measurement, 33*(1), 56-70.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Engelhard, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling, 60(1),* 33-52.

Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Malone, A. S., Wang, A., ... & Changas, P. (2016). Effects of intervention to improve at-risk fourth graders' understanding, calculations, and word problems with fractions. *The Elementary School Journal, 116*(4), 625-651.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., ... & Borman, G. D. (2016). Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development. NCEE 2016-4010. *National Center for Education Evaluation and Regional Assistance*.

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.

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. *What Works Clearinghouse*.

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.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings*, Center for Educator Compensation Reform.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055-2100.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*(1), 1–9.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83(2),* 371-386.

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.

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of Observation: Considerations for Developing a Classroom Observation System That Helps Districts Achieve Consistent and Accurate Scores*. MET Project, Policy and Practice Brief. Bill & Melinda Gates Foundation.

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2019). Validity of a special education teacher observation system. *Educational Assessment.*

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. Z. (2018). Using Evidence-Centered Design to Create a Special Educator Observation System. *Educational Measurement: Issues and Practice*, 35-44.

Johnson, E. S., Zheng, Y. Z., Crawford, A., & Moylan, L. A. (2018). Developing an Explicit Instruction Special Education Teacher Observation Rubric. *The Journal of Special Education*, 1-13.

Johnson, E. S., Moylan, L. A, Crawford, A., & Zheng, Y. Z. (2019). Developing a Comprehension Instruction Observation Rubric for Special Education Teachers. *Reading & Writing Quarterly,* 1-19.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: The Guilford Press.

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

Kane, M. T. (2006) Validation. In R. L. Brennan (Ed.), *Educational Measurement (4th ed., pp. 17-64*). Westport, CT: Praeger.

Kane, M. T. (2013). The argument-based approach to validation. *Social Psychology Review, 42*(4), 448-457.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). The strands of mathematical proficiency. *Adding it up: Helping children learn mathematics*, 115-155.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*, 234-249. doi:10.3102/0013189X17718797

Lawson, J. E., & Cruz, R. A. (2018). Evaluating special educator's classroom performance: Does rater "type" matter? *Assessment for Effective Intervention, 43*(4), 227-240.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*, 25-47. doi: 10.1007/s10857-010-9140-1

Linacre, J. M. (2017). *Facets 3.80* [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.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015).
Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice, 34(2),* 34-46.

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.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-67.

Murphy, K., & Cleveland, J. (1991). Performance appraisal: An organizational perspective. Boston, MA: Allyn & Bacon.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, *4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189-227.

National Center for Education Statistics. (2015). Kena, G., Musu-Gillette, L., Robinson, J., Wang, X., Rathbun, A., Zhang, J., Wilkinson-Flicker, S., Barmer, A., & Dunlop Velez, E. (2015). *The condition of education 2015* (NCES 2015-144). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS).* Baltimore: Brookes.

Qi, Y., Bell, C. A., Jones, N. D., Lewis, J. M., Witherspoon, M. W., & Redash, A. (2018). Administrators' Uses of Teacher Observation Protocol in Different Rating Contexts. *ETS Research Report Series*, *2018*(1), 1-19.

Razynski, K., Engelhard, G., Cohen, A., & Lu, Z. (2015). Comparing the effectiveness of Self-paced and collaborative frame-of-reference training on rater accuracy in a large scale writing assessment. *Journal of Educational Measurement, 52*, 301-318.

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. doi: 10.1111/j.1949-8594.2002.tb17883.x

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.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*(3), 497–506.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice, 31*(1), 21-30.

Taylor, E. S., & Tyler, J. H. (2012). The effect pf evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651.

Wang, J., Engelhard, G., & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments with an unfolding model. *Educational and Psychological Measurement*, *76*, 1005–1025.

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47.

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.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26*, 283-306.

Wind, S. A., & Engelhard, G. (2016). Exploring rating quality in rater-mediated assessments using Mokken scaling. *Educational and Psychological Measurement, 76*, 685-706.

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37.

Table 1

Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater	Severity (Logits)	Model *SE*	Infit *MNSQ*	Outfit *MNSQ*	Exact Obs Agreement %
6	.47	.09	1.04	1.05	51.2
2	.43	.09	.76	.77	54.1
10	.18	.09	.97	.96	52.4
1	.07	.09	1.48	1.48	48.1
3	-.01	.09	.61	.62	55.2
5	-.02	.09	1.15	1.15	52.0
4	-.04	.09	.98	.98	53.4
8	-.04	.09	1.28	1.26	47.5
9	-.41	.09	1.00	1.00	46.7
7	-.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-square = 115.5; df = 9; significance < .001.

Table 2

*Rater Accuracy by Lesson by Item**

	Rater 1		Rater 2		Rater 3		Rater 4		Rater 5		Rater 6		Rater 7		Rater 8		Rater 9		Rater 10		Item Total	
Lesson	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Item																						
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	10	0
2	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	1	10	7
3	1	0	0	1	1	1	0	0	1	1	1	0	0	0	0	1	1	0	0	1	5	5
4	1	0	1	1	1	0	1	0	1	0	1	1	1	0	1	1	1	0	1	1	10	4
5	0	1	1	0	0	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	6	8
6	0	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	1	0	0	3	4
7	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	4	1
8	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	7	0
9	0	0	1	1	1	1	0	1	0	1	1	1	0	0	0	0	1	1	0	1	4	7
10	0	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	4	9
11	0	0	1	1	0	0	1	1	0	1	0	0	0	1	1	0	1	0	0	1	4	5
12	0	0	0	1	1	1	0	1	0	1	1	1	0	0	0	0	1	1	1	1	4	7
13	0	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	0	1	5	6
14	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	1	1	0	0	0	2	6
15	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	7	7
16	0	1	0	1	0	0	1	1	0	1	1	1	0	1	0	0	1	1	0	1	3	8
17	0	0	0	1	1	0	0	1	0	1	0	1	1	1	0	0	1	1	1	0	4	6
% Exact	29	47	53	65	59	47	53	65	41	59	65	59	41	41	53	35	100	47	47	65	54	53
Overall %Exact	38		59		53		59		50		62		41		44		74		56			

Note. Rater scores are computed so a 1 means that their score agreed with the expert rater and a 0 means they did not agree.

Table 3.

Count of Rater's Rationales and Scores Aligned with those Provided by Expert Raters Across Two Lessons

	Rater Number											
Category and Code	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total Count**	**%**
1. Score and rationale same as expert rater	15	19	14	18	15	18	12	14	19	14	158	46
2. Score same but rationale different than expert rater												
2.1 Rater added extra criteria.	1		1						1		3	1
2.2 Rater confused the target item with another item.		2	1		1	1	1	1		2	9	3
2.3 Rater provided rationale which supports a higher score or lower score.	1		1			1				2	5	1
2.4 Rater provided evidence that is inconsistent with the rationale.		1			1		1			1	4	1
2.5 Rater did not address key terms in the rubric that were emphasized by the expert raters.	1		2	1	3		1				8	2
3. Score different but rationales were similar to expert rater	1	2	1		2	3	2	1	1	1	14	4
4. Score and rationale were both different from expert rater												
4.1 Rater thinks the teacher conducted or partially conducted teaching activities, while the master raters thought the teacher partially did or did not conduct; Or reverse.	9	4	6	5	3	3	5	4	4	4	47	14
4.2 Rater thought differently about some subjective terms like adequate.	4	3	8	9	4	6	8	9	9	6	66	19
4.3 Rater confused the target item with another item.		1			1						2	1
4.4 Rater adds an extra criteria		1					1			1	3	1

4.5 Rater did not address key terms in the rubric that were emphasized by the expert raters.	2	1		1	3	1	1	4	1	1	15	5
4.6 Rater provided evidence that is inconsistent with the rationale.							1				1	1
5. Rater did not provide a rationale.					1	1	1	1		1	5	1

Table 4

Count of Rater's Rationales and Scores Aligned with those Provided by Expert Raters Across Two Lessons

	Lesson Number	
Category and Code	**1**	**2**
1. Score and rationale same as expert rater	75	83
2. Score same but rationale different than expert rater		
2.1 Rater added extra criteria.	2	1
2.2 Rater confused the target item with another item.	4	5
2.3 Rater provided rationale which supports a higher score or lower score.	4	1
2.4 Rater provided evidence that is inconsistent with the rationale.	3	1
2.5 Rater did not address key terms in the rubric that were emphasized by the expert raters.	5	3
3. Score different but rationales were similar to expert rater	1	13
4. Score and rationale were both different from expert rater		
4.1 Rater thinks the teacher conducted or partially conducted teaching activities, while the master raters thought the teacher partially did or did not conduct; Or reverse.	13	34
4.2 Rater thought differently about some subjective terms like adequate.	51	15
4.3 Rater confused the target item with another item.	1	1
4.4 Rater adds an extra criteria	1	2
4.5 Rater did not address key terms in the rubric that were emphasized by the expert raters.	4	11
4.6 Rater provided evidence that is inconsistent with the rationale.	0	1
5. Rater did not provide a rationale.	5	0