Understanding Rater Behavior in Observations of Special Education Teachers

Evelyn S. Johnson, Yuzhu Zheng, Laura A. Moylan, and Angela Crawford

Boise State University

Author Note

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

Citation: Johnson, E. S., Zheng, Y., Moylan, L. A., & Crawford, A. R. (2020). Understanding rater behavior in observations of special education teachers. *Studies in Educational Evaluation. 64* 100827.

Abstract

In this study, we investigated factors that influence raters' application of the scoring criteria of an Explicit Instruction (EI) observation protocol using many-faceted Rasch measurement (MFRM) and think aloud analysis. Specifically, we investigated the extent to which raters are able to consistently represent the scoring criteria in the EI rubric, how raters discriminate among levels of performance on each instructional element, and the consistency with which the raters applied evidence to support their scoring decisions. Video observations of instruction from 30 special education teachers across three states were collected. External raters (n = 15) observed and scored videos, providing rationales for item level scores. MFRM analyses showed that raters differed in their severity, but each individual rater was able to apply scoring criteria in a consistent manner. Think aloud analyses showed that raters varied in how they interpreted items and in the evidence they used to support their scores. Implications for research are discussed.

Keywords: rater behavior, observation of special education teachers, explicit instruction, Many-facet Rasch measurement, Think aloud

Understanding Rater Behavior in Observations of Special Education Teachers

With the reform in teacher evaluation systems, observation of teacher practice has become a promising way to provide direct measures of teaching effectiveness, and provide teachers with feedback on how to improve instruction (Johnson et al., 2018). Observations of teacher practice rely on the use of protocols that capture the salient elements of instruction. *Effective* observation protocols create information that can be used to evaluate, identify and improve specific teaching practices (Gitomer et al., 2014). Studies on classroom observations suggest that when teachers are objectively evaluated and supported to improve instruction, there is a positive impact on student growth (Biancarosa et al., 2010; Taylor & Tyler, 2012). In recent years, a number of observation protocols have been developed such as, the Danielson Framework for Teaching (FFT; Danielson, 2013), the Classroom Assessment Scoring System (CLASS, Pianta et al., 2007); Mathematical Quality of Instruction (MQI; Hill et al., 2011), and the Protocol for Language Arts Teaching Observation (PLATO; Grossman et al., 2009). Although these observation protocols differ in their specifics, what they share is an underlying theory of change that through the use of detailed scoring criteria and procedures implemented by raters, teachers' instructional practice will improve (Bell et al., 2015).

A consistent finding in studies of teacher observation is that the *instructional* dimensions of observation protocols are the most challenging for raters to score reliably (Bell et al. 2015, Bill and Melinda Gates Foundation, 2011; Gitomer et al, 2014). Across multiple large-scale studies of teacher observation, raters account for between 25 to as much as 70% of the variance in scores assigned to the *same* lesson (Casabianca et al., 2015). Low levels of exact agreement among raters on instructional dimensions (ranging from 47 – 52%) have been reported across a number of studies using a variety of observation instruments (Bell et al., 2015; Cash et al., 2012;

Jones, 2019; Kane & Staiger, 2012). Methods to improve rater reliability and consistency have been investigated, but low levels of agreement persist even when raters are required to meet certification requirements and engage in calibration and validation efforts (Casabianca et al., 2015; Jones, 2019). These findings present a challenge to the theory of change that teacher observations can lead to instructional improvement, and warrant a closer examination of how rater behavior, particularly raters' understanding and application of scoring criteria, impacts results.

Although some would argue that consistency in observing teachers to provide feedback is not as important as it is for evaluation, we disagree. The potential for observation protocols to lead to improved instructional practice depends on the extent to which different raters make the same judgments given the same evidence (Gitomer et al., 2014). The rater must be able to consistently use a protocol to distinguish among different instructional elements and the levels of performance of implementation to ensure that the scores assigned and the feedback provided to the teacher are not unduly influenced by the rater assigning them (Hill & Grossman, 2013). If teachers receive inconsistent messages about their teaching, they will be less likely to have confidence in the observation process, and may be less likely to change their practice (Johnson et al., 2016). Drawing on the research on rater behavior on performance assessments and an analysis of two large-scale observation studies, Bell et al (2015) identified three aspects of scoring that contribute to the accuracy and consistency of observations: 1) the raters' maintenance of a consistent representation of the scoring criteria, 2) the raters' judgments used to discriminate the levels of performance on each instructional element, and 3) the raters' collection and application of evidence to support scoring decisions. Understanding rater behavior in the context of these three aspects is a critical consideration to improve the accuracy and consistency with which teacher practice is evaluated. In the present study, these three aspects of rater behavior are examined within the context of a special education teacher observation protocol. First, a description of the observation protocol is provided, followed by the specific research questions to be addressed.

Recognizing Effective Special Education Teachers (RESET)

The Recognizing Effective Special Education Teachers (RESET) observation system is a federally funded project to create a special education teacher observation system aligned with evidence-based instructional practices (EBPs) for students with high incidence disabilities (SWDs). SWDs are defined as those with mild emotional/behavioral disorders, learning disabilities, high functioning autism, other health impairment (ADHD) or language impairment. The goal of RESET is to leverage the extensive research on EBPs for this population of students and to develop observation protocols aligned with these practices in order to: (a) determine the extent to which special education teachers are implementing EBPs with fidelity, (b) provide feedback to special education teachers to improve their practice and, (c) improve outcomes for SWD.

The theory of change underlying RESET states that improving teacher practice depends upon having a clear target for quality instruction that is articulated through the alignment of an observation instrument with the salient characteristics of the instructional practices that have been demonstrated to be effective for SWD (see Figure 1). The RESET observation system was designed to provide this clear target through psychometrically sound observation protocols aligned with EBPs that provide reliable evaluations of teachers’ instruction and allow for the provision of feedback that is specific and actionable. Through this process, it is anticipated that the teachers' ability to implement EBPs improves, and this instructional improvement will result in their students' accelerated academic growth.

RESET was developed using the principles of Evidence-Centered Design (ECD; Mislevy et al., 2003) and consists of 21 observation protocols that detail evidence-based instructional practices organized in three categories: a) instructional methods, b) content organization and delivery, and c) individualization (see Table 1). A complete description of how the RESET protocols were developed is detailed in Johnson et al (2018). In this study, the focus is on the Explicit Instruction protocol. The Explicit Instruction protocol was selected because explicit instruction was recently identified as one of 22 high leverage practices for SWD (McLeskey et al., 2017), and is supported by nearly 50 years of research as an effective approach for SWD across grade levels and content areas (Hughes et al., 2017; Stockard et al., 2018).

The Explicit Instruction protocol includes 25 items that comprise the salient elements of this instructional practice (see Table 4). The entire Explicit Instruction protocol and training manual is available at https://education.boisestate.edu/reset. The protocol provides raters with descriptions of the specific instructional elements of Explicit Instruction and includes detailed performance level descriptors for each item to improve the consistency with which raters interpret and evaluate instructional practice. In our work examining how the various facets of the rubric function (e.g. item, teacher, rater), the protocol has been demonstrated to be psychometrically sound. For example, generalizability theory studies conducted on the Explicit Instruction rubric indicate an overall *g* coefficient of .74 with raters accounting for less than 5% of the variance (Crawford et al., 2019). In two studies using many-faceted Rasch measurement (MFRM) analyses, fit statistics, reliability and separation indices and bias analyses were all within acceptable ranges (Johnson et al., 2016; Johnson et al., 2018). While these indicators of reliability suggest that the observation protocol produces generally consistent scores, across these studies, the *exact* level of rater agreement ranged from 50 - 51%. This is disconcerting since the goal is to create an observation protocol that results in consistent evaluations and feedback to teachers to improve their ability to implement explicit instruction. Placing the findings from the RESET studies into the broader teacher observation research, it is important to note that several large-scale observation studies report similar, low levels of exact agreement (range from 47 – 57%) across a number of widely-used observation protocols (see for example, Bell et al., 2015; Cash et al., 2012; Jones, 2019; Kane & Staiger, 2012). These studies employed rigorous rater training and certification methods, including periodic calibration and validation efforts, yet were unable to maintain consistently high levels of exact inter-rater agreement.

If a rater's observations of teacher instruction are used to evaluate, to provide feedback and to plan professional development opportunities, then the data informing these decisions must be robust (Mantzicopoulos et al., 2018). In order to inform improvements to the rater training process, it is important to better understand the factors that influence raters' application of the scoring procedures and criteria of an observation protocol. Therefore, the purpose of this study was to further investigate: 1) the extent to which raters are able to consistently represent the scoring criteria in the Explicit Instruction protocol and associated training manual, 2) how raters discriminate among levels of performance on each instructional element, and 3) the consistency with which the raters collected and applied evidence to support their scoring decisions.

Methods

The study reported in this manuscript is part of a larger study to examine the functioning of the various facets of the Explicit Instruction rubric using MFRM analysis. Drawing on data collected to examine the psychometric properties of the Explicit Instruction rubric (Johnson et al., 2018), here we report specifically on the rater facet, and the additional data collected to further understand how raters applied the scoring procedures and criteria to evaluate special education teachers' instructional practice.

Participants

Special Education Teachers. Thirty special education teachers from three states participated in this study. They were recruited by sending study information and recruitment letters to the special education district directors, who identified strong special education teachers and shared recruiting materials with them. Data collection took place during the 2015-16 and 2016-17 school years. All participants provided video recorded lessons that reflected their use of explicit instruction in either reading or math intervention. All participants were female, teaching from 2nd to 8th grade levels in a resource room context. Two of 30 teachers were Asian, and the remaining 28 teachers were white. Their number of years' experience ranged from 1 to 29 years (M = 9.2, SD = 4.7). Teachers had a range of education credentials. All participants held a Bachelor's degree in special education, and 16 teachers also had a Master's degree in either Special Education or Literacy Education. Teachers worked across a variety of school settings, with a summary of the school demographics presented in Table 2.

Raters. One male and fourteen female raters were recruited from seven states to score videos provided by the 30 teachers. 12 raters were white, 2 Asian, and 1 Pacific Islander. Criteria for raters included having five or more years of experience working with SWD, and strong knowledge of explicit instruction. All raters were special education professionals with between 5-20 years of working experience. Two raters had a Bachelor's degree in Special Education, 11 had a Master's Degree, and two had Doctoral Degrees. At the time of the study, eight raters worked as classroom teachers, three were mentor teachers or instructional coaches, two were special education graduate students, one was a specialist at a state Department of Education, and one was a school psychologist and RTI coordinator within her district.

Procedures

Rater Training. Over a four-day training period, raters were first provided with an overview of the RESET project goals and a description of how the EI rubric was developed. Research project staff then explained each item of the EI rubric 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 watched and scored a video that had been scored by project staff. The scores were reviewed and discussed to include the rationale for the score that each item received. Raters then watched and scored three videos independently, and scores were reconciled with the master coded rubric for each video. Any disagreements in scores were reviewed and discussed. To determine rater agreement, Kendall's coefficient of concordance, W, was used to allow for ordinal data with multiple raters. For the first training video, W = .191 p < .001. For the fourth video, W = .303, p < .001. During the training, minimum rater performance standards were not established because some studies have shown that raters still account for large portions of variance, and issues, such as drift, persist even with establishing minimum performance standards (Cash et al., 2012; Jones, 2019; Kane & Staiger, 2012). Instead, 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.

After training, raters were then assigned a randomly ordered list of videos 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. Raters were reminded to consult the training manual as they completed their observations and were given a timeframe of four weeks to complete their ratings. Completed evaluations were submitted using an electronic version of the rubric developed in the Qualtrics ® survey system.

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). We randomly selected two teachers to have their first and last video scored by every rater. Remaining videos were randomly assigned so that each video was scored by four raters. This created a design in which 13 raters scored 28 videos each, one rater scored 32 videos, and one rater scored 8 videos.

Rater Think-Aloud. For one of the video-recorded lessons that all raters scored, we asked raters to audio record a think aloud as they used the rubric and the training manual to score the lesson. Raters were asked to score a video as they normally would for the study, but to think out loud and audio record their rationale and thought process as they did so. Although we did not provide specific questions or prompts for every item, we did provide examples of what raters might include in their think aloud (see Figure 2).

Data analysis

The scores assigned to the recorded lessons were analyzed through MFRM analysis, which is a model including all sources of variability (facets) that are thought to influence the scores in the analysis (Eckes, 2011). All facets are calibrated simultaneously and receive a common score on a linear scale (the logit scale) that represents the latent construct. Each facet can be examined independently to assess levels of reliability, precision, and consistency to help determine whether or not the rating system is functioning as intended (Vogel & Engelhard, 2011). One advantage of using MFRM to analyze rater behavior is that it can account for differences in rater severity by adjusting the observed score and computing an average fair score for teachers. This is different than other approaches to examining rater behavior that expect raters to function as scoring machines, achieving perfect agreement against a master set of scores (Eckes, 2012; Myford & Wolfe, 2003). Research on rater behavior suggests that achieving perfect agreement across human raters who judge complex performances is an elusive goal and that 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, 2012). This study was designed as a four-facet model including item, teacher, rater, and lesson. The model used for the MFRM analysis in this study is given by:

lnPnijokPnijo(k-1)= Bn-Di-Cj-To-Fk

where Pnijok 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*. Pnijo(k-1) is the probability of teacher *n*, when rated on item *i* by judge *j* in occasion *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 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 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 (Johnson et al., 2018).

Think Aloud Analysis. The think aloud data were first analyzed to investigate the raters' exact agreement with a master scored rubric of the lesson. Then, rationales provided for scores that differed from the master score were examined and summarized to better identify ways in which raters were interpreting the performance level descriptors for the items.

Next, the think aloud data were analyzed in two stages to address the third research question. At stage one, the think-aloud protocols were transcribed and then sorted by item (e.g. all of the think alouds for item one were grouped by rater, then the think alouds for item two and so on). The transcriptions were then tagged according to each performance level descriptor of the rating scale in the rubric. The focus of the analysis centered on the rationales the raters provided to support their scores, and whether the rationales were consistent with the criteria as provided in the rubric and training manual. Key words, short phrases or sentences from the raters' explanations for their scores were identified as the unit of analysis. Selective coding of these units was conducted based on the following guiding questions:

1. What are the rationales provided for each score?
2. Are the stated rationales consistent with the criteria as defined in the rubric and training manual?
3. If the stated rationales are not consistent with the scoring criteria defined in the rubric, in what way are they inconsistent?

As a result of this analysis, seven categories were developed to summarize the consistency of the raters' rationales with the scoring criteria (see Table 5). At stage two of the analysis, we used the categories developed in stage one to analyze a random selection of 20% of the scores and responses (1,925) from the entire data set to determine the extent to which raters were consistent with the guidance provided in the protocols, training manuals and training sessions.

Results

To what extent are raters able to consistently represent the scoring criteria in the EI rubric and associated training manual?

The statistics for the rater facet from the MFRM analysis are presented in Table 3. The rater severity ranges from -.31 logits (*SE*=.03) for Rater 14 who is the most lenient to .52 logits (*SE*=.03) for Rater 9 who is the most severe. The rater fixed chi-square value tests the assumption that all the 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 fixed chi-square value of 659.1 with 14 degrees of freedom is statistically significant (p < .01), which signifies that raters differed in severity when evaluating the teachers. 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 .97 further confirmed the variability in rater severity. As can be seen in Table 3, the fit statistics for the rater facet fell between .62 to 1.34, which is within the acceptable ranges and indicates the consistency of rater severity 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.

How do raters discriminate among levels of performance on each instructional element?

Table 4 presents a by item exact agreement and a summary of the rationales provided for scores that differed from the master scored lesson. Only two items (number 3 and 7) had assigned scores that spanned all levels of performance descriptors. Items with the highest percentage of agreement tended to be lower inference items, or items focused on the materials or content as opposed to the teacher actions. For example, the item with the highest level of agreement was Item 1*, The goals of the lesson are clearly communicated to students*; the item with the lowest level of agreement was Item 3, *The teacher clearly explains the relevance of the stated goal to the students*. The difficulty with consistent scoring of item 3 centered around what is meant by 'relevance'. Some raters interpreted relevance as a real-world application, whereas other raters were consistent with the way the item is described in the training manual for the explicit instruction rubric, *This item assesses whether the teacher explains to students the value of the stated goal to their overall course of study or to their lives* (Johnson et al., 2016, p.10). In other cases, it appears that the assignment of a score was a possible data entry error. For example, on Item 9, the raters who assigned a score of 3 provided rationales consistent with a score of 2.

The most challenging items to score consistently (Items, 10, 16, 17, 22, 23, 24) included the phrase, *throughout the lesson*. It is unclear the extent to which raters made their judgments on the entire lesson, as the documented rationales only mentioned *throughout the lesson* in items 17 and 24. For example, 46% raters assigned a score of 3 instead of the master score of 2 to the teacher on Item 24 (*Feedback is specific and informative throughout the lesson*) because the differences centered on whether feedback was specific and whether it was provided throughout the lesson. Item 18 also includes the phrase *throughout the lesson* but the question is related to an alignment of focus on the stated goal rather than on a more specific instructional practice.

What is the consistency with which the raters collected and applied evidence to support their scoring decisions?

Seven categories which summarize the raters' rationales for their scores comparing with the criteria of the protocol emerged from the analysis of the think-aloud data and are presented in Table 5. These seven categories were used to further examine a random sample of the total responses (including the score assigned to the item and the brief, written rationale and evidence provided by the rater) for the entire set of scored videos.

The analysis shows that 59% of raters' rationales for the given scores were fully consistent with the scoring criteria. For example, one rater gave the teacher a score of 3 on Item 7 because "Before reading the story, the teacher works to bring them back up to speed on what they've read so far to help them understand better what they're about to read." This rationale is fully consistent with the criteria for a score of 3 on this item, which is "The teacher effectively reviews prior skills and/or engages background knowledge before beginning instruction." Approximately 20% of raters' explanations were partially consistent with the rubric, but some important components of the criteria were missing in the raters' evidence. For example, one rater assigned a score of 2 to Item 8 because "The teacher provides demonstrations of identifying the message found in the story of the boy who cried wolf after she asked students what they thought the message was." One component of the criteria for a 2 in the rubric was missing, which is "The teacher does not provide **clear** demonstrations of proficient performance." The rater only mentioned that the teacher provided demonstrations, but she did not provide any information about whether the demonstrations were clear, which is a critical component for a score 2 in the rubric. Approximately 2% of the raters' rationales were partially consistent with additional criteria added by the raters themselves. For example, one rater gave a score 2 on Item 10 because "language is clear and precise but [the teacher] needed better questioning techniques." The explanation was partially consistent with the criterion for score 2 in the rubric, which is "The teacher uses language that is not always clear, precise, and accurate." In the explanation, the rater commented on whether the teacher's language was clear or precise, however, she added an extra criterion by focusing on the teacher's questioning technique, which is not part of the criteria for this item.

Seven percent of raters' explanations were not consistent with the criteria in the rubric. The raters provided explanations, which were not relevant to the target item or other items in the rubric. For example, one rater gave a score of 3 to the teacher on Item 6 because "all the examples or material provided to the students are aligned with using the four ways to solve the multiplication facts." However, the criteria for a score 3 in the rubric is "Examples or materials selected are aligned to the *instructional level* of most or all of the students." The rater did not address the alignment to instructional level, she focused on alignment of the examples with the taught strategy. Approximately 3% of explanations were not for the target item, but for another item in the rubric. For example, one rater gave a score 2 to the teacher on Item 9 because "Each student gets a turn to come to the board and answer questions as others follow along." The criteria for this score for this item is "The teacher does not provide an adequate number of demonstrations given the nature and complexity of the skill or task." It is unclear whether the rater interpreted the students' turn at the board as the way in which the teacher was providing demonstrations or whether she was confusing this with providing opportunities for students to engage or respond, which is Item 19.

Some raters used the same rationale for several items. Approximately 5% of the explanations provided by the raters were used to support multiple items. Approximately 4% of raters' evidence was consistent with the criterion for a lower or higher score described in the protocol, but it was unclear whether this was a data entry error or whether the rater misinterpreted the criteria. For example, one rater gave a score 2 on Item 25 because "She did adjust her instruction based on student response. When something new was introduced, she made sure there was understanding before adding a new concept or moving on." The criteria for a score 3 in the rubric is "The teacher makes adjustment to instruction as needed based on the student response." The rater did not provide any additional information to explain why she assigned a 2 instead of a 3.

Discussion

The validity of an observation protocol relies in large part on the consistency with which raters use the scoring criteria and procedures to evaluate instructional practice. Consistent with the larger body of research on teacher observation, the results of this study suggest that there is still much work to do to support raters' ability to consistently evaluate instruction. While one approach to achieving higher reliability would be to eliminate items that are more difficult for raters to agree on, "the reliability of scores is not simply a psychometric hurdle" (Gitomer et al., 2014, p. 24). Eliminating items that reflect critical elements of an instructional practice will not support the end goal of improving teacher practice. Instead, improving reliability requires raters to develop shared understandings of the instructional practices to be observed, and to consistently translate what the observe to a protocol and its related scoring criteria and procedures (Bell et al., 2014; Gitomer et al., 2014; Hill & Grossman, 2013). To better understand the type of training that raters might need to achieve the desired level of consistency, it is critical to understand rater behavior.

The results of the MFRM analysis suggest that the raters differed in severity when evaluating teachers but also show that 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, 2012). 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. But there is little assurance that a state or district has the capacity and resources to employ such methods in the context of teacher observation. Additionally, correcting scores for rater severity may be theoretically possible when observation protocols are being used for evaluation, but they are not feasible when they are used as formative assessments to support teachers in improving their practice. While a score may be of less consequence in a formative assessment situation, the larger question is whether a teacher would receive consistent feedback about her current level of performance.

The results of this study indicate that even when the rater is internally consistent in scoring, it is important to examine the rater's thinking process and decision-making to ensure consistency with an observation protocol's scoring procedures and criteria. The Explicit Instruction 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 item's performance level descriptors 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; Goe et al., 2008; 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 those items across a variety of teachers and lessons, and consistently mapping these performances to a set of scoring criteria remains a challenge. The items that were most problematic were those that require a careful judgment (e.g. *Scaffolding is provided when needed to facilitate learning*) and those that demand a continuous focus on instruction across an entire lesson. The cognitive demand of attending to a practice throughout a lesson may be too high for raters to score reliably. However, if the desired level of implementation includes the need to employ a practice for a sustained period of time, then it is necessary to determine a way to reliably measure and provide feedback on these practices (Goe et al., 2008).

Results from this study also have practical implication for selection of raters in the teacher evaluation system. The potential for using observation tools to lead to improved instructional practice relies in large part on raters who are able to accurately understand the elements of the observation tools, and accurately and consistently apply scoring criteria and procedure to practice (Bell et al., 2012; Gitomer et al., 2014; Hill & Grossman, 2013; Johnson et al., 2020). Raters are found to be more reliable if they have strong background or expertise in the subject of the instrument (Blazar et al., 2017; Lawson et al., 2018). There is some evidence to suggest that raters with expertise in special education are more accurate in evaluating special education teachers than those without (Lawson et al., 2018). Therefore, it is also critical to select the right raters in the observation systems.

Limitations and Future Directions

In addition to the small sample of raters and teachers, which limits the generalizability of our findings, there are two important limitations of this study. First, the coding system used was grounded in one observation of a teacher's instructional practice. The general nature of the coding categories and their emphasis on the consistency of rater evidence with the scoring criteria however, limit this concern. Second, although the RESET Explicit Instruction protocol is designed for use as both an evaluative and formative instrument, the context in which raters conducted their scoring and provided evidence was as a teacher evaluation, not as teacher feedback. It is possible that the thinking-process, focus, and accuracy may differ if raters scored the lessons under a different context.

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 consistency 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 Explicit Instruction implementation. 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 rater specific 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. In addition to training raters, it could prove more effective to also invest in improving teachers' understanding of EBPs, and to support teachers' ability to use the observation protocols aligned with EBPs to work collaboratively with raters and coaches to set goals, self-evaluate, plan, and make progress towards those goals.

References

Bell, Courtney A., Yi Qi, Andrew J. Croft, Dawn Leusner, Daniel F. Mccaffrey, Drew H. Gitomer, and Robert C. Pianta. (2014). Improving observational score quality: Challenges in observer thinking T.J. Kane, K.A. Kerr, R.C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*, Jossey-Bass, San Francisco, CA, pp. 50-97.

Biancarosa, G., Bryk, A., & Dexter, E. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal, 111*(1), 7-34. http://harringtonmath.com/wp-content/uploads/2013/11/Content-knowledge-for-teachers.pdf

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains*. Seattle, WA: Author.

Blazar, D., Braslow, D., Charalambos, Y. C., & Hill, H. C. (2017). Attending to general and mathematics specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment, 22*(2), 71-94. https://doi.org/10.1080/10627197.2017.1309274

Bond, T. G., & Fox, C. M. (2007). Fundamental measurement in the human sciences. *Chicago, IL: 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. https://doi.org/10.1177 /0013164414539163

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 of raters associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529-542. https://doi.org/10.1016/j.ecresq.2011.12.006

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. Z. (2019). Variance and reliability in special education observation rubrics. *Assessment for Effective Intervention, 45*(1), 27-37. https://doi.org/10.1177/1534508418781010

Danielson, C. (2013). *The framework for teaching evaluation instrument*. Princeton, NJ: The Danielson Group.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270-292. https://doi.org/10.1080/15434303.2011.649381

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

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191. https://doi.org/10.1207/s15324818ame0503_1

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. https://www.academia.edu/34327254/The_instructional_challenge_in_improving_teaching_quality_Lessons_from_a_classroom_observation_protocol

Goe, L., Bell, C. A., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality. http://www.gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf

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. https://tedd.org/wp-content/uploads/2014/03/Grossman-et-al-Teaching-Practice-A-Cross-Professional-Perspective-copy.pdf

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. https://doi.org/10.3102/0013189X12437203.

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. https://doi.org/10.17763/haer.83.2.d11511403715u376

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831. https://doi.org/10.3102/0002831210387916

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical and contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140-148. https://doi.org/10.1111/ldrp.12142

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2016). *Explicit Instruction Rubric Training Manual*. https://education.boisestate.edu/reset

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2018). Using evidence-centered design to create a special educator observation system. *Educational Measurement: Issues* *and Practice*. https://doi.org/10.1111/emip.12182

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2020, accepted for publication). The relationship of special education teacher performance on observation instruments with student outcomes. *Journal of Learning Disabilities*.

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, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Lawson, J. E. & Cruz, R. A (2018). Evaluating special educators' classroom performance: Does rater "type" matter? *Assessment for Effective Intervention, 43*(4),1-14. https://doi.org/10.1177/1534508417736260

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. https://doi.org/10.1080/10627197.2017.1408407

McLeskey, J., Barringer, M. D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M.,& Ziegler, D. (2017). High-leverage practices in special education. *Arlington, VA: Council for Exceptional Children & CEEDAR Center*. *Google Scholar*.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

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.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). Classroom assessment scoring system (CLASS). Baltimore: Brookes.

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479-507. https://doi.org/10.3102/0034654317751919

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651. doi: 10.1257/aer.102.7.3628

Vogel, S. P., & Engelhard, G., Jr. (2011). Using Rasch measurement theory to examine two instructional approaches for teaching and learning of French grammar. *Journal of Educational Research, 104*(4), 267–282. https://doi.org/10.1080/00220671003733815

Table 1

Organization and Structure of RESET

Subscale	Content Area	Rubrics
Instructional Methods		Explicit Instruction Cognitive Strategy Instruction Peer Mediated Learning
Content Organization and Delivery	Reading	Letter Sound Correspondence Multi-Syllabic Words and Advanced Decoding Vocabulary Reading for Meaning Comprehension Strategy Instruction Comprehensive Reading Lesson
	Math	Problem Solving Conceptual Understanding Procedural Understanding of Automaticity
	Writing	Spelling Sentence Construction Self Regulated Strategy Development Conventions
Individualization		Self-Regulation Data-Based Decision Making Universal Design/Assistive Technology

Table 2

School Demographics

School	Grade	Enrollment (%Female)	White	Hispanic	Asian	Multi-race	Black	American Indian	%FRL	%SWD
1	K – 6	523 (48)	85	6	4	3	1	1	27	7
2	K – 5	470 (47)	75	14	4	4	2	1	54	9
3	6-8	1230 (50)	88	6	2	2	1	1	19	8
4	6-8	990 (49)	81	8	4	3	3	1	35	9
5	K– 5	729 (48)	76	9	7	4	3	1	63	11
6	K-6	358 (53)	79	11	4	4	1	1	63	7
7	K-5	664 (46)	52	45	1	1	1	1	68	10
8	K-7	810 (44)	88	6	4	2	1	1	16	4
9	K-6	368 (52)	79	7	5	5	2	2	98	10
10	K – 5	668 (49)	87	6	3	3	1	1	21	7
11	K-5	429 (44)	72	19	6	1	1	1	67	8
12	6-8	699 (44)	31	67	1	1	1	1	90	10
13	K-5	350 (51)	86	12	1	1	1	1	46	8
14	9-12	1369 (50)	87	9	1	1	1	1	34	8
15	K-6	511 (49)	59	27	2	1	9	1	100	10
16	K-5	498 (49)	28	70	1	2	1	1	95	8
17	K-6	518 (50)	89	4	3	1	2	1	31	9
18	K-8	359 (52)	85	8	2	3	1	1	16	5
19	6-8	906 (50)	63	32	1	1	1	1	64	8
20	6-8	711 (46)	41	55	1	2	1	1	87	7
21	K-6	163 (44)	91	4	1	3	1	1	53	9
22	K-5	643 (51)	65	31	1	2	1	1	64	8
23	K-3	292 (48)	69	21	4	2	3	1	40	9
24	4-8	345 (48)	70	19	1	2	7	1	49	8
25	K-12	60 (35)	90	9	0	0	1	0	33	45
26	K-5	508 (46)	65	31	2	1	1	1	40	8
27	K-8	252 (44)	88	5	5	2	1	1	38	5

Note. FRL = Free and Reduced Lunch, SWD = students with disabilities

Table 3

Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater Number	Severity (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
9	.52	.03	.62	.62
3	.27	.03	1.15	1.17
4	.20	.03	.80	.77
15	.19	.06	.75	.81
6	.17	.03	.96	1.01
5	.03	.03	1.24	1.19
8	.02	.03	.81	.84
10	-.02	.03	.99	.97
1	-.06	.03	1.01	1.09
12	-.13	.03	1.34	1.34
7	-.18	.03	1.06	1.00
11	-.21	.04	.96	.98
2	-.22	.03	1.02	1.04
13	-.25	.04	1.16	1.14
14	-.31	.03	1.07	1.06
Mean (count = 15)	.00	.04	.99	1.00
SD	.23	.01	.19	.19

Note. Root mean square error (model) = .04; adjusted *SD* = .22; separation = 6.13; reliability = .97; fixed chi-square = 659.1; df = 14; significance = .00.

Table 4

Analysis of rater scores and rationales across a common lesson

Item	3	2	1	Explanation for Different Scores* Assigned
1. The goals of the lesson are clearly communicated to students.	**92%**	8%*		The teacher did not have students repeat the goal
2. The stated goal(s) is/are specific.	**69%**	31%*		No details provided on how to achieve goal
3. The teacher clearly explains the relevance of the stated goal to the students.	**38%**	54%*	8%*	Relevance to real world use not provided; did not see
4. Instruction is completely aligned to the stated or implied goal.	**77%**	23%*		Teacher introduced new idea at end of lesson
5. All of the examples or materials selected are aligned to the stated or implied goal	**77%**	23%*		Students cannot solve without help
6. Examples or materials selected are aligned to the instructional level of most or all of the students.	**85%**	15%*		Not aligned to **all** students (instead of most)
7. The teacher effectively reviews prior skills and/or engages background knowledge before beginning instruction.	15%*	**77%**	8%*	Statement of goals was the review; rater did not observe
8.The teacher provides clear demonstrations of proficient performance.	31%*	**69%**		Indicated there was not enough demonstration but felt it was appropriate anyway

9. The teacher provides an adequate number of demonstrations given the nature and complexity of the skill or task.	23%*	**77%**		Demonstrations were not adequate (but assigned a 3)
10. The teacher uses language that is clear, precise, and accurate throughout the lesson.	46%*	**54%**		No details provided, simply said language was clear
11. Scaffolding is provided when it is needed to facilitate learning	62%*	**38%**		Differences in the focus on **quality** of scaffolding
12. Complex skills or strategies are broken down into logical instructional units to address cognitive overload, processing demands, or working memory	38%*	**62%**		Differences in the focus of effectiveness of how the skills are broken down
13. The teacher systematically withdraws support as the students move toward the independent use of the skills.	38%*	**62%**		Focus on whether the withdraw was systematic
14.Guided practice is focused on the application of skills or strategies related to the stated or implied goal	77%*	**23%**		Extent to which practice is on the **application** of skills
15. The teacher consistently prompts students to apply skills or strategies throughout guided practice	77%*	**23%**		Teacher prompts students to complete work, but prompts do not focus on **application** of taught strategies

16. The teacher maintains an appropriate pace throughout the lesson.	54%*	**46%**		Differences based on whether pacing is focused on finishing the lesson or on responding to students' needs
17. The teacher allows adequate time for students to think or respond throughout the lesson.	54%*	**46%**		Raters who assigned a 3 did not focus on 'throughout the lesson'
18. The teacher maintains focus on the stated or implied goal throughout the lesson.	**85%**	15%*		Raters who assigned 2 said the teacher was inconsistent with the focus on strategies
19. The teacher provides frequent opportunities for students to engage or respond during the lesson.	69%*	**31%**		Differences based on whether the opportunities were there for all students and in multiple ways of engagement
20. There are structured and predictable instructional routines throughout the lesson.	**62%**	38%*		Rationales provided were not clear – most raters used terms like "I feel like routines are predictable"
21. The teacher monitors students to ensure they remain engaged.	69%*	**31%**		Differences based on consistency of monitoring throughout the lesson
22. The teacher consistently checks for understanding throughout the lesson.	46%*	**54%**		Raters who gave a 3 equated asking questions with checking for understanding, even if no time allowed for student response
23.The teacher provides timely feedback throughout the lesson.	**62%**	38%*		No rationales provided to explain why a 2 was assigned

24. Feedback is specific and informative throughout the lesson.	46%*	**54%**		Differences centered around whether feedback was specific, and whether it was provided throughout the lesson
25. The teacher makes adjustments to instruction as needed based on the student responses.	14%*	**86%**		Raters who gave a 3 said the teacher responded to student questions

Note. Bolded responses are consistent with the master scores.

Table 5

Consistency Summaries

Category	**Count**	**Percentage**
1. Provided rationale is fully consistent with scoring criteria	1132	59
2. Provided rationale is partially consistent with scoring criteria but with missing components	378	20
3. Provided rationale is partially consistent but with additional criteria added by rater	47	2
4. Provided rationale is not consistent with scoring criteria and irrelevant evidence is cited	133	7
5. Provided rationale is related to another item	65	3
6. Provided rationale is the same across multiple items	100	5
7. Provided rationale is consistent with a different performance descriptor (possible data entry error)	70	4
Total	**1925**	**100**

Figure 1

Theory of Change for RESET

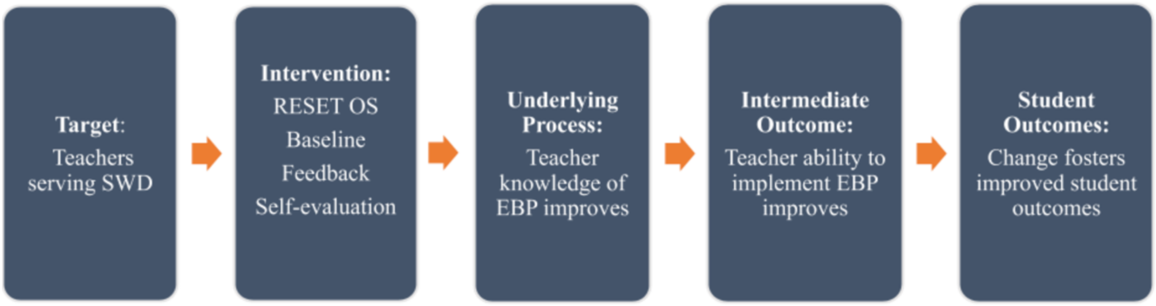
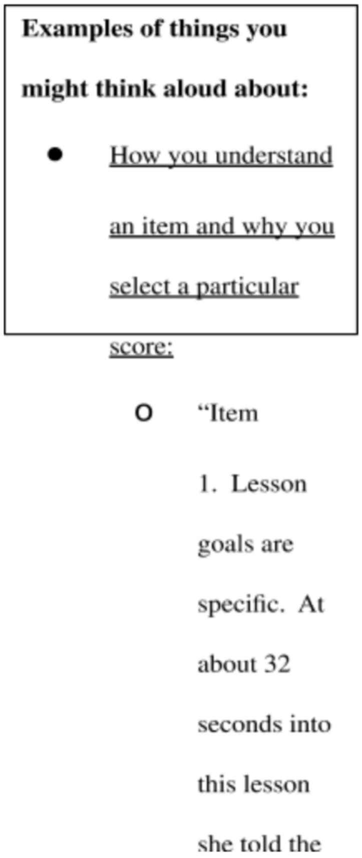


Figure 2

Guidance provided for rater think-aloud



Examples of things you might think aloud about:

- How you understand an item and why you select a particular score:
  - "Item 1. Lesson goals are specific. At about 32 seconds into this lesson she told the