Developing an Explicit Instruction Special Education Teacher Observation Rubric

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

Boise State University

July 2018

Author Note

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

Abstract

In this study, we developed an Explicit Instruction special education teacher observation rubric that details the elements of explicit instruction, and tested its psychometric properties using many-faceted Rasch measurement (MFRM). Video observations of classroom instruction from 30 special education teachers across three states were collected. External raters (n = 15) were trained to observe and evaluate instruction using the rubric, and assigned scores of 'implemented', 'partially implemented' or 'not implemented' for each of the items. Analyses showed that the item, teacher, lesson and rater facets achieved high psychometric quality for the instrument. Implications for research and practice are discussed.

Keywords: special education teacher evaluation, explicit instruction, observation systems, Many-facet Rasch measurement

Developing an Explicit Instruction Special Education Teacher Observation Instrument

Explicit instruction was recently identified as one of 22 high leverage practices (HLPs) for students with disabilities by the Council for Exceptional Children (McLeskey et al., 2017). Explicit instruction is an instructional approach that is highly effective for students with high incidence disabilities (SWD) and is supported by nearly 50 years of research (Hughes, Morris, Therrien & Benson; 2017; Stockard, Wood, Coughlin & Rasplica Khoury, 2018). Despite the strong evidence base supporting the relationship of explicit instruction with higher achievement for SWD in both reading (Baker, Gersten, Haager, & Dingle, 2006; Smolkowski & Gunn, 2012; Stockard et al., 2018) and math (Doabler et al, 2017; Gersten et al., 2009; Stockard et al., 2018), observation studies of special education instructional practice suggest that explicit instruction may not be implemented on a large scale (Ciullo, Lembke, Carlisle, Newman Thomas, Goodwin & Judd, 2016, McKenna et al., 2015; Swanson, 2008).

Teacher Observation and Evaluation Systems

Teacher evaluation systems that include observation of teacher practice are seen as a promising way to close the research to practice gap because they have the potential to evaluate and provide teachers with feedback on how to improve instruction. Emerging analyses of teacher observation systems suggest that when teachers are objectively evaluated and supported to improve instruction, there is a positive impact on student growth (Biancarosa, Bryk, & Dexter, 2010; Taylor & Tyler, 2012). Observation systems therefore, can be conceptualized as serving two primary purposes: 1) evaluating performance to make high-stakes decisions regarding a teacher's employment status, and 2) providing structured feedback to improve a teacher's instructional practice (Adnot, Dee, Katz & Wyckoff, 2017). Within most state and district evaluation systems, observation tools are used for both purposes (Anderson, Butler, Palmiter &

Arcaira, 2016; Herlihy et al., 2014). However, many states are using observation tools designed primarily for evaluation purposes, which tend to rely on general instruments that purport to capture the common, broad characteristics of effective teaching regardless of content, grade level or student population (Anderson et al., 2016; Blazar, Braslow, Charalambous & Hill, 2017).

General observation instruments have been criticized for their limited alignment with the best practices within the relevant instructional or content area. Criticisms center around 1) the constraint on the quality of feedback that can be provided (Anderson, et al., 2016; Grossman, Compton, Igra, Ronfeldt, Shahan & Williamson, 2009; Johnson & Semmelroth, 2014) and 2) the limited overlap between general and content specific factors (Blazar et al., 2017; Kane & Staiger, 2012; Lockwood, Savitsky & McCaffrey, 2015; McClellan, Donoghue & Park, 2013). Special education teachers routinely report perceiving a misalignment of general observation tools with best practices for SWD (Anderson et al., 2016; Holdheide, 2013) and researchers have argued that general teacher observation systems may not be well suited for special education teachers (Johnson & Semmelroth, 2014; Jones & Brownell, 2014).

For example, in an analysis comparing Danielson's Framework for Teaching (FFT; Danielson, 2011) with the elements of effective special education practice, Jones and Brownell (2014) reported that explicit instruction, an HLP in special education, was absent from FFT. They also reported a misalignment in the instructional characteristics to receive a score of 'Distinguished' on FFT which emphasizes a constructivist approach to learning, and the characteristics of explicit instruction, which is more teacher-directed, repetitive and systematic. If special education teachers are evaluated with general instruments that do not capture the elements of explicit instruction, and in fact, emphasize instructional practices that are inconsistent with this HLP, it is unlikely that explicit instruction will be implemented with the

fidelity required to improve outcomes for SWD. There is a need to develop content specific

observation tools for special education teachers aligned with HLPs.

Recognizing Effective Special Education Teachers (RESET) Observation System

RESET is a federally funded project to create teacher observation rubrics aligned with

HLPs for SWD. The goal of RESET is to leverage the research on instructional practices to

develop observation instruments that can be used to evaluate special education teacher

effectiveness and to improve instruction. RESET was developed using the principles of

Evidence-Centered Design (ECD; Mislevy, Lukas & Almond, 2003) and consists of 21 rubrics

that detail instructional practices organized in three categories, 1) instructional methods, 2)

content organization and delivery, and 3) individualization. A complete description of how the

RESET rubrics were developed is provided elsewhere (see Johnson, Crawford, Moylan &

Zheng, 2018). In this study, the focus is on the *Explicit Instruction* rubric, which has been

designed to evaluate and support teachers' ability to effectively implement explicit instruction.

To begin rubric development, the critical components of explicit instruction were

extracted from the literature, then reviewed and synthesized into a coherent set of elements.

Across studies, the detailed descriptions of explicit instruction vary, which limits our

understanding of the specific elements that impact student achievement. However, a recent

synthesis of the research on explicit instruction identified five essential and seven common

components. The five essential components include 1) segment complex skills, 2) modeling and

think-alouds, 3) systematically faded supports/prompts, 4) opportunities to respond and receive

feedback, and 5) purposeful practice opportunities. The additional, common components include

1) selecting critical content, 2) sequencing skills logically, 3) ensure students have prerequisite

skills and background knowledge, 4) provide students with a clear statement of goals, 5) present

a wide-range of examples and non-examples, 6) maintain a brisk pace, and 7) help students to

organize knowledge (Hughes et al., 2017).

A set of detailed items to describe the proficient implementation of these components of

explicit instruction were developed. Because the purpose of RESET is to both evaluate and

provide feedback to special education teachers, we developed a set of scoring rules that define

and describe varying levels of implementation for each item (e.g. proficient implementation,

partial implementation, not implemented; Crawford, Johnson, Moylan & Zheng, in press).

Generalizability theory studies conducted on the two versions of the rubric (one with general

descriptors and one with detailed descriptors) indicate improvements to the overall g coefficient

when detailed descriptors were used (from .61 to .74; Crawford et al., in press). Although the

final version of the rubric (see Appendix A) reflects the current research on explicit instruction,

over time, it will be important to understand whether *all* of the items on the rubric are critical,

and to investigate whether there are items that account for greater variance in student growth.

In addition to better understanding how the items of the rubric function, it is also

important to remember that raters who observe teacher practice also play a critical role in the

observation and evaluation process. The explicit instruction rubric is a high-inference

observation instrument, designed to capture a complex instructional practice and to be used by

observers with high levels of expertise. As a result, it can be difficult to obtain consistent

interpretation and application of the scoring criteria to observations of multiple teachers' lessons

across multiple raters scoring multiple items. In fact, it has been reported 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, Lockwood & MCCaffrey, 2015).

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). Research on rater behavior suggests that

achieving perfect agreement across 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, 2011; Linacre, 1994).

Many-faceted Rasch measurement (MFRM) is an approach to data analysis that allows

for the investigation of multiple facets (e.g. teachers, lessons, items, raters) of a complex

performance assessment to understand how these facets function within the measurement

process, and to examine their interactions. For example, with an MFRM analysis it is possible to

model two aspects of rater behavior: 1) severity, and 2) stochastic differences. One can also

investigate bias interactions among raters and other facets of the observation, such as

rater/teacher interactions or rater/item interactions (Linacre, 1994). In MFRM analyses, rater

behavior is captured through a "severity" parameter, and that parameter characterizes the rater in

the same way that an ability parameter characterizes the teacher being evaluated, and a difficulty

parameter characterizes an item of the rubric (Linacre, 1994). MFRM also reports on the amount

of error that raters display. All raters are expected to demonstrate some degree of error, but too

much threatens the validity of the evaluation (Linacre, 1994). By examining rater severity, error,

and bias, MFRM analyses provide important insights that can be used to improve rater training

efforts, leading to more consistent evaluations and feedback over time (Wigglesworth, 1993).

MFRM analysis also 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). Examining these statistics at the item level allows assessment developers to understand the extent to which items accurately measure teachers along the full continuum of the construct (in this case, their ability to implement explicit instruction). The in- and out-fit statistics at the item level also inform whether construct irrelevant variance may be problematic for certain items. This allows assessment developers to revise or eliminate items that are not functioning as intended. If evaluation and feedback provided through the use of observation rubrics are meant to drive changes in instructional practice, it is imperative that the rubrics contain the 'right' items, and that a teacher's performance not be entirely dependent upon the rater who is observing the lesson.

Teacher observation instruments are high stakes assessments because they are used to make critical decisions about teachers and more importantly, because they should be used to inform and improve the quality of instruction that students receive. Considering the stakes, we argue, as others have (e.g., Bell, Gitomer, McCaffrey, Hamre, Pianta, & Qi, 2012; Herlihy et al., 2014), that it is imperative to apply the same assessment standards to teacher evaluation and observation systems as have been applied to other areas of educational assessment. Developing valid and reliable teacher observation tools is essential to ensure that decisions about teacher effectiveness are fair, and that teachers are provided with consistent and meaningful feedback on how to improve (Hudson, 2015; Johnson & Semmelroth, 2015). Over time, observation tools that are aligned with specific instructional practices can also contribute to our understanding of the elements of practice that have the most impact on student outcomes. To meet these demands, observation instruments require a deliberate development approach and a rigorous evaluation of their psychometric properties, including item level analyses. Therefore, the purpose of this study was to examine the psychometric quality of the Explicit Instruction rubric.

Methods

Participants

Special education teachers. Thirty special education teachers from three states participated in this study. 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. Finally, teachers worked across a variety of school settings. Table 1 provides the demographics of the schools in which data was collected.

Raters. One male and fourteen female raters were recruited from seven states. 12 raters were White, 2 Asian, and 1 Pacific Islander. Criteria for raters included having five or more years of experience working with SWD. 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

Video collection. All special education teacher participants were asked to video record weekly lessons with a consistent group of students using the Swivl® video capture and upload

system. To decide on appropriate lessons for recording, research project staff contacted each teacher to discuss the lessons they were planning to record. Based on the information provided, the teachers then targeted a specific instructional group to record. Research staff viewed the first lesson submitted to ensure that it reflected explicit instruction. Teachers were sent a short video and set of instructions that demonstrated how to use the Swivl ® system and were provided with project staff contact information for technical support. The Swivl ® system is a small 'bot' that rotates a tablet recorder when paired with a Bluetooth ® audio marker that the teacher clips on. The marker is a microphone for high quality audio, and also allows for a video recording that 'follows' the teacher as she moves around the classroom. All teachers were provided with an Asus Nexus 7 Model K008 with 16 GB of memory.

The lessons ranged in length from 20-50 minutes. Each teacher contributed a total of 20 videos over the school year. Videos are used by the RESET research team to test and refine the rubrics that comprise the RESET observation system. Three videos from each teacher were selected, resulting in a total of 90 videos for this study. Observation studies have demonstrated that between two to four lessons are needed for reliable observations of a teacher's instruction (Hill, Charalambos, & Kraft, 2012; Johnson & Semmelroth, 2015). The first video (Lesson 1) was the initial video collected at the beginning of the year. The second video (Lesson 2) was selected from the middle of the school year (submitted between January – February; selections were based on audio and video quality), and the third video (Lesson 3) was the final video that teachers submitted at the end of the school year. These video selection criteria (e.g. three observations, across a school year) were adopted because they are consistent with how an administrator would typically use the observation system. Each video was assigned an identification number and listed in random order for each rater to control for order effects.

Rater training. Over a four-day training period, raters were provided with an overview of the RESET project goals and a description of how the rubric was developed. Research project staff then explained each item of the EI rubric and clarified any questions the raters had. Raters were provided with a training manual that includes 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. Disagreements in scores were reviewed and discussed. Raters were then assigned a randomly ordered list of videos and asked to evaluate the videos in the assigned order, to score each item, to provide time stamped evidence 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 observation load, we developed a rating scheme that allowed for scores across raters and videos to be linked 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. One rater was randomly selected to score at least one video of each teacher. Remaining videos were randomly assigned and 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.

Data Analysis

We first examined reliability by calculating the internal consistency and exact agreement across raters. Then data were analyzed through many-faceted Rasch measurement (MFRM) analyses. Observation instruments typically achieve relatively low levels of exact agreement across raters (see for example, Cash, Hamre, & Pianta, 2012; Kane & Staiger, 2012). 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. The model used for the MFRM analysis in this study is given by:

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

where P_{nijok} is the probability of teacher n, when rated on item i by judge (rater) j on occasion (lesson) o, being awarded a rating of k. $P_{nijo(k-1)}$ is the probability of teacher n, when rated on item i by judge 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 3.71 (Linacre, 2014). 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; Englehard, 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. Finally, MFRM allows for bias analysis of the scores to examine the discrepancy between

observed and expected scores according to the severity levels of the raters. In this study, the

biased interactions between teachers and raters were examined. Significant differences between

expected and observed scores ($p < .05$) indicate the presence of bias (Linacre, 2014).

Results

The internal consistency of the items was .929, and exact rater agreement was 51%. The

results of the analysis are shown in Figure 1 and Tables 2 through 5. All analyses are based on a

total of 10,010 assigned scores. Category statistics showed that of the assigned scores, 40% were

a 3 (implemented), 51% were a 2 (partially implemented), and 9% were a 1 (not implemented).

Item Difficulty and Fit Statistics

Figure 1 is the Wright map which plots the measures for the four facets (a) item, (b)

teachers, (c) raters, and (d) lessons on a common scale. The scale along the left of Figure 1

represents the logit scale, ranging from -2 to +2, which is estimated from the pattern of the data.

Placing the facets on a common scale allows for the comparisons within and among the facets

items, teachers, raters and lessons (Smith & Kulikowich, 2004).

The column heading for "Items" ranks the items from most to least difficult, with the

lowest scoring items (Item 3, 7, and 13) at the top and the highest scoring items (Item 19 and 23)

at the bottom. Table 2 shows the analysis report for the item measures. As shown in Figure 1 and

supported by the data in Table 2, Item 3 on the rubric (*The teacher clearly explains the relevance*

of the stated goal to the students), with a logit value of .91 was the most difficult item, and Item

19 (*The teacher provides frequent opportunities for students to engage or respond during the*

lesson), with a logit value of -.59, was the least difficult item. Table 2 provides information about

the fit and separation of the items. Item fit statistics indicate whether raters have scored items in

a consistent manner. Items that are not scored in a consistent manner may need to be removed or

revised. The fit statistics for all of the items are within the acceptable range, which means that raters consistently scored easier items with higher scores, and more difficult items with a lower score. The item reliability of separation of .98 demonstrates that item difficulties are separated along the continuum of difficulty of explicit instruction implementation. This separation was statistically significant with a chi-square of 1262.2 and 24 degrees of freedom (*p< .001)*.

The item difficulty ranking provides information to assess the construct validity of an instrument, through the evaluation of the logic of the ordering of items (Smith, 2001). In examining the items on the rubric (see Appendix A) and their ranked order on the variable map, as well as the overall percentages of scores received for each item the rank order seems logical. For example, item 3 includes explaining the relevance of the stated learning objective to students. Across the total number of times this item was scored, only 14% of possible responses were scored as implemented, 44% as partially implemented, and 42% scored as not implemented. When reviewing the raters' explanations and evidence for scoring this item as partially implemented, most comments reflected that relevance was stated in a very general or vague way such as "That's why we are doing more examples, so more students get it" or "Learning common multiples is new and difficult". Item 7 focuses on reviewing prior skills and engaging background knowledge, with 24% scored as implemented, 55% as partially implemented and 21% as not implemented. Rater comments for scoring items as partially implemented primarily focused on the ineffectiveness of the review with comments such as, "The teacher reviews the question answering strategy at the start of the lesson, but does it very quickly, and as a result, she needs to reteach to her students throughout the lesson, because they did not have the strategy down". Item 13 includes the systematic withdrawal of teacher support with 26% scored as implemented, 52% as partially implemented, and 22% as implemented. Rater

comments related to partially implemented scores included "The teacher started by leading the discussion of how to solve the math problem, then she has the students do one on their own and explain what they did. The students' lack of understanding was evident, so the teacher had to jump back in – she was missing the 'we do' part of practice that is so important."

The 'easier' items included items 19 and 23. Item 19 is *The teacher provides frequent opportunities for students to engage or respond during the lesson*. 58% of all responses were scored as implemented, 41% as partially implemented and less than 1% as not implemented. This item focused on frequency, and not quality, and nearly all of the lessons (90%) were conducted in teacher:student ratios of 1:6 or fewer. Evidence used to support implementation of this item included the use of frequent questioning, providing students with opportunities to practice a skill, and engaging students through multiple response techniques such as writing on a white board or giving a thumbs up when they knew the answer. Item 23 is *The teacher provides timely feedback throughout the lesson*. 57% of all responses were scored as implemented, 41% as partially implemented and 2% as not implemented. As is the case with item 19, the focus of this item is on the immediacy of the feedback and not the quality. Item 24 focuses on the specificity of the feedback, and this item was more difficult for teachers than was item 23, with 34% implemented, 61% partially implemented, and 5% not implemented.

Teacher Proficiency and Fit Statistics

The teacher column on Figure 1 lists the teachers from most proficient (Teacher 5) at the top to least proficient (Teacher 11 and 17) at the bottom. Teachers who are more proficient are expected to score higher than teachers who are less proficient on items that are more difficult. Table 3 gives fit and separation information for the teacher facet. The reliability of separation is .98, with a statistically significant chi square of 1796.3 and 29 degrees of freedom (*p<.001*).

This indicates that teachers differ in their ability to proficiently implement explicit instruction as measured by this rubric, beyond what can be attributed to measurement error. The fit statistics measure the extent to which a teacher's pattern of responses matches that predicted by the model, and therefore can be used to identify teachers who have not been evaluated in a consistent manner or for whom the rubric is not appropriate (i.e. the teacher's lesson may not have been delivered using explicit instruction). Table 3 shows that all fit statistics are within acceptable ranges (+/- .5 to 1.5), suggesting that the evaluation with the rubric has been consistently applied to determine teachers' ability to implement explicit instruction.

Rater Severity and Fit Statistics

The rater column ranks the raters from most severe (Rater 9) at the top to the most lenient (Raters 13 and 14) at the bottom. The fit statistics help to determine whether raters are consistent with their own ratings on the rubric and can be used to identify severe or lenient ratings that are not expected given a rater's overall scoring pattern, or used to identify biases for a particular item or teacher. Fit values greater than 1 show more variation than expected (misfit), and values less than one show less variation than expected in their ratings (overfit). Misfit is generally thought to be more problematic than overfit (Myford & Wolfe, 2003). The fit statistics for raters are within the acceptable range (.5 to 1.5; Linacre, 2008). The reliability of separation of .97, on a chi-square of 659, degrees of freedom 14, is significant (p < .001) and along with the spread from -.31 to .52 logits suggests that raters differ in their overall ratings and severity level.

MFRM analyses can account for differences in rater severity by adjusting the observed score and computing an average fair score for teachers. A fair score is the score that a particular examinee would have obtained from a rater of average severity (Eckes, 2011). Table 3 includes a comparison of a teacher's average observed score across all items, lessons and raters who

observed them and their fair average score. There are minimal differences between the observed

and fair average scores, with no set of scores resulting in a different level of proficiency rating

for a teacher. Additionally, while there are some differences in the rank ordering of teachers

based on observed versus fair average scores, there are no changes in the identification of the top

10% (Teachers 5, 10, 4) or the bottom 10% (6, 17, 11) of performers.

Lesson Rating and Fit Statistics

The lesson facet is somewhat difficult to interpret because we did not specify the content

or focus of the lessons. Figure 1 and Table 5 shows that each of the three lessons were

approximately of the same difficulty, ranging from -.07 to .04 logits. The reliability of separation

of .89 is statistically significant ($p < .001$), suggesting that lesson 'difficulty' differed across the

three time periods, with lesson 2 being the highest scoring lesson on average. The fit statistics for

the lesson facet indicate interrater consistency for that lesson evaluation. This finding is

consistent with teacher observation studies that report a need for multiple observations of a

teacher (Hill et al., 2012; Johnson et al., in press; Kane & Staiger, 2012; Lei, Li & Leroux, 2018)

<div align="center">

Discussion

</div>

The results of our analyses suggest that we have developed an Explicit Instruction

observation rubric that can be used to provide consistent evaluations of a special education

teacher's ability to effectively implement this HLP. As indicated throughout this study, teacher

observation is a complex performance assessment that is influenced by multiple factors that must

be accounted for when determining the psychometric quality of an observation instrument.

Teachers provide a performance (video recorded lesson) designed to represent the underlying

construct (in this case, implementation of explicit instruction), and raters judge the quality of the

lesson based on their understanding of that construct and the use of a detailed scoring rubric

(items). This evaluation process highlights the need to carefully investigate the psychometric quality of complex, multi-faceted performance assessments to ensure that teacher observation systems are designed and implemented in a manner that results in fair and effective use.

Capturing the elements that comprise the HLP of explicit instruction at a level of specificity that will be helpful for teachers and that results in consistent evaluation and scoring by raters was challenging. Our findings indicate that the 25 items as currently detailed on the Explicit Instruction Rubric resulted in their consistent evaluation across raters, and that the rank order of difficulty of the items was logical. This represents a critical first step in developing a valid, content-specific observation instrument for special education teachers. Although we could not find research to inform a desired number of items for an observation instrument, our work to date and feedback received from raters suggests that a 25-item rubric is too lengthy to feasibly use in practice. Given the clustering of some items in terms of their difficulty levels, it is likely that the Explicit Instruction rubric can be shortened and still result in reliable evaluation of a teacher's proficiency with this practice. However, it will be important to better understand the level of specificity needed to provide actionable feedback to teachers, and continued research that examines the relationship of teacher performance on the items with student growth and outcome measures will also inform which items may have greater predictive value, and whether there are items that do not measurably impact student achievement.

Our results are consistent with teacher observation research indicating that multiple observations of a teacher are needed to result in reliable evaluations of teacher practice (Hill et al., 2012; Ho & Kane, 2013; Semmelroth & Johnson, 2014). The levels of exact agreement reported in this study are also consistent with those reported in teacher observation research that have employed observation rubrics with three and four level scales (Cash et al., 2012; Kane &

Staiger, 2012). This is an important finding as these larger studies employed rater certification

requirements, where raters were not allowed to score unless they were able to meet required

thresholds of agreement with master scored videos. Our project is not equipped to create this

type of rater certification process, and yet, the resulting levels of agreement are similar.

The role of the rater in teacher observation systems is an important consideration for their

development because the scores provided are a function not only of the teachers' ability but also

of the severity and consistency of the raters evaluating them. A rater's understanding of the

observed construct and the way it is defined and explained in the associated items can impact the

way a teacher's performance is evaluated, resulting in construct irrelevant variance (Messick,

1994), and can threaten the validity and fairness of teacher evaluations. To control for this, rater

training was designed to develop common understandings of the rubric items, and we provided a

detailed manual that included specific examples of levels of implementation for every item.

While these are necessary components of rater training, the confound of observed scores with

teacher proficiency and rater severity presents a non-trivial problem.

Importantly, the adjustments in the MFRM analysis that result from using the fair average

instead of the observed score show that in this data, no changes to a teacher's overall categorical

evaluation occurred (e.g. implemented, partial implemented, or not implemented), and there were

minimal changes in the overall rank order of teachers, with no changes to the top 10% or bottom

10% performers, criteria which are routinely used to make merit and retention decisions. As part

of instrument development, MFRM analyses offer important insights into rater behavior, but of

course the challenge will be the translation of these processes from research to practice. To

ensure fairness to teachers, it may be necessary to employ similar analyses to observations used

to inform high-stakes decisions. Given the likely limited capacity to conduct these analyses, an

additional investment to partner with psychometricians or evaluators familiar with these procedures may present an additional but necessary cost to developing a sound evaluation process.

The results of the present study are promising, however, there are limitations that warrant caution in the generalization of results. The most significant limitation is that the sample sizes of both special education teachers (n = 30) and raters (n = 15) are small, and somewhat limited in their representativeness of the larger population of special education teachers and potential raters (e.g. nearly all participants were White females). One benefit of using video observations however, is that over time, we can develop a video bank that will include a larger and more diverse pool of teachers. Continued studies with larger samples of teachers and raters can be conducted to verify the results of the studies reported in this study. Additionally, although our larger pool of RESET teacher participants includes teachers across the K-12 grade levels, to test the Explicit Instruction rubric, only second through eighth grade level teachers could be included, as there were no videos at the high school level that displayed explicit instruction.

Implications for Research

We believe that our findings have important implications for the continued development of content specific observation rubrics aligned to the HLPs for SWDs. First, the identification of explicit instruction as a HLP and the common finding across observation studies that it is not widely used in practice suggests a need for a special education teacher evaluation system that reflects the practices found to be effective in improving the achievement of students with disabilities. If the goal of teacher evaluation through observation instruments is instructional improvement, then it is important to provide teachers and administrators with specific information that can lead to individualized professional development support (Blazar et al.,

2017). The findings of this study suggest that the explicit instruction rubric could provide teachers with a consistent benchmark against which they can engage in continuous improvement.

Continued investigation of the rater's role in teacher observation and the factors that influence scoring decisions is critical. Limited research on content specific observation rubrics in mathematics suggest that raters with strong content knowledge are more accurate in scoring than are those without (Hill et al., 2012), but it is unclear whether this is the case for the Explicit Instruction rubric. Understanding the way that raters interpret the rubric items and match it to observed evidence within the video recorded lesson may also provide important insight into both the wording of items as well as to the development of training that leads to greater consistency in scoring. Finally, the current Explicit Instruction rubric is long, and rater fatigue is a well-documented phenomenon in observation research (Casabianca, Lockwood & McCaffrey, 2015). Continued research to identify items that are strongly predictive of student growth can help focus attention to a smaller set of elements.

Researchers and practitioners cannot afford to ignore the effectiveness research on explicit instruction (Stockard et al., 2018), and the critical need to support special education teachers in proficient implementation of this HLP. The RESET rubric offers one way to help close the research to practice gap by providing teachers with detailed descriptions of how to effectively implement explicit instruction.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and

 student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54-76.

Anderson, L. M., Butler, A., Palmiter, A., & Arcaira, E. (2016). *Study of emerging teacher*

 evaluation systems. Prepared for: Policy and program studies service. Office of Planning,

 Evaluation and Policy Development, Washington, DC.

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading

 growth of first-grade English learners: Validation of an observation instrument. *The*

 Elementary School Journal, 107 (2), 199-219.

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument

 approach to observation protocol validity. *Educational Assessment, 17*, 62-87.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy

 collaborative professional development on student learning. *The Elementary School Journal,*

 111, (1), 7-34.

Blazar, D., Braslow, D., Charalambos, Y. C., & Hill, H. C. (2017). Attending to general and

 specific dimensions of teaching: Exploring factors across two observation instruments.

 Educational Assessment, 22(2), 71-94.doi: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.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when

observational assessment occurs at large scale: Degree of calibration and characteristics associated with calibration. *Early Childhood Research Quarterly*, *27*(3), 529-542.

Ciullo, S., Lembke, E. S., Carlisle, A., Newman Thomas, C., Goodwin, M., & Judd, L. (2016). Implementation of evidence-based literacy practices in middle school response to intervention: An observation study. *Learning Disability Quarterly, 39* (1), 44-57.

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (in press). Sources of variance in two versions of a special education teacher observation rubric. *Assessment for Effective Intervention*.

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

Doabler, C. T., Clarke, B., Stoolmiller, M., Kosty, D. B., Fien, H., Smolkowski, K., & Baker, S. K. (2017). Explicit Instructional Interactions: Exploring the Black Box of a Tier 2 Mathematics Intervention. *Remedial and Special Education*, *38*(2), 98-110.

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.

Ewing, M., Packman, S., Hamen, C., & Thurber, A. C. (2010). Representing targets of measurement within Evidence-Centered Design. *Applied Measurement in Education, 23 (4)* 325-341. doi:10.1080/08957347.2010.510959.

Gersten, R. M., Chard, D., 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*, 1202–1242.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., & Williamson, P. (2009). Teaching practice:

A cross-professional perspective. *Teachers College Record, 111*(9), 2055-2100.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S.

(2014). State and local efforts to investigate the validity and reliability of scores from

teacher evaluation systems. *Teachers College Record*, *116*(1), 1-28.

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. doi:10.3102/0013189X12437203.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities

posed by new teacher evaluation systems. Harvard Educational Review, 83(2), 371-384.

Holdheide, L. (2013). *Inclusive design: Building educator evaluation systems that support students

with disabilities*. Avalable from www.gtlcenter.org

Hudson, P. (2015). Feedback consistencies and inconsistencies: eight mentors' observations on

one preservice teacher's lesson. *European Journal of Teacher Education, 37*(1), 63-73

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction:

Historical contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140-148.

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. & Semmelroth, C. L. (2015). Validating an observation protocol to measure

special education teacher effectiveness. *Journal of the American Academy of Special Education

Professionals*

Johnson, E. S., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it

matters, what makes it challenging. *Assessment for Effective Intervention*. *39* (2), 71-82.

Jones, N. D., & Brownell, M. T. (2014). Examining the Use of Classroom Observations in the

Evaluation of Special Education Teachers. *Assessment for Effective Intervention*, *39*(2), 112–

124. doi:10.1177/1534508413514103

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*.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating

scales. *Language Testing*, *26*(2), 275-304.

Lei, X., Li, H., & Leroux, A. J. (2018). Does a teacher's classroom observation rating vary across

multiple classrooms?. *Educational Assessment, Evaluation and Accountability*, *30*(1), 27-46.

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

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

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective

teaching from classroom observations: An application of Bayesian exploratory factor

analysis without restrictions. *The Annals of Applied Statistics, 9,* 1484-1509.

McClellan, C., Donoghue, J., & Park, Y. S. (2013). *Commonality and uniqueness in teaching

practice observation*. Retrieved from http://www.clowderconsulting.com

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.

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*.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13-23.

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.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617-639.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the Classroom Observations of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly, 44*, 48–57.

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, ## (X)*, 1-29.

Swanson, E. A. (2008). Observing reading instruction for students with LD: A synthesis. *Learning Disability Quarterly, 31*, 1–19. doi:10.1177/0022219411402691

Taylor, E. S. & Tyler, J. H. (2012). Can teacher evaluation improve teaching? Evidence of systematic growth in the effectiveness of teachers. *Education Next, 12,*(4), 78-84.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, *27*(2), 237-246.

Wright, B. D., & Stone, M. H. (1999). Measurement essentials. *Wilmington*. *Wide Range Inc*, 221.

Table 1

School Demographics

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

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

*These schools each had 2 teacher participants

Table 2

Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
19	-.59	.05	.81	.84
23	-.49	.05	.99	1.06
5	-.34	.05	.93	.95
21	-.33	.05	.90	.95
6	-.32	.05	1.10	1.13
20	-.27	.05	.90	.90
17	-.25	.05	.87	.92
10	-.13	.04	.90	.95
4	-.10	.04	.86	.87
18	-.07	.04	.77	.76
14	-.07	.04	1.05	1.05
11	-.06	.04	.85	.87
22	-.04	.04	.89	.92
15	-.03	.04	.91	.95
16	.01	.04	1.00	1.07
24	.03	.04	.89	.92
12	.09	.04	.96	.96
1	.15	.04	1.30	1.33
25	.19	.04	.96	.99
9	.24	.04	1.00	1.02
2	.27	.04	1.25	1.31
8	.28	.04	1.02	1.02
13	.46	.04	1.15	1.13
7	.46	.04	1.21	1.19
3	.91	.04	1.29	1.23
Mean (count = 25)	.00	.04	.99	1.01
SD	.49	.00	.15	.14

Note. Root mean square error (model) = .04; adjusted *SD* = .32; separation = 7.16;

reliability = .98; fixed chi-square = 1262.2; df = 24; significance = .00.

Table 3

Teacher Measure Report from Many-Facet Rasch Measurement Analysis

Teacher Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ	Observed Average	Fair Average
5	1.53	.08	1.17	1.11	2.77	2.79
10	1.14	.06	1.17	1.11	2.58	2.66
4	1.06	.06	1.12	1.03	2.60	2.63
1	.93	.03	1.01	1.07	2.53	2.56
7	.86	.06	1.30	1.18	2.49	2.50
8	.80	.06	1.12	1.10	2.48	2.49
9	.73	.06	1.42	1.35	2.46	2.46
25	.73	.05	1.10	1.06	2.48	2.45
28	.72	.05	1.19	1.24	2.42	2.42
24	.67	.05	.71	.69	2.34	2.41
12	.62	.05	1.23	1.40	2.42	2.39
13	.62	.05	1.11	1.04	2.41	2.38
22	.58	.05	1.17	1.23	2.38	2.37
26	.56	.05	1.09	1.09	2.35	2.34
16	.50	.03	.88	.87	2.32	2.34
19	.44	.05	.88	.87	2.30	2.29
21	.40	.05	1.10	1.14	2.30	2.25
2	.37	.05	.86	.87	2.24	2.24
14	.34	.05	1.03	1.05	2.24	2.23
23	.28	.05	1.03	1.04	2.26	2.21
20	.20	.05	.98	.97	2.20	2.16
29	.18	.05	.75	.75	2.11	2.11
18	.15	.05	1.04	1.03	2.05	2.11
27	.12	.05	.75	.73	2.10	2.09
3	.11	.05	.81	.80	2.06	2.08
15	.09	.05	1.16	1.15	2.07	2.06
30	.04	.05	.89	.89	2.09	2.04
6	-.13	.05	.84	.85	1.95	1.90
17	-.28	.05	.88	.86	1.80	1.85
11	-.34	.05	.93	.92	1.80	1.79
Mean	.47	.05	1.02	1.02	2.29	2.29
SD	.42	.01	.17	.16	.23	.23

Note. Root mean square error (model) = .05; adjusted *SD* = .41; separation = 8.03;

reliability = .98; fixed chi-square = 1796.3; df = 29; significance = .00.

Table 4

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 5

Lesson Measure Report from Many-Facet Rasch Measurement Analysis

Lesson Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
2	-.07	.02	1.07	1.08
1	.03	.02	1.01	1.03
3	.04	.01	.94	.93
Mean (count = 3)	.00	.02	1.00	1.01
SD	.05	.00	.07	.07

Note. Root mean square error (model) = .02; adjusted *SD* = .04; separation = 2.88;

reliability = .89; fixed chi-square = 26.4; df = 2; significance = .00.

Measr	-Items	+Teacher	-Rater	-Lesson	Scale
2					(3)
		Teacher 5			
		Teacher 10, 4			____
1	3	Teacher 1, 7			
		Teacher 8			
		Teacher 24, 25, 28, 9			
		Teacher 12, 13, 22, 26			
	13,7	Teacher 16	Rater 9		
		Teacher 19, 2, 21			
	2,8	Teacher 14, 23	Rater 3		
	1, 25, 9	Teacher 20, 29	Rater 15, 4, 6		
	12	Teacher 15, 18, 27, 3			
0	15, 16, 22, 24	Teacher 30	Rater 10, 5, 8	1, 3	2
	10, 11, 14, 18	Teacher 6	Rater 1, 12	2	
	17		Rater 11, 2, 7		
	20, 21, 5, 6	Teacher 11, 17	Rater 13, 14		
	23				
	19				
-1					____
-2					(1)

Figure 1. *Variable map of the EI rubric facets items, teachers, raters, and lessons.*