Validity of a Special Education Teacher Observation System

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

Boise State University

May 2019

Author Note

Evelyn S. Johnson, Department of Early and Special Education, Boise State University; Angela Crawford, Project RESET, Boise State University; Laura A. Moylan, Project RESET, Boise State University; Yuzhu Zheng, 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 S. Johnson, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: evelynjohnson@boisestate.edu

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

Abstract

This manuscript describes the comprehensive validation work undertaken to develop the Recognizing Effective Special Education Teachers (RESET) observation system, which was designed to provide evaluations of special education teachers' ability to effectively implement evidence-based practices and to provide specific, actionable feedback to teachers on how to improve instruction. Following the guidance for developing effective educator evaluation systems, we employed the Evidence-Centered Design framework, articulated the claims and inferences to be made with RESET, and conducted a series of studies to collect evidence to evaluate its validity. Our efforts and results to date are described, and implications for practice and further research are discussed.

Keywords: special education teacher evaluation, observation systems, validity argument

Validity of a Special Education Teacher Observation System

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 (SWD). In our work, we define students with high incidence disabilities as those with mild emotional/behavioral disorders, learning disabilities, high functioning autism, other health impairment (ADHD) or language impairment.

The goal of the RESET project is to leverage the extensive research on EBPs for this population of students and to develop rubrics 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 ultimately, (c) improve outcomes for SWD. Assessment systems are intended to facilitate defensible decisions about the people being assessed. In the case of RESET, the decisions to be made include identifying a teacher's baseline level of performance, providing targeted feedback based on the teacher's observed strengths and weaknesses as measured by their item level scores on the rubrics, and determining whether a teacher has demonstrated sufficient growth in her implementation of the practice.

Validity determines the extent to which the decisions made from an assessment are defensible and is considered the foundational principle that guides the development, administration, interpretation, evaluation and use of tests (Standards for Educational and Psychological Testing, 2014). The validation of an assessment system has been described as a process that begins with a clearly articulated theory of action, followed by a statement of the proposed use of the assessment that leads to a carefully planned argument defining its key claims

and assumptions, and finally, proceeds with the collection and organization of evidence into a substantiated validity argument (Bell, Gitomer, McCaffrey, Hamre, Pianta & Qi, 2012; Cook, Brydges, Ginsburg & Hatala, 2015; Kane, 2006; Kane 2013). One approach to this process is Kane's (2006) argument-based validity approach, consisting of the interpretive/use argument, and the validity argument. The interpretive/use argument (IUA) presents the "network of assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2013, p. 23). The validity argument evaluates those assumptions through the collection of empirical data and analytic reasoning. In this manuscript, we present the theory of action and the IUA for RESET, then describe the process of validation both undertaken and planned, discuss some of the challenges encountered, and conclude with implications for further research.

RESET Conceptual Framework and Theory of Action

The theory of action underlying RESET rests on the idea that improving teacher practice depends upon 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 rubrics 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 results in accelerated student growth. The theory of action provides testable assumptions about RESET that allow us to determine the extent to which it achieves these goals.

RESET was developed using the principles of Evidence-Centered Design (ECD; Mislevy, Steinberg & Almond, 2003) and consists of 21 rubrics that detail evidence-based instructional practices organized in three categories: a) instructional methods, b) content organization and delivery, and c) individualization (see Table 1). For each rubric, an extensive synthesis of research was conducted to create sets of items that detail each of the EBPs listed in Table 1 (see Johnson, Crawford, Moylan & Zheng (2018) for a complete description). These items served as the performance level descriptors for proficient implementation, which were refined through an iterative process of testing the items with video recorded lessons and discussing the clarity and utility of each item. We then sent each rubric to subject matter experts for review, synthesized their feedback, and completed revisions to create a final set of items that described proficient implementation of the practice reflected in the rubric.

As is the case with many teacher observation systems, RESET has been conceptualized as eventually serving two primary purposes: (a) evaluating a teacher's ability to effectively implement evidence-based practices, and (b) providing structured feedback to improve a teacher's instructional practice (Adnot, Dee, Katz, & Wyckoff, 2017). Therefore, our validation efforts to date have been centered around RESETs proposed interpretation and use as a vehicle to observe and evaluate a teachers' ability to implement the specific steps of the EBPs.

To use RESET, teachers submit video recordings of their lessons, which are then evaluated by trained raters using the relevant rubric. The scoring rules are based on the special education teacher's level of implementation of each item, evaluated on a three-point scale in which a score of 3 is implemented, a 2 is partially implemented, and a 1 is not implemented. Raters are trained through the use of exemplars and elaborated descriptions and examples of practice at each of the three levels of performance. Raters then view recorded lessons between

20-45 minutes in length, and assign a score for each item on the rubric, citing the evidence they used within the observed lesson to reach their scoring decision. Both item scores and an overall lesson score are reported to the teacher. Given the intended use of RESET, item-level scores are important because we anticipate that teacher performance across items will not be consistent - some items will be well implemented, and some items will not - and different teachers will show different abilities across items. Reporting scores at the item level is intended to provide the teacher with specific, actionable feedback about which elements of the EBP they may need to improve. Lesson scores are based on the average performance across the items to provide an overall assessment of a teacher's ability to implement the specific EBP reflected in the lesson.

To support these uses of RESET, several assumptions need to be met. Following Kane's argument-based approach to validity (Kane, 2004) and its application to teacher observation instruments as described by Hill, Charalambous, & Kraft (2012), and Bell et al. (2012), we organized the assumptions around the four areas of scoring, generalization, extrapolation and decisions, and these assumptions are summarized in Table 2. The articulated assumptions then serve as a blueprint for carrying out a research agenda intended to critically evaluate the extent to which RESET can meet its intended use. In the next section, we describe the studies conducted to date and evaluate the results in light of the stated purposes of RESET, then briefly describe the studies planned or in progress to collect evidence for the remaining assumptions. For each study conducted or planned, we indicate how the results are used to evaluate the various assumptions.

Generalizability Study

After each of the rubrics within the RESET system were drafted, we needed to create the scoring criteria to describe the various performance levels of implementation for each practice. Following the model of the National Professional Development Center on Autism (Wong et al.,

2015), we used the general descriptions of implemented (3), partially implemented (2), and not implemented (1) to correspond with the three-point scale. However, we were uncertain as to the need for developing detailed descriptors to describe partially or not-implemented for each item, or whether the general categories would suffice. Although instruments with context specific descriptors are more time-consuming and expensive to develop and implement, the research suggests they may result in greater reliability (Knoch, 2009; Norris & Borst, 2007), greater construct validity (Knoch, 2009), and more actionable feedback to teachers (Fulcher, Davidson, & Kemp, 2011). We therefore conducted a study to compare two versions of the Explicit Instruction rubric that examined the following research questions through the use of generalizability theory: 1) Do the ratings produced with the two versions of the rubric differ in terms of the relative contribution of sources of variance? 2) Do the ratings produced differ in terms of their indices of generalizability and dependability? And 3) How many raters and lessons are needed to achieve strong levels of dependability? (Crawford, Johnson, Moylan & Zheng, 2018).

The study included a sample of 10 special education teachers from 3 states who each contributed four videos of their instruction, ranging in length from 20-45 minutes (see Crawford et al., 2018 for a detailed description of this study). The videos were evaluated by two sets of raters (four raters in each set) using the Explicit Instruction Rubric from the RESET Observation System. The Explicit Instruction rubric with item-specific performance descriptors demonstrated less unwanted error associated with raters, therefore, we present the results for that version of the rubric only. Table 3 includes the results of the ANOVA and is organized by each facet and facet interaction, and includes the sums of squares (*SS*), degrees of freedom (*df*), mean squares (*MS*), percentage contribution of each source to the total variance, and the standard error associated

with each variance component (*SE*). In our model, the facets include teachers (T), raters (R), items (I), and lessons (L), with lessons nested within teachers, items as a fixed facet, and teachers, raters and items crossed. The variance for teachers (T) shows the amount of systematic variance in teachers' implementation of explicit instruction. Ideally, this component would have the highest variance. Variance related to items (I) is acceptable, as one would expect some items to be more difficult than others. The interaction of teachers with items (TI) indicates that there is potential for RESET to provide item-level diagnostic information for individual teachers, an important finding given its intended use to provide feedback at this level. As can be seen in Table 3, the percentage of variance attributable to the rater facet (R) was 4.5%, showing some promise that the inter-rater and intra-rater scores were consistent with the specific descriptors.

The G-study data are used to compute reliability as the ratio of differentiation variance (the object of measurement, in this case, teachers) to the instrumentation variance (L:T, R, and interactions), expressed in a generalizability coefficient with 4 lessons per teacher and 4 raters per lesson, which was .74 for the overall rating. The item facet was assumed to be fixed. If the system were used operationally with fewer lessons or fewer trained raters, the generalizability coefficient would be expected to be lower.

Generalizability coefficients > .70 are generally considered to be acceptable reliability estimates for observation instruments in the early stages of construct validation research (Erlich & Shavelson, 1978 p. 80; Nunnally & Bernstein, 1994 pp. 264-265). Because the data in this study were ordinal but analyzed as though continuous, these calculations are attenuated and represent lower-bound estimates (Ark, 2015).

Our overall results suggest that with the empirically derived performance level descriptors we were able to achieve levels of reliability acceptable for observation instruments.

The finding that less than 5% of variance was attributable to the rater main effect provides some support for the generalizability assumption (Assumption 2.1) of our interpretive use argument (see Table 2). However, Table 3 shows that a considerable amount of variance is attributable to the combinations and interactions of the teacher, rater and item facets, which warrants closer examination, and these interaction components (particularly the TR and LR:T) interfere with generalizability.

Teacher observation is a form of rater-mediated performance assessment (Eckes, 2011; Engelhard, 2002) in which the raters who observe teacher practice play a critical role in the observation and evaluation process. Teachers record a lesson evaluated through items designed to represent the salient characteristics of the EBP, and raters judge the quality of instruction based on their understanding of the EBP and interpretation of the scoring rules and criteria (Bejar, Williamson & Mislevy, 2006, Eckes, 2011, Gitomer et al., 2014). The RESET rubrics are high-inference observation instruments, each designed to capture a complex instructional practice and to be used by raters with high levels of expertise. As a result, a significant challenge for RESET is to obtain consistent interpretation and application of the scoring criteria to observations of multiple teachers' lessons across multiple raters scoring multiple items. In other studies of teacher observation, 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 & Melinda Gates Foundation, 2011; Gitomer et al., 2014). Raters have been shown to account for between 25 to as much as 70% of the variance in scores assigned to the same lesson (Casabianca, Lockwood & McCaffrey, 2015).

In our analyses, although the main effect for raters accounted for 4.5% of the variance, the interaction of raters with teachers accounted for 6.2%, the raters with lessons accounted for

7.2%, and raters with items also accounted for 7.2%. So the variability associated with the TR and LR:T variance components, and the residual variance (LRI:T) are the main sources of error. Because items are treated as fixed, they are not a source of error. Generalizability analyses can help identify important sources of variation, and then item response theory (IRT) techniques can be used to diagnose specific facets or combinations of them to guide test revision and rater training (Smith & Kulikowich, 2004; Webb, Shavelson & Haertel, 2006).

Many-faceted Rasch Measurement (MFRM) Studies

To more closely examine the rater facet of RESET, we conducted MFRM studies for several of the RESET rubrics. Methods to improve rater reliability and consistency such as increased training and calibration requirements have been investigated in the research, 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 a more attainable goal is to acknowledge that raters will differ in their severity but can be trained to be consistent in their own scoring (Eckes, 2011; Linacre, 1994).

MFRM is an approach 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. In MFRM analyses, rater behavior is captured through a "severity" parameter, which characterizes the rater in the same way that an ability parameter characterizes the teacher being observed, and a difficulty parameter characterizes an item of the rubric (Linacre, 1994). MFRM also reports on the amount of rater error. Interactions among raters and other facets of the observation, such as rater/teacher or rater/item interactions can also be investigated (Linacre, 1994). By examining rater severity,

error, and bias, MFRM analyses provide insights that can be used to improve rater training, leading to more consistent evaluations and feedback (Wigglesworth, 1993).

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). Examining these statistics at the item level allows us 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, so they can be revised or eliminated. If feedback provided through the use of observation rubrics is meant to drive changes in instructional practice, it is imperative that the rubrics contain the 'right' items. A teacher's performance also should not be dependent upon the rater observing the lesson. We have tested multiple RESET rubrics using MFRM analyses. To illustrate how these analyses inform the assumptions of the validity argument, we briefly describe the methods used to test the Explicit Instruction Rubric (Johnson et al., 2018) and summarize the results in relation to the assumptions included in Table 2.

MFRM Methods Summary. Thirty special education teachers across grade levels 2-8 from three states each provided three video recorded lessons (one from the beginning, the middle and the end of the school year) for a total of 90 videos. We included lessons as a facet because several studies show a difference in teacher performance depending on when during the school year the observation is conducted (Mantzicopoulos, French, Patrick, Watson & Ahn, 2018). Fifteen raters from seven states were recruited and trained by RESET project staff to use the Explicit Instruction rubric to observe the lessons, assign a score to each item of the rubric, provide time-stamped evidence of what they used as a basis for the score, and provide a brief

rationale for their score to be shared as feedback to teachers. Data were analyzed through MFRM analyses. The model used for the MFRM analysis in our studies is given by:

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

where P_{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 performance of teacher n, D_i is the difficulty of item i, C_j is the severity of the rater r, T_o is the stringency of the lesson l, and F_k is the difficulty overcome in being observed at the rating k relative to the rating k-1 (Eckes, 2011).

MFRM Results. The results of the analysis showed an internal consistency of items of .93, and exact rater agreement across a total of 10,010 assigned scores of 51%. 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). Figure 2 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 represents the logit scale, which is estimated from the pattern of the data. Placing the facets on a common scale allows for the comparisons within and among the four facets (Smith & Kulikowich, 2004).

The column heading for "Items" ranks the items from those least commonly performed well to those most commonly performed well, 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. As shown in Figure 2 and supported by the results of the MFRM analysis, 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 item

least well-performed, 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 most well-performed. Item fit statistics indicate whether raters have scored items in a consistent manner. The fit statistics for all of the items are within the acceptable range, which means that there were items that teachers tended to do well on, and other items that teachers tended to struggle with. This is an important finding given that feedback to teachers is provided at the item level. The identification of "more difficult" items can help raters focus their feedback to teachers. The MFRM results, along with the results of the G-study which indicated that 12.3% of the variance is with items, and 7.8% of the variance is explained by teacher x item interactions, provide evidence to support assumptions 1.1, 1.3, 1.4 (Table 2).

The third column of Figure 2 contains the teacher facet, with more proficient teachers having higher logit values. Teacher 5 is the most proficient teacher, and teachers 11 and 17 are the least proficient. The reliability of separation is .98, which indicates that teachers differ in their ability to implement explicit instruction as measured by this rubric, beyond what can be attributed to measurement error (Assumption 2.1 of Table 2). 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. The results of our MFRM analyses indicated that all fit statistics were within acceptable ranges, suggesting that the scoring criteria have been consistently applied to determine teachers' implementation of explicit instruction (Assumption 2.1).

The rater column of Figure 2 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 reported in Table 4 help to determine whether raters are consistent with their own ratings and can be used to identify 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 (overfit). Misfit is generally thought to be more problematic than overfit (Myford & Wolfe, 2003). The fit statistics for raters are within the acceptable range, providing evidence for Assumptions 1.4, 2.1, although our low levels of exact agreement indicate a need for additional approaches to investigate rater behavior. In addition to the MFRM analyses, we conducted think-aloud studies, described later in this paper.

MFRM analyses can account for differences in rater severity by adjusting the observed score and computing a "fair average" score for teachers (Eckes, 2011; Linacre, 2017). A "fair average" score is the score that a particular examinee would have obtained from a rater of average severity (Eckes, 2011). We compared the teachers' average observed score across all items, lessons and raters and their "fair average" score as computed by the FACETS program (Linacre, 2017). There were 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 were some minor differences in the rank ordering of teachers based on observed versus fair average scores, there were no changes in the identification of the top 10% or the bottom 10% of performers (Assumptions 2.1, 4.2).

As indicated, the results of our MFRM analyses provide evidence for several of the scoring assumptions of the IUA. In addition to the findings presented thus far, another important source of evidence provided by the MFRM analyses is found in the score distributions, which are summarized in Table 5 for the various RESET rubrics we have tested to date. Across rubrics, we find a distribution of scores suggesting that raters are using the items to differentiate across

various levels of performance (Assumption 1.1). Throughout our work with RESET we find that the score distribution tends to include lower percentages of scores of 'implemented' when teachers are evaluated with the content rubrics instead of with the more general instructional delivery rubrics. This finding is consistent with observational studies of special education instruction that report that teachers struggle to deliver instructional practices in the ways in which they were intended (Ciullo, Lembke, Carlisle, Thomas, Goodwin & Judd, 2016). This result is also consistent with research that identifies the multidimensional nature of instruction, including both general and content-specific practices, and suggests the need for observation systems that reflect this complex structure (Blazar, Braslow, Charalambos & Hill, 2017). Overall, the MFRM analyses provide evidence to support many of the scoring assumptions for RESET. However, the 51% exact agreement across raters is disconcerting, and warrants further examination.

Feedback Study

As described, the RESET observation system must facilitate the provision of accurate, reliable feedback about the specific instructional adjustments teachers need to make. While it is critical to develop an instrument with adequate psychometric properties, it is also important to investigate whether feedback provided as a result of observations scored with the rubric leads to improvements in teachers' ability to implement the relevant EBP (Assumption 4.1). Therefore, we conducted a feedback study, in which a total of 30 special education teachers participated.

Fifteen teachers were assigned to the treatment condition, using the RESET rubric to self-evaluate their instruction and receive item-level feedback from RESET project staff for each of six lessons, and 15 teachers were in the comparison group. Treatment group teachers received scores and feedback at the item level on each of six lessons they recorded over an eight-month

period. At the end of the study period, 15 external raters who were unaware of assignment to condition or to time of observation used the Explicit Instruction rubric to evaluate the videos. We selected three videos (one from the beginning, middle and end of the school year) from each teacher to evaluate. Data were analyzed through MFRM analyses as already described, and repeated measures MANOVA, in which teacher condition served as the between subject factor, and lesson number served as a measure of time, used in this analysis as the within subject factor. As noted previously, Lesson 1 was recorded in the beginning of the school year, Lesson 2 in the middle, and Lesson 3 at the end of the school year. For this study, we used component scores on the rubric, which consisted of the total sum of scores for the items that comprise the various components of the Explicit Instruction rubric. The overall results for the between factor, group, was not significant. However, there was a significant interaction effect for lesson number (time) * group, $F(1, 28) = 2.386$, $p = .034$. Follow up contrasts examining the interaction effect revealed that the treatment group made significantly greater gains on Component 1, Identifying and Communicating Goals (this component score is comprised of the first 3 items of the Explicit Instruction rubric), than the comparison group, $F(1, 28)$, $p = .049$.

These findings provide preliminary but promising evidence for Assumption 4.1. Through the process of observation and feedback provided with the Explicit Instruction rubric teachers were able to improve their ability to implement this EBP as measured by external raters. These initial findings are encouraging and suggest the need for continued studies that examine the impact of feedback on teachers' ability to implement EBPs. To conduct this study we relied on two RESET project staff to observe and provide feedback to participating teachers. We did this to limit any variability due to the various ways in which raters could interpret teachers' performance for our initial investigation. In practice however, the potential for using observation

protocols to improve instructional practice depends on the extent to which different raters make the same judgments given the same evidence (Gitomer et al., 2014). An observer must be able to consistently use a protocol to distinguish among different instructional elements and the levels of performance in their implementation, and they must have a common understanding to ensure that the scores assigned and feedback provided are not unduly influenced by the observer assigning them (Hill & Grossman, 2013).

Rater Behavior Studies

Therefore, to better understand the factors that influence raters' application of the scoring procedures and criteria, for several of the RESET rubrics, we are investigating: 1) the extent to which raters are able to consistently represent the scoring criteria in the rubrics 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. To date, we have completed a rater behavior study for the Explicit Instruction rubric (Johnson, Zheng, Moylan & Crawford, under review), with studies for our Reading for Meaning and Understanding Math Procedures in progress. To examine rater related issues, in addition to the MFRM analyses reported previously, we also asked the 15 rater participants to record a think aloud on a common video recorded lesson, in which they articulated how they were interpreting what they saw in the video with the scoring criteria, and how they reached their decision about final scores. RESET project staff also completed a master scored rubric for this video, that was used to obtain a measure of raters' consistency with expert rated observations.

We first analyzed the level of rater agreement with the master scored rubric of the lesson (Table 6 shows the first six items to provide an example). Only two rubric items (number 3 and

7) had assigned scores that spanned all three levels of performance. 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 appeared to be 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 Explicit Instruction rubric training manual, *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, Crawford, Moylan, Zheng, 2016, p. 10). The most challenging items to score consistently included the phrase, *throughout the lesson*. For example, Item 10, *The teacher uses language that is clear, precise, and accurate throughout the lesson,* had agreement levels of 54% with the master scored lesson.

In addition to this analysis of rater agreement across a commonly scored video, we also coded the raters' think aloud data 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? and 3) If the stated rationales are not consistent with the scoring criteria, 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 7). At stage two of the analysis we used these categories to analyze a random selection (20%) of scores and responses to determine the extent to which raters were consistent

with the guidance provided in the rubrics, training manuals and training. The analysis shows that 59% of raters' rationales for the given scores were fully consistent with the scoring criteria.

The results of this study also indicate that even when the rater is internally consistent in scoring (as indicated by MFRM fit statistics), 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 rubric includes items that are quite specific, but the variability with which raters interpreted them 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. It appears that the items that were most problematic were 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).

We are also conducting these analyses for our content area rubrics with the expectation that the resulting analyses will inform Assumptions 1.2, 1.3, and 4.1. These analyses also inform areas of need for rater training to ensure that as RESET is implemented, teachers receive consistent and accurate feedback that allows them to improve their ability to implement EBPs. If

a rater's observation of teacher instruction is used to guide instructional improvement with the end goal of improving student outcomes, then the data informing these decisions must be robust (Mantzicopoulos, French, Patrick, Watson & Ahn, 2018).

Planned Studies: Examining Student Outcomes in Relation to Teacher Performance

To collect evidence to evaluate the extrapolation assumptions 3.1 and 3.2, we are in the process of testing the premise that if special education teachers effectively implement an EBP, they will realize gains in student performance that are consistent with those reported in the research. In our initial conceptualization of RESET, we argued that state assessment systems were not sensitive enough to capture the growth of students with disabilities, a sentiment that has recently been echoed by other researchers (Fuchs et al., 2018). Therefore, the measurement of student outcomes within a special education teacher system must be based on standardized measures that are more proximal to the outcome of interest and that are flexible enough to capture the diverse needs of the heterogeneous special education population (Johnson & Semmelroth, 2014). We plan to do this by comparing the rates of student growth on standardized academic assessments typically used within special education with teacher performance on RESET, with the student measures converted to a common scale (e.g. z scores). We currently have a study underway in which 23 teachers are providing video recorded lessons that will be scored by external raters. Teachers are also providing student level data collected across three time points (beginning, middle, end of school year) for between 3-5 students. Data will be analyzed through growth curve, correlational and descriptive analyses so that we can examine the relationship between teacher performance and student outcomes.

Once data are analyzed, moderate correlations would suggest that RESET captures an important source of influence on student performance, whereas lower correlations could suggest

that other factors are at play. For example, in our work, we have observed that school schedules are frequently disrupted for a variety of reasons. Students with disabilities frequently have their instruction cancelled when special education teachers attend trainings or meetings that are part of the legal requirements of the special education system. Additionally, instruction is often not provided with the same frequency, duration or intensity as described in the research base, and it is important to understand the influence of these factors on student growth. In recognition of these variables, we have recently added an 'opportunity to teach' component to the RESET system that asks special education teachers to monitor and document these variables so we can better understand their influence on student achievement and make decisions that are aligned with addressing the underlying causes of low student growth.

Implications for Research and Practice

The complexities of developing teacher observation systems that realize their promise as effective drivers of education reform have been well-documented (Blazar et al., 2017; Bell et al., 2012; Hall, 2014; Hill & Grossman, 2013; Johnson, Ford, Crawford & Moylan, 2016; Shepard, 2012). In general, these include: 1) the need to align observation tools with the desirable practices at a level of specificity that allows teachers to receive actionable, specific feedback, 2) the challenge of training raters to develop shared understanding and consistent application of the scoring criteria and procedures, and 3) the time and resources needed to effectively observe and provide feedback to teachers. We have grappled with these challenges during the development and validation of the RESET observation system.

The primary goal of the RESET observation rubrics is to detail practices at a level of specificity that, when used to provide an evaluation of the teachers' ability to implement a specific EBP, gives teachers a clearer and more consistent target for instruction. As Hill and

Grossman (2013) note, "the absence of [specific] practices from most observation instruments limits the snapshot of teaching that emerges, the nature of feedback teachers receive, and the diagnostic information districts can glean about subject-specific needs for professional development" (p. 375). In our efforts to design the RESET system, we were heavily influenced by Hill and Grossman's (2013) call for observation instruments that provide a level of specificity to teachers about their instructional implementation that is actionable, and that overtime, will lead to significant, positive changes in teachers' ability to implement evidence-based instructional practices, and ultimately, gains in student performance. To develop RESET in a way that was responsive to this call, we synthesized the research to detail the specific elements of a variety of EBPs to assess teachers' implementation of these practices, to provide them with specific and actionable feedback on how to improve, and to hold teachers accountable for making those improvements.

To support this use of RESET, several assumptions must be met, and we have collected evidence to investigate several of these assumptions. Our generalizability studies and MFRM analyses conducted to date provide evidence to support several of the assumptions outlined in Table 2, in particular Assumptions 1.1, 1.3, 1.4, and 2.1. Our feedback study, while limited in terms of sample size and scope, is encouraging. However, more data are needed to fully investigate the decision assumption (4.1) and determine whether feedback at the item level will provide teachers with an accurate evaluation of their strengths and challenges in implementing EBPs. Closely related to this issue is our finding of low exact rater agreement. Across all of the studies of RESET rubrics, exact rater agreement ranges from 50-52%, which, while consistent with other reports of instructional observation studies (Bell et al., 2015; Casabianca Lockwood & McCaffrey, 2015; Cash, Hamre Pianta, & Myers, 2012; Jones, 2019), is problematic. Our

examination of rater behavior has provided some insight into how we can tailor training efforts to identify and then support raters to become more accurate and consistent in their application of the observation rubrics. Finally, the ultimate goal of this process is to improve outcomes for students with high incidence disabilities. Our studies to examine the relationship of a teacher's performance on RESET to student growth will provide evidence of the extrapolation assumptions (3.1, 3.2, 3.3).

Conclusion

Teacher observation systems are being used to make high-stakes decisions, yet few systems have been examined to determine their psychometric defensibility (Herlihy et al., 2014). Implementing teacher observation systems without sufficient data can have long-lasting negative consequences. In the short term, ill-informed decisions can misdirect efforts to improve instruction for students, and in the longer term, a disconnect between what was promised and what is delivered can completely undermine the support for and confidence in the system (Hall, 2014). To defend the use of system-based results, comprehensive validity efforts based on the collection of evidence aligned to system-based claims must be defined and conducted for observation systems prior to implementation, once they are in place, and for the subsequent years of use (Hall, 2014).

This paper described RESET, a special education teacher observation system, explained its intended use as a tool to improve special education teachers' ability to improve their implementation of EBPs, described the assumptions that would need to be met to support this use, and examined the evidence collected to date to determine the extent to which claims made from its use are warranted. The results are promising, but significantly more work is needed to develop a system that is both useful and fair. If we are to be successful in improving the practice

of special education teachers, we will need to further investigate how to improve rater accuracy and consistency, examine the impact of feedback provided with RESET on teachers' practice, and analyze the relationship of teacher performance with student outcomes.

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.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ark, T. K. (2015). Ordinal generalizability theory using an underlying latent variable framework (Doctoral dissertation, University of British Columbia). Retrieved from https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0166304

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In Williamson, Bejar, & Mislevy (Eds). *Automated scoring of complex tasks in computer-based testing*, 49-81.

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

Bell, C A., Yi Q, Croft, A J., Leusner, D., McCaffrey, D. F., Gitomer, D. H. & Pianta, R. C. (2014). Improving Observational Score Quality: Challenges in Observer Thinking. In Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta (Eds) *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, 50-97. San Francisco: Jossey-Bass.

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,doi:10.1080/10627197.2017.1309274.

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 of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542.

Ciullo, S., Lembke, E. S., Carlisle, A., Thomas, C. N., 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.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49* 560-575. Doi: 10.1111/medu.12678

Crawford, A. R., Johnson, E. S, Moylan, L. A., & Zheng, Y. (2018). Variance and reliability in a special educator evaluation instrument. *Assessment for Effective Intervention*. doi: 10.1177/1534508418781010

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.

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

Erlich, O., & Shavelson, R. (1978). The search for correlations between measures of teacher

behavior and student achievement: Measurement problem, conceptualization problem, or both? *Journal of Educational Measurement, 15*, 77–89.

Fuchs, D., Hendricks, E., Walsh, M. E., Fuchs, L. S., Gilbert, J. K., Zhang T. W., Patton, S., Davis-Perkins, N., Kim, W., Elleman, A. M. and Peng, P. (2018). Evaluating a multidimensional reading comprehension program and reconsidering the lowly reputation of tests of near-transfer. *Learning Disabilities Research & Practice, 33(1)* 11–23.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*, 5–29.

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.

Goe, L., Bell, C. A., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/

Hall, E. (2014). *A framework to support the validation of educator evaluation systems*. National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org

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.

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.

Johnson, E. S., Crawford, A., Moylan, L. A., & Ford, J. W. (2016). Issues in evaluating special education teachers: Challenges and current perspectives. *Texas Education Review 4*(1), 71-83.

Johnson, E. S., Crawford, A., Moylan, L. A., Zheng, Y. (2016). *Explicit Instruction Rubric Technical Manual*. Boise State University, Boise, ID.

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*. *37*(2), 35-44.

Johnson, E. S., Moylan, L. A., Crawford, A. R., & Zheng, Y. Z. (in press). Developing a comprehension instruction observation rubric for special education teachers. *Reading and Writing Quarterly*.

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2018). Developing an explicit instruction special education teacher observation instrument. *Journal of Special Education*. doi: 10.1177/0022466918796224.

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

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

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

64). Westport, CT: Praeger.

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

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Linacre, J. M., (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (2014). *Facets 3.71. 4* [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.

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

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

Norris, C. E., & Borst, J. D. (2007). An examination of the reliabilities of two choral festival adjudication forms. *Journal of Research in Music Education, 55*, 237–251.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

Shepard, L. (2012). *Evaluating the use of tests to measure teacher effectiveness: validity as a*

theory of action framework. A paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, British Columbia.

Smith, 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.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of statistics*, *26*, 81-124.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305-319.

Wong, C., Odom, S. L., Hume, K. A., Cox, C. W., Fettig, A., Kurcharczyk, S., . . . Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders, 45*, 1951–1966.

Table 1

Organization and Structure of RESET

Subscale	Content Area	Rubrics
Instructional Methods	N/A	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 of: Number Sense & Place Value, Operations, Fractions, Algebra
		Procedural Understanding of: Number Sense & Place Value, Operations, Fractions, Algebra
		Automaticity
	Writing	Spelling
		Sentence Construction
		Self Regulated Strategy Development
		Conventions
Individualization		Executive Function/Self-Regulation
		Cognitive Processing Accommodations
		Assistive Technology

Duration/Frequency/Intensity

Table 2

Interpretive Use Argument and Assumptions for RESET

1. Scoring assumptions

    1.1 The scoring rule is appropriate.

    1.2 Raters' understanding of the items are accurate and consistent with the developers' understanding.

    1.3 Raters can consistently apply the scoring criteria.

    1.4 Raters use the items without bias in that the same instructional behaviors and quality observed across different teachers would be scored similarly.

2. Generalizability assumption

    2.1 Overall teacher scores are generalizable over items, raters, and lessons.

3. Extrapolation assumptions

    3.1 RESET consists of a set of distinct rubrics that detail the elements of evidence-based practices for students with high incidence disabilities. Performance across a set of items on an individual rubric represents the teacher's ability to implement the specific practice detailed in the rubric (trait interpretation).

    3.2 Higher scores on a RESET rubric is positively related to student gains in the specific academic area (e.g. performance on decoding instruction is related to a student's reading growth)

    3.3 Items accurately represent the evidence-based practices.

4. Decision assumptions

    4.1 Feedback to teachers based on item level scores appropriately reflects key teacher strengths and weaknesses.

Table 3

Analysis of Variance for the Explicit Instruction Rubric

Source of Variation	SS	df	MS	% of Total Variance	SE
Teachers (T)	213.99	9	23.77	7.5	0.026
Lessons:Teachers (L:T)	88.12	30	2.94	3.3	0.007
Raters (R)	91.29	3	30.43	4.5	0.019
Items (I)	350.76	24	14.61	12.3	0.025
TR	124.67	27	4.62	6.2	0.012
TI	249.90	216	1.16	7.8	0.007
LR:T	93.60	90	1.04	7.2	0.006
LI:T	221.44	720	.31	4.9	0.004
RI	142.36	72	1.98	7.2	0.008
TRI	210.25	648	.32	5.6	0.005
LRI:T	419.21	2160	.19	33.6	0.006
Total	2205.59	3999			

Note. SS = sums of squares; MS = mean squares; EBP = evidence-based practice.

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

Score Distributions Across RESET Observation Rubrics

Rubric Name	Implemented	Partially Implemented	Not Implemented
Explicit Instruction	40%	51%	9%
Reading for Meaning	28%	31%	41%
Comprehensive Decoding	32%	41%	26%
Understanding Procedures	19%	57%	24%

Table 6.

Analysis of rater scores and rationales across a common lesson for a sample of items from the Explicit Instruction RESET rubric.

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)

Note. Bolded responses are consistent with the master scores.

Table 7

Consistency Summaries of Rater Evidence with the RESET Rubric Training Manual

Coded Categories to Explain Rater Consistency	N*	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

Note. N=the total count of coded observations that fit within this category of rater consistency