Using Evidence-Centered Design to Create a Special Educator Observation System

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

Boise State University

January 2018

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. (2018) Using Evidence-Centered Design to Create a Special Educator Observation System. *Educational Measurement: Issues and Practice,* **https://doi.org/10.1111/emip.12182**

Abstract

The Evidence-Centered Design (ECD) framework was used to create a special education teacher observation system, Recognizing Effective Special Education Teachers (RESET). Extensive reviews of research informed the domain analysis and modeling stages, and led to the conceptual framework in which effective special education teaching is operationalized as the ability to effectively implement evidence-based practices for students with disabilities. In the assessment implementation stage, four raters evaluated 40 videos and provided evidence to support the scores assigned to teacher performances. An inductive approach was used to analyze the data and to create empirically derived, item level performance descriptors. In the assessment delivery stage, four different raters evaluated the same videos using the fully developed rubric. Many-facet Rasch measurement (MFRM) analyses showed that the item, teacher, lesson and rater facets achieved high psychometric quality. This process can be applied to other content areas to develop teacher observation systems that provide accurate evaluations and feedback to improve instructional practice.

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

Using Evidence-Centered Design to Create a Special Educator Observation System

Teacher observation systems are increasingly seen as an important component of education reform because they offer the opportunity to evaluate teaching practice and to 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). However, in the effort to adopt observation systems on a broad scale, many states and districts are using evaluation tools that are very generic in nature, or that have been designed primarily for accountability and therefore do not provide teachers with extensive feedback on practice (Hill & Grossman, 2013). If teacher observation systems are to fulfill their promise of improving instruction, considerable work remains to ensure that they are developed and implemented in ways that address the shortcomings of current tools.

To be useful, a teacher observation system must facilitate accountability, support growth and development of professional practice, and provide accurate, reliable ratings and feedback about the specific instructional adjustments teachers need to make (Hill & Grossman, 2013). Many observation systems however, are poorly aligned with the evidence-based instructional practices (EBPs) within the relevant content area, limiting the quality of the feedback provided to teachers through this mechanism (Grossman, Compton, Igra, Ronfeldt, Shahan & Williamson, 2009). This is especially the case for special education teachers, who are routinely evaluated with observation instruments designed for the general education setting (Johnson & Semmelroth, 2014). Additionally, large scale studies of current observation systems have indicated a propensity for bias in scores, in which the majority of teachers are discovered to be proficient or

better (Kane & Staiger, 2012). Recent state level reports confirm that in practice, the tendency for bias in teacher observation systems is significant (Farley, 2017).

Effective teacher observation systems require deliberate construction and thorough psychometric evaluation. An assessment that seeks to measure something as complex as instructional practice must be designed around the inferences that are to be made, the observations that will be used to draw these inferences, and the chain of reasoning that connects them (Messick, 1994). Evidence-Centered Design (ECD) provides a conceptual design framework to create complex, coherent assessments based on the principles of evidentiary reasoning (Mislevy, Steinberg & Almond, 2003). In brief, ECD consists of five stages: 1) domain analysis, 2) domain modeling, 3) conceptual framework, 4) assessment implementation, and 5) assessment delivery. Designing assessment products through the ECD framework ensures that the way that evidence is gathered and interpreted is consistent with the underlying construct the assessment is intended to address (Mislevy, et al, 2003).

ECD has been applied to several, significant large-scale *student* assessment systems (Plake, Huff & Reshetar, 2010), but has not been used extensively to develop teacher observation instruments. In this manuscript, we describe the development of a special education teacher (SET) observation instrument that has been developed through the ECD framework with the goal of providing SETs clear and actionable signals about ways to improve their teaching practice, minimizing bias in the resulting evaluations, and providing reliable results across raters.

Recognizing Effective Special Education Teachers (RESET)

RESET is a federally funded project to create observation rubrics aligned with EBPs for students with high incidence disabilities. The goal is to leverage the extensive research on EBPs for this population of students to inform the development of observation instruments that provide

feedback to SETs to improve their practice and ultimately, to improve outcomes for students with disabilities (SWD). To create the RESET observation system, we followed the five-stage ECD framework (Mislevy et al., 2003). Below, we describe each stage as it applies to the development of RESET, followed by a reporting of the studies undertaken to inform the assessment implementation and delivery stages.

Domain Analysis

The domain analysis stage involves collecting substantive information about the domain being assessed (Mislevy & Haertel, 2006); in this case, effective special education teaching. We reviewed the research on teacher impact to determine the salient aspects of the teacher's role in affecting student outcomes to create a definition of special education teaching. Drawing on the research on instructional practice, we identified common elements of effective instruction such as: 1) maintaining rigorous expectations; 2) creating an effective, engaging learner environment; 3) making content area knowledge relevant, and 4) providing learning experiences using effective research-based strategies (Hattie, 2009). Next, we engaged in a meta-review of the research on effective special education instructional practice, organizing our search through these four elements. Several meta-analyses of EBPs provided useful starting points for conducting our review (see for example: Bellini, Peters, Benner & Hopf; 2007; Berkeley, Scruggs & Mastropieri, 2009; Gersten et al., 2009; Swanson & Sachse-Lee, 2000). The result of this review led to a definition of effective special education teaching as the ability to assess a student's learning needs and implement EBPs to support academic and social/emotional growth.

Domain Modeling

We then moved to the domain modeling stage, in which the information and relationships identified in domain analysis are translated into assessment design options. Based on our

definition of effective special education teaching, we concluded it is best assessed through observations of a SETs instruction that are evaluated using rubrics detailing the essential elements of the EBPs we expect to see in the classroom. To create assessment design options within the domain modeling stage, both characteristic and variable features are used to specify how SETs will produce performance tasks (Mislevy & Haertel, 2006). The *characteristic* tasks common across SETs include video recording the SET directly working with students in an instructional setting. However, because teaching contexts and instructional settings are highly variable in special education, the *variable* features of RESET include establishing criteria for evaluating a range of EBPs depending on the specific context in which the SET is working. SETS are responsible for providing instruction across content areas, grade levels, and various arrangements such as pull-out models or co-teaching. SETs also work with students who require specially designed instruction that is individualized depending on student need. SETs must be well-versed in numerous EBPs and be cognizant of various disability types to plan and implement effective instruction (Odom, Brantlinger, Gersten, Horner, Thompson & Harris, 2005). Therefore, an effective SET observation system must capture a broad range of EBPs, delivered in a variety of contexts and adapted to meet individual student needs.

Conceptual Assessment Framework

With the framework developed in the domain modeling stage, we moved to create the blueprint for RESET – or the conceptual assessment framework, which is divided into models that bridge the assessment argument with the operational activities of the assessment system (Mislevy et al., 2003). The models included within RESET include the 1) teacher model; 2) evidence model, 3) task model, and 4) presentation model (Mislevy et al., 2003). The *teacher* model in RESET consists of a single variable, a SETs proficiency in the implementation of

EBPs. Through our review of literature undertaken in the domain analysis stage, we organized EBPs into three main areas: 1) instructional methods, 2) content organization and delivery, and 3) individualization. Within each category, we outlined the rubrics associated with the EBPs to create an overall blueprint for RESET. The list of rubrics organized by category is included in Table 1. Through RESET, we obtain evidence that provides an estimate of a SETs proficiency to effectively deliver instruction, to organize and support content area learning, and to individualize instruction based on the students' presenting needs.

SETs submit video recordings of their lessons which are then evaluated using the appropriate rubric from each subscale. This process comprises the *evidence* model. The scoring rules are based on the SETs level of implementation of EBPs, and evaluated as implemented, partially implemented, or not implemented. The *task* model for RESET is any lesson delivered by the SET to SWD. The *presentation* model for RESET relies on the use of video recorded lessons and electronic versions of the relevant RESET rubrics. Observations are self-evaluated by the SET and evaluated by raters who have been trained in the use of RESET.

Assessment Implementation

The operational model derived from the conceptual assessment framework leads to the assessment implementation stage (Mislevy et al., 2003), the stage at which assessment items are created. As described above, RESET consists of a set of rubrics, each rubric reflects the items and performance-level descriptors for a specific EBP. To create individual items for each rubric, we conducted extensive reviews of the research on the EBPs included within RESET, then synthesized the descriptions of these practices across studies to create a set of items that detailed each EBP. To illustrate the item development process in more detail, we will use the Explicit Instruction rubric as an example (see Appendix A).

A number of studies and meta-analyses have identified explicit instruction as one of the most effective approaches to teaching students with disabilities (see for example, Archer & Hughes, 2010; Brophy & Good, 1986; Christenson, Ysseldyke & Thurlow, 1989; Gersten, Schiller & Vaughn, 2000; Rosenshine & Stevens, 1986; Swanson, 1999). We first extracted the critical elements of explicit instruction from the literature, then reviewed and synthesized them into a coherent set of elements. Then, drawing on this review, we drafted a set of items to describe proficient implementation of explicit instruction. We refined the descriptors for proficient implementation by reviewing video recorded lessons collected from SETs, and discussing the clarity and utility of each item as written. We sent the rubric to subject matter experts for review, synthesized their feedback and completed revisions to create a set of elements that described proficient implementation of explicit instruction.

Because the purpose of RESET is to both evaluate and provide feedback to SETs, we needed to create a set of scoring rules that define and describe varying levels of implementation (e.g. implemented, partially implemented, not implemented). Initially, we considered using general descriptor levels, however, rating scales can be imprecise when general descriptors are used (Hill & Grossman, 2013). Additionally, a key focus of ECD is to identify observable evidence to create performance-level descriptors (PLDs) that result in a transparent evidentiary argument and consistent evaluations of performance (Ewing, Packman, Hamen & Thurber, 2010). PLDs communicate what various levels of performance should look like, and serve a critical role in setting cut scores that ultimately determine the categorization of a person's performance (Ewing et al., 2010). Ewing et al (2010) describe an iterative process for articulating PLDs in which performances are mapped onto the performance continuum, with items that best target the meaning of a specific performance category as well as clearly

differentiating the adjacent performance levels. An analysis is then undertaken to provide a synthesis of the salient content and skills that characterize and differentiate the categories along the performance continuum, and this analysis will reveal where more evidence may be needed to inform the PLDs. In this initial work to develop RESET, we began the process of PLD development through a study designed to create analytically developed descriptors (Knoch, 2009), with the intent in future studies to engage in the iterative process described by Ewing et al (2010) to further refine the rubrics.

Assessment Delivery

In the assessment delivery stage items are piloted, feedback is collected, and psychometric analyses are conducted, the results of which are integrated into the final design of the assessment tool. Our primary objective was to create an observation instrument that provides reliable results across raters that provides SETs with clear and actionable signals about legitimate ways to improve their teaching practice. Because there are multiple variables that can impact a SETs score within RESET, we employed many-facet Rasch measurement (MFRM) analysis to conduct a substantive investigation of the teacher, lesson, rater and item facets, as well as the teacher and item difficulty. MFRM is an extension of the Rasch model that conceptualizes the expected performance of individuals as a function of their ability and the item difficulty (Smith & Kulikowich, 2004). MFRM allows us to include additional assessment variables such as rater severity into the analysis. MFRM also allows us to identify particular elements within a facet that are problematic and to conduct a bias analysis that identifies specific combinations of facet elements – particular rater-teacher combinations, for example - that are consistently different from the overall identified pattern (Eckes, 2011).

Teacher observation systems are high stakes assessments. They are used both to inform the instruction that students receive as well as to make critical decisions about teachers. To meet these demands, observation systems require a deliberate approach to development and a rigorous evaluation of their psychometric properties. The ECD framework provides a useful heuristic for creating observation systems suited for these purposes. In this review, we have described the application of the first three stages of the ECD process for creating RESET, an observation system specifically designed for SETs. Using one of the rubrics within the RESET system, the Explicit Instruction (EI) rubric, we now detail two studies undertaken that informed the assessment implementation and assessment delivery stages of the ECD process.

Methods

In this section, we describe two studies. The first study describes the processes undertaken to create an initial set of PLDs and the second study details the procedures used to analyze the reliability of the EI rubric.

Study 1. Performance-Level Descriptor Study.

Participants

Special education teachers. A total of ten special education teachers from three states provided four video recorded lessons each for a total of 40 videos. All teachers were female, with an average experience level of 11.55 years (8.46 SD). Nine teachers taught at elementary and one at the middle school level. All teachers had their special education certification, five had undergraduate degrees and five had graduate degrees.

Raters. A total of four raters participated in the descriptor development study. Two of the raters were instructional coaches, and two were veteran special education teachers who served as

department chair and lead teacher within their schools. Raters had an average of 15 years of experience. All four raters were female.

Procedures

Video collection. During the 2015-16 school year, SETs provided weekly video recorded lessons from a consistent instructional period. Videos were recorded and uploaded using the Swivl® capture system and ranged in length from 20-35 minutes. Each teacher contributed a total of 20 videos over the school year. From this video bank, four videos from each teacher were selected for inclusion in the study. To be included in the data set, videos had to have adequate video and audio quality (of the 800 total videos, 42 were found to be not usable due to poor video quality or lack of sound), and had to depict an instructional lesson for which the use of the EI rubric was applicable. If a teacher had more than four videos that met these criteria, we randomly selected four. Videos were assigned an ID number and listed in unique, random order for each rater to control for order effects.

Rater training. Rater training took place over two days. Raters were provided with an overview of the RESET project goals, and a description of how the EI rubric was developed. Project staff then explained each item of the EI rubric and clarified any questions the raters had about the items. Then, raters watched a video that had been scored by project staff, scored the video with the EI rubric, and then the scores were reviewed and discussed to include the rationale for the score that each item received. Raters then watched and scored two videos independently, and scores were reconciled with a master coded rubric for each video. Any disagreements in scores were reviewed and discussed. Raters were then assigned a randomly ordered list of videos. Raters were asked to score each item, to provide time stamped evidence

that they used as a basis for the score, and to provide a brief explanation of the rationale for their score. Raters were given a timeframe of four weeks to complete their ratings.

Data Analysis

Performance-Level Descriptor (PLD) development. To create the PLDs for each item, we compiled the evidence and explanations provided by the raters after they scored the videos. We used a general inductive approach (Thomas, 2006) to condense their input into themes and categories that emerged as key terms identified as influencing scoring decisions. The coding process included several phases: initial reading, identifying segments of information, labeling segments of information, creating categories, selecting categories, and creating themes. First, the evidence and explanations were reviewed until the researchers were familiar with their content and gained an understanding of the text. Then, text segments that contained meaningful units were identified. The identified segments were labeled as codes by using words, phrases, or sentences directly used in the segments to capture their key elements as closely as possible. Codes which had the same or similar key elements were grouped together to generate categories. Then, categories were selected to develop descriptors relevant to the rating scale of 1) not implemented, 2) partially implemented and 3) implemented, or N/A) not applicable.

Several strategies were used to address the trustworthiness of the item level descriptors including consistency checking, peer debriefing, and stakeholder checking. Consistency checking involved independent parallel coding by two researchers (Thomas, 2006). Two researchers analyzed the raters' evidence and explanations, then compared their analysis until they reached consensus in codes, categories, and descriptors. Peer debriefings were conducted with the research team (Creswell & Miller, 2000). The RESET team reviewed the codes and categories while referring to the evidence and explanations of raters, and participated in

consensus building of descriptors. Stakeholder checking was conducted by requesting teachers and raters to review the descriptors. The researchers also kept procedural and analytic memos about the meaning of the data (Esterberg, 2002). The end result of this extensive process was a full set of descriptors for each item, a revision of the item descriptors for 'implemented' and paring down the number of items from 27 to 25. The final rubric is in Appendix A.

Study 2. Many-facet Rasch Measurement Analysis.

Participants.

Special education teachers. The same teacher participants from Study 1 participated in Study 2.

Raters. A total of four raters participated in the MFRM reliability analysis study. One rater was a post-doctoral researcher, one a school-psychologist and RTI coordinator in her school, one a special education faculty member, and the fourth a special education teacher completing graduate studies in special education. Raters had an average of 17 years experience. Three raters were female and one was male.

Procedures

Video collection. The same video set from Study 1 was used during Study 2.

Rater training. Rater training was conducted as described in Study 1, with the exception that raters in this study were trained using the fully developed EI rubric with performance-level descriptors for each item.

MFRM Analyses. We analyzed the data collected by the raters using the fully developed EI rubric through MFRM analyses. The raw scores assigned to the EI rubric are ordinal, making valid comparisons between teachers or items difficult, as equal raw score differences between pairs of points do not imply equal amounts of the construct under investigation (Smith &

Kulikowich, 2004). With Rasch models, the ability estimates of teachers are freed from the distributional properties of the items, and the particular raters used to rate the performance (Eckes, 2011). Additionally, the estimated difficulty of items and severity of raters are freed from the distributional properties of the other facets of the assessment (Smith & Kulikowich, 2004). The model used for this analysis 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 extra 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 (Eckes, 2011; Linacre, 2014b).

Results

Data collected from the raters who used the fully developed EI rubric was analyzed with the FACETS (Linacre, 2014a) program. The results of the analysis are shown in Figure 1 and Tables 2 through 6. Figure 1 includes the variable map and rank order of each facet. Tables 2-5 report the fit statistics and reliability and separation indices for each of the facets. Bias analysis results are reported in Table 6. All analyses are based on a total of 3952 observations. Category statistics showed that of the 3952 assigned scores, 51% were a 3 (implemented), 33% were a 2 (partially implemented) and 16% were a 1 (not implemented). Only 1% of items received an N/A.

The far left column of Figure 1, titled "Measr," is the logit measure for the elements within each facet of the design. The second column contains the item measures, with more difficult items having larger logit values. Items 3, 13 and 12 were the most difficult, and items 21, 5, and 19 the least. Examining the items on the EI rubric (see Appendix A), the rank order of items is logical. For example, items 12 and 13 require teachers to task analyze and to deliver instruction in ways that support the individual needs of their students. This is a difficult skill that likely develops over time and with training. Items that were the least difficult included #5, alignment of instruction to the state goal, which, if the teacher is using an evidence-based program to guide her instruction, will meet this criterion. Additionally, item 19 focuses on providing students with opportunities to respond. Low teacher:student ratios may make implementing this item significantly easier than it might be in larger classrooms.

The third column contains the teacher facet, with more proficient teachers having higher logit values. Teacher 9 is the most proficient teacher (proficiency = 1.64 logits, *SE* = .10), and teacher 10 is the least proficient (proficiency = -.17 logits, *SE* = .08). The fourth column contains the lesson facet. In our data collection design the rank ordering of the lesson facet is somewhat difficult to interpret, because we did not specify the content or focus of the lessons but instead had the teachers select which lessons to submit. The fifth column contains the rater facet, with more severe raters having higher logit values. Rater 4 was our most severe rater (severity = .49 logits, *SE* = .05), and Rater 1 our most lenient (severity = -.64 logits, *SE* = .06).

Tables 2 through 5 report the fit statistics and the reliability and separation indices for the item, teacher, rater, and lesson facets. For all facets, all fit statistics fell within .8 to 1.2, which are within acceptable levels (Eckes, 2011). In addition to the fit statistics, reliability and separation information indices are reported. For items, the reliability coefficient was .97, separation = 5.62; for teachers, the reliability coefficient was .98, separation = 7.39. These statistics demonstrate reliable differences in item difficulty and teacher proficiency. For lessons, the reliability coefficient was .93, separation = 3.72, showing a discrimination across lessons, but lessons 1 and 2 have almost the same logits, providing some indication that we may be able to obtain reliable ratings with just three lessons instead of four. The reliability coefficient for raters was .98, separation = 9.07, suggesting differences in rater severity. The bias analysis (Table 6) indicated that a total of 31.5% of the variance in the observations (n = 3952) was explained by the model. 5.54% was explained by teacher/rater interactions, with 3.55% explained by teacher/lesson interactions, leaving 59.42% of the variance remaining in residuals. Table 6 presents only the rater/teacher pairs that showed bias and reports observed and expected scores, bias size in logits, t value and its probability. Of 40 possible teacher/rater interactions, 23 are

biased. Teacher 3 was the only teacher with no biased interactions. Rater 4 had the fewest number of interactions. There was almost an even number of negative bias (n = 11) as positive (n = 12) interactions, with no clear pattern attributable to a specific teacher, rater or teacher/rater pair. As a whole, despite the presence of biased pairs, the EI rubric does not appear to exhibit a great deal of bias and the overall MFRM results suggest the facets function effectively.

Discussion

ECD is a framework that can guide efforts to create assessment systems that measure the complex construct of teaching, the inferences to be made about a teacher's ability to implement instruction, the observations that will be used to draw these inferences, and the chain of reasoning that connects them (Messick, 1994). In this manuscript, we described how the ECD framework was applied to create RESET, a special education teacher observation system (Johnson, Crawford, Moylan & Zheng, 2016). The process described can be applied to other content areas to develop observation instruments of the caliber needed to realize the goal of improving practice.

We used a rigorous process in the assessment implementation stage that included having expert raters provide the evidence and rationale they used to assign scores. Then we created detailed performance level descriptors for each item. In the assessment delivery stage, we tested these descriptors with another set of raters to evaluate how well the EI rubric functioned. Through MFRM analyses, we were able to assess the reliability of the rubric and review how the various facets of the observation tool function.

Overall, our analyses provide strong evidence that we have created a rubric that will provide consistent evaluations of a SETs ability to implement Explicit Instruction. The psychometric reliability of items and teacher ability measures is supported by high reliability and

separation statistics. That is, the RESET EI rubric reliably divided the items and teachers into statistically different strata, indicating the sensitivity of the instrument (Wright & Stone, 1999).

Although the results of the studies reported in this manuscript are promising for the continued development of the RESET observation rubrics, there are several limitations that warrant caution in interpreting the results. The most significant limitation is that the sample sizes of both special education teachers (n = 10) and raters (n = 8 total) are small, and also limited in their representativeness of the larger population of special education teachers and potential raters. The benefit of using video observations however, is that over time, we can develop a video bank that will include a larger pool of teachers. Continued studies with larger samples of teachers and raters will be needed to verify the results of the studies reported in this manuscript.

A second limitation in the study reported here includes the process used to develop PLDs. Although we collected a significant amount of evidence from raters during our first study to inform descriptor development, within the process of ECD, the identification of claims and evidence to create PLDs should be iterative, with the goal of creating a transparent evidentiary argument (Huff, Steinberg & Matts, 2010). Future studies that continue this cycle of generating evidence and applying the mapping process to ensure that score interpretations are well-matched with the evidence and resulting PLDs are needed to further refine the RESET observation rubrics (Ewing et al., 2010; Plake, Huff & Reshetar, 2010).

Finally, scores provided on observation systems are a function not only of the teachers' ability but also of the severity of the rater evaluating them. Our analyses indicate that raters differed in their severity, but that the fit statistics for raters were within acceptable levels, suggesting no evidence of halo effects or noisy scoring. 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 a fair score for teachers. This is different than other approaches to examining rater behavior that expect raters to function as scoring machines, achieving perfect agreement against a master set of scores (Eckes, 2011). Research on rater behavior however, suggests that achieving perfect agreement across human raters who judge complex performances is an elusive goal and that acknowledging that raters will differ in their severity but can be trained to be consistent in their own scoring may be a more attainable reality (Eckes, 2011). The training provided to our raters appears to have achieved this goal, but further studies examining whether these findings will hold when raters who will likely serve as evaluators but who have less experience in special education (e.g. principals) are needed.

Despite these limitations, the results of our current analyses are promising. To fully realize the benefit of the RESET observation system, continued research on a variety of assessment aspects is needed. For example, the processes described in this manuscript must be applied to the other rubrics within the RESET system. Given the focus of RESET on improving teacher performance, we will also need to examine the impact of feedback and self-evaluation. Finally, the connection of teacher performance on RESET will need to be connected to student measures. A significant amount of research is needed to fully inform the development of teacher observation systems, but the ECD process is a useful blueprint for this undertaking.

Conclusion

Teacher observation systems are high stakes assessments. They are expected to significantly impact teacher behavior in ways that will lead to improved instruction and greater student gains. To achieve this vision, teachers must be held accountable through evaluation systems expressly designed for this purpose.

The development of RESET has been guided by the ECD framework to respond to the need for better teacher observation tools. Through adherence to the five stage process, we have adequately modeled the domain of effective special education teaching, created a conceptual assessment framework based on the research, and devised assessment items that reflect EBPs, result in reliable evaluations of teacher implementation and are at a grain size sufficient to provide actionable feedback. Next steps in the process include collecting validity evidence for RESET through studies that examine the impact of receiving feedback, and studies that correlate teacher performance to student growth. The processes undertaken to create RESET could be applied to create observation systems across other content areas to support the improvement of instructional practice.

References

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching*. Guilford Press: New York.

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.

Bellini, S., Peters, J. K., Benner, L., & Hopf, A. (2007). A meta-analysis of school-based social skills interventions for children with autism spectrum disorders. *Remedial and Special Education, 28*(3), 153-162.

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2010). Reading Comprehension Instruction for Students With Learning Disabilities, 1995—2006: A Meta-Analysis. *Remedial and Special Education, 31*(6), 423-436.

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.

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

Brophy, J., & Good, T.L. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

Christenson, S. L., Ysseldyke, J. E., & Thurlow, M. L. (1989). Critical instructional factors for students with mild handicaps: An integrative review. *Remedial and Special Education, 10*(5), 21-31.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry.

Theory into practice, *39*(3), 124-130.

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.

Esterberg, K. G. (2002). *Qualitative methods in social research*. London: McGraw-Hill.

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.

Farley, A. N. (2017). Review of For Good Measure? Teacher Evaluation Policy in the ESSA Era. Boulder: National Education Policy Center. http://nepc.colorado.edu/

Gersten, R., Schiller, E. P., & Vaughn, S. (2000). *Contemporary special education research.* Mahwah, NJ: Erlbaum.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, *79*(3), 1202-42.

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

Hattie, J. A. C. (2009). Visible learning: A synthesis of 800+ meta-analyses on achievement. *Abingdon: Routledge*.

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.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23 (4),* 310-324. doi: 10.1080/08957347.2010.510956

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2016). *Recognizing Effective Special Education Teachers: Technical Manual*. Boise: Boise State University.

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

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. http://www.metproject.org

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

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

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

Linacre, J. M. (2014b). *A user guide to Facets, Rasch-model computer programs*. Chicago, IL: Winsteps.com.

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

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

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

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K, R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*(2), 137–148.

Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education*, *23*(4), 342-357.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M.C. Wittrock (Ed), *Handbook of research on teaching, 3rd ed (pp. 376-391)*. New York: Macmillan.

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.

Swanson, H. L., (1999). Instructional components that predict treatment outcomes

for students with learning disabilities: Support for a combined strategy and direct instruction model. *Learning Disabilities Research & Practice, 14*, 129-140.

Swanson, H. L., Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, *33*(2), 114–136.

Taylor, E. S. & Tyler, J. H. (2012). Can teacher evaluation improve teaching? Evidence of systematic growth in the effectiveness of midcareer 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

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

Item Measure Report from Many-Facet Rasch Measurement Analysis

Item Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
19	-1.61	.20	.81	.85
5	-.99	.16	.81	.80
21	-.80	.15	.86	1.03
18	-.72	.15	.83	.90
23	-.69	.14	.91	.84
6	-.53	.14	1.12	1.16
17	-.53	.14	.89	.91
4	-.48	.14	.77	.82
22	-.44	.14	.86	.87
10	-.39	.13	1.04	1.02
14	-.20	.13	1.11	1.09
20	-.15	.13	1.11	1.04
16	-.01	.12	.98	1.00
1	.16	.12	1.23	1.26
15	.20	.13	.84	.82
24	.34	.12	.91	.97
7	.38	.12	.93	.95
9	.38	.12	.97	.95
2	.44	.12	1.32	1.34
8	.47	.12	.96	.95
11	.49	.12	.95	.93
25	.60	.12	.92	.92
12	.93	.12	.92	.89
13	1.30	.12	1.11	1.09
3	1.86	.13	1.38	1.52
Mean (count = 25)	.00	.13	.98	1.00
SD	.76	.02	.16	.17

Note. Root mean square error (model) = .13; adjusted *SD* = .75; separation = 5.62; reliability = .97; fixed chi-square = 714.4; df = 24; significance = .00.

Table 3

Teacher Measure Report from Many-Facet Rasch Measurement Analysis

Teacher Number	Ability (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
10	-.17	.08	.87	.97
3	.26	.07	.86	.89
1	.27	.08	.95	.93
4	.50	.08	1.16	1.10
8	.78	.08	1.10	1.06
7	.79	.08	.90	.88
5	1.29	.09	1.03	1.16
6	1.42	.09	1.07	1.02
2	1.52	.09	.94	.83
9	1.64	.10	1.14	1.12
Mean (count = 10)	.83	.08	.1.00	1.00
SD	.62	.01	.11	.11

Note. Root mean square error (model) = .08; adjusted *SD* = .61; separation = 7.39; reliability = .98; fixed chi-square = 492.7; df = 9; significance = .00.

Table 4

Lesson Measure Report from Many-Facet Rasch Measurement Analysis

Lesson Number	Difficulty (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
3	-.26	.05	.99	.97
4	-.04	.05	1.02	1.05
1	.15	.05	1.04	1.04
2	.16	.05	.93	.93
Mean (count = 4)	.00	.05	.99	1.00
SD	.20	.00	.05	.06

Note. Root mean square error (model) = .05; adjusted *SD* = .19; separation = 3.72; reliability = .93; fixed chi-square = 43.4; df = 3; significance = .00.

Table 5

Rater Measure Report from Many-Facet Rasch Measurement Analysis

Rater Number	Severity (Logits)	Model SE	Infit MNSQ	Outfit MNSQ
1	-.64	.06	.84	.96
2	-.03	.05	1.17	1.13
3	.17	.05	.92	.85
4	.49	.05	1.02	1.05
Mean (count = 4)	.00	.05	.99	1.00
SD	.48	.00	.14	.12

Note. Root mean square error (model) = .05; adjusted *SD* = .47; separation = 9.07; reliability = .98; fixed chi-square = 232.7; df = 3; significance = .00.

Table 6

Bias Analysis Results – Teacher x Rater Interaction

Teacher - Rater	Observed Score	Expected Score	Bias Size	*t*	*p*
1 – 3	158	205.23	-1.06	-6.65	.000
10 – 3	157	184.68	-.65	-4.02	.000
5 – 2	234	255.21	-.57	-3.66	.000
4 – 2	188	212.15	-.56	-3.73	.000
7 – 3	205	228.46	-.52	-3.54	.000
2 – 1	266	277.13	-.48	-2.50	.014
6 – 4	222	241.33	-.46	-3.08	.002
5 – 1	254	264.52	-.42	-2.23	.027
8 – 3	210	228.01	-.40	-2.73	.007
8 – 1	245	258.38	-.39	-2.38	.019
9 – 4	234	247.21	-.35	-2.22	.028
8 – 4	228	213.68	.32	2.11	.037
7 – 2	250	236.67	.35	2.08	.039
1 – 4	206	188.70	.37	2.54	.012
1 – 2	231	211.43	.46	2.89	.004
8 – 2	247	229.91	.48	2.71	.008
10 – 2	197	176.89	.48	3.07	.002
7 – 1	272	258.70	.49	2.37	.019
4 – 1	262	242.15	.69	3.33	.001
6 – 3	274	253.03	.76	3.53	.000
2 – 3	277	256.34	.80	3.55	.000
9 – 3	287	260.03	1.33	4.58	.000
5 – 3	286	248.45	1.60	5.69	.000

Note. Observed and expected scores are based on the total possible number of points (300) across the observed count of items (100 = 25 items x 4 lessons).

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

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