12-2019

Variance and Reliability in Special Educator Observation Rubrics

Angela R. Crawford
Boise State University

Evelyn S. Johnson
Boise State University

Laura A. Moylan
Boise State University

Yuzhu Zheng
Boise State University

Sources of Variance in Special Educator Observation Rubrics

Angela R. Crawford*
Boise State University
angelacrawford1@boisestate.edu

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

Laura A. Moylan
Boise State University

Yuzhu Zheng
Boise State University

February 2018

Author Note

Angela Crawford, Project RESET, Boise State University; Evelyn S. Johnson, Department of Early and Special Education, 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. Angela Crawford, Boise State University, 1910 University Dr., MS 1725, Boise, Idaho 83725-1725. Email: angelacrawford1@boisestate.edu

Citation: Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (2018). Variance and Reliability in Special Educator Observation Rubrics. *Assessment for Effective Intervention*, 1534508418781010.

Abstract

This study describes the development and initial psychometric evaluation of a Recognizing Effective Special Education Teachers (RESET) teacher observation instrument. Specifically, the study uses generalizability theory to compare two versions of a rubric, one with general descriptors of performance levels and one with item-specific descriptors of performance levels, to evaluate special education teacher implementation of explicit instruction. Eight raters participated in viewing and scoring videos of special education instruction. Data collected from raters were analyzed in a three facet, crossed, mixed-model design to estimate the variance components and reliability indices. Results show lower unwanted sources of variance and higher indices of reliability with the rubric with item-specific descriptors of performance levels. Contributions to the field of teacher evaluation are discussed.

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

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. To be useful, a teacher observation system must facilitate accountability, support growth and development of professional practice, and provide accurate, reliable ratings and specific feedback (Hill & Grossman, 2013). To accomplish this, observation systems must meet two criteria: 1) they must be context specific to provide concrete guidance on how to improve practice and 2) they must be psychometrically sound, providing accurate and consistent evaluations of a teacher's ability to implement the desired instructional practices (Hill & Grossman, 2013). Many observation systems, however, are poorly aligned with the evidence-based instructional practices (EBPs) within the relevant content area or context, limiting the quality of the feedback provided to teachers through this mechanism (Grossman, Compton, Igra, Ronfeldt, Shahan, & Williamson, 2009). This is especially the case in special education, a field for which there are few instruments that detail the EBPs that are effective for students with disabilities (SWD) (Johnson & Semmelroth, 2014).

Research examining teacher observation instruments has indicated significant limitations with many current observation tools. Current tools often use general descriptors designed for broad application across contexts which limit their usefulness as tools for feedback, limit the accuracy with which performance is evaluated, and can contribute to biased ratings (Hill & Grossman, 2013). Large scale studies of these more general observation systems have indicated a propensity for bias in scores, in which the majority of teachers are rated as proficient or better (Kane & Staiger, 2012). Although teacher evaluations are often interpreted as a 'true' measure of teacher quality, a number of studies have indicated that many factors contribute to variance in the scores, suggesting that multiple facets of observation systems should be investigated (Hill, Charalambous & Kraft, 2013; Johnson & Semmelroth, 2015; Kane & Staiger, 2012). Further, research showing significant variance around theoretically meaningful cut scores suggests that there remains lack of clarity about quality instruction and what that looks like in practice (Cohen & Goldhaber, 2016; Polikoff, 2015).

Effective teacher observation systems require deliberate construction and thorough psychometric evaluation. Deliberate construction of context specific rubrics can be accomplished through extensive review and synthesis of the research that describes the elements of the evidence-based practices and empirical analysis of data describing observable levels of implementation in classrooms. Thorough psychometric evaluation can be undertaken through rigorous analysis during the development phase. One possible reason for the lack of context specific instruments that have been validated for large-scale use is that this degree of depth during development can be a time-consuming and expensive process. More research is needed to ensure that this time and expense results in psychometrically sound instruments describing observable levels of performance that provide accurate, concrete guidance to improve instructional practice.

In this manuscript, we describe the development and psychometric evaluation of a special education teacher (SET) observation system, Recognizing Effective Special Education Teachers (RESET). We first provide an overview of the RESET observation system and methods commonly used for developing rubrics. We then describe a study that employs generalizability theory (G-theory; Brennan, 2001; Cardinet, Johnson & Pini, 2010; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991) to compare the sources of variance and reliability indices in two versions of an observation rubric—one with general descriptors of performance levels and one with item-specific descriptors of performance levels.

RESET Observation Rubrics

RESET is a federally funded project with the goal of leveraging the research and literature describing best practices for students with high incidence disabilities to create observation rubrics that are specific and that will provide SETs with actionable feedback. The theory of change that drives RESET is depicted in Figure 1. By providing SETs with baseline evaluations of their instruction, setting goals, and providing them with feedback that is actionable, we expect to see improvements in instructional practice, and ultimately in student outcomes. To develop the RESET system, we followed the principles of Evidence Centered Design (Mislevy, Steinberg & Almond, 2003; see Johnson, Crawford, Moylan & Zheng, in press for a detailed description).

Several sources served to inform the starting points for developing this observation system, including the Council for Exceptional Children and CEEDAR Center's High-Leverage Practices (McLeskey et al., 2017), IES practice guides (Gersten et al., 2008; Gersten et al., 2009), meta-analyses of instructional practice for SWD (see for example: Berkeley, Scruggs & Mastropieri, 2010; Dennis et al., 2016; Dexter & Hughes, 2011; Gersten, Chard, Jayanthi, Baker, Morphy & Flojo, 2009; Gillespie & Graham, 2014; Jitendra, Lein, Im, Alghamdi, Hefte & Mouanoutoua, 2018; Stockard, Wood, Coughlin & Rasplica Khoury, 2018; Swanson, 1999), and descriptions of practice based on the research (Archer & Hughes, 2011). After identifying the practices for inclusion in RESET, we organized them into three domains: 1) instructional methods, 2) content organization and delivery, and 3) individualization. Within each category, we outlined the rubrics to create an overall blueprint for RESET. The list of RESET rubrics is included in Table 1.

To create individual items for each rubric, we first extracted the critical components of that practice from the literature, then reviewed and synthesized them into a coherent set of elements. Then, we drafted a set of items to describe proficient implementation of that practice. We refined the descriptors by reviewing video recorded lessons collected from SETs, and by 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 items that described proficient implementation.

Development of Rubric Rating Scales

The process just described was followed for all RESET rubrics, for the remainder of this manuscript, we will focus on the Explicit Instruction rubric to further describe the development of descriptors of performance for the RESET rubrics. Once the items describing proficient implementation were developed, we needed to create the scoring rules to describe the various levels of implementation of that practice. Following the model of the National Professional Development Center on Autism, we used the general descriptions of 'implemented', 'partially implemented' and 'not implemented' (Wong et al., 2015). However, we were uncertain as to the need for developing detailed descriptors for each item of the rubric across levels of implementation, or whether the general categories of *partially implemented* and *not implemented* would suffice. Although the research base on the development and psychometric evaluation of rating scales is limited, there are studies that report on the development of rating scales in contexts other than teacher observation, for example, in writing (Knoch, 2009), language (North & Schneider, 1998; Papageorgiou, Xi, Morgan & So, 2015), and music (Norris & Borst, 2007). This body of work reports on comparisons of general descriptors versus specific descriptors, arguing that specific descriptors (a) enable test users to more readily interpret to test results, (b) provide a common standard to raters, thus enhancing the reliability and validity of high inference assessments, and (c) transmit diagnostic information to the examinee (Alderson, 1991; Papageorgiou et al., 2015; Pollitt & Murray, 1996). Empirical analyses support these arguments, showing that specific descriptors have resulted in higher reliability across raters (Knoch, 2009; Norris & Borst, 2007) and higher construct validity (Knoch, 2009).

These findings have important implications for teacher observation instruments. Teacher observation instruments are often developed with the purposes of maximizing rater reliability, reporting results to external assessors. However, they are also designed for efficient implementation, typically with general descriptors of performance that can be applied across varied settings (Hill, et al, 2012; Hill & Grossman, 2013). While it is likely that context-specific instruments are more time consuming and costly 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; Hill & Grossman, 2013). Given the importance of sound development, psychometric evaluation, and the ability to provide actionable feedback, there is a need for research that reports on the development process, the rationale for decisions, and psychometric properties of teacher observation instruments (Garcia Gomez, et al., 2007; Hill, Charalambous, & Kraft, 2012; Papageorgiou, et al., 2015).

Therefore, the purpose of the current study is to compare a rating scale with general descriptors of performance levels of implementation of to a rating scale with item-specific descriptors of performance levels derived from performance data. Through the use of a generalizability (g) and decision (d) study, we examined the following research questions:

1. Do the ratings produced across the two versions of the explicit instruction rubric differ in terms of the relative contribution of sources of variance?

2. Do the ratings produced from these two versions of the rubric differ in terms of their indices of generalizability and dependability?

Methods

Participants

Special Education Teachers. A total of 10 special education teachers were recruited from across three states to participate in this study (see Table 2). Teacher participants were paid a $500 stipend for providing 20 videos across the 2015-16 school year. All participants provided video recorded lessons that reflected their use of explicit instruction in either reading (n = 5) or math (n = 5) intervention. All instruction took place in intervention or pull-out settings. Nine teachers taught in elementary schools and one in a middle school. All teachers were female, with an average experience level of 11.55 years (8.46 SD). Five had undergraduate degrees and five had graduate degrees.

Raters. A total of eight raters participated, with different raters assigned to phase one (n = 4) or phase two (n = 4) to control for bias (see Table 3). Raters were recruited and selected on the basis of experience with instruction for SWDs, experience with explicit instruction, and experience with teacher observation. Although the decision to use different raters for each phase of the study confounds raters with the different versions of the rubric, this was determined to be less problematic than allowing interpretations from scoring with the phase one rubric to influence interpretations using the phase two rubric and to limit the possibility of rater fatigue. Additionally, in G theory, reliability is understood as

the degree to which we can generalize from one observation to a universe of observations (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 15). Hence, G theory supports the inference that the observed score is a *universe score*, and permits generalizing from a specific sample to the universe of interest (Shavelson & Webb, 1981, pp. 133-137).

Measures. In both phases of the study we used the RESET Explicit Instruction rubric. In phase one, the rubric contained items that described the proficient level of implementation, and then general descriptors of 'partially implemented' or 'not implemented' were used. Each item is scored on a three-point scale where a score of 3 is proficient implementation, a 2 is partial implementation and a 1 is not implemented. In phase two, the rubric included the same items along with the fully developed descriptors for each item for each level of implementation. The methods used to develop these descriptors is described elsewhere (Johnson et al., in press). Figure 2 contains a sample of items to demonstrate the differences in the two versions of the rubric.

Procedures

Video Collection. All special education teacher participants were asked to video record weekly lessons of their instruction with a consistent group of students using the Swivl® video capture and upload system. Each teacher contributed a total of 20 videos over the 2015-16 year. Videos are used by the RESET research team to test and refine the rubrics that comprise the RESET observation system. For this study, after first ensuring that the videos had adequate audio and video quality, four videos from each teacher were randomly selected, resulting in a total of 40 videos. Videos ranged in length from 20-40 minutes. The videos were edited to remove any time at the beginning or end that did not reflect instruction (e.g., the teacher began recording a few minutes before students entered the classroom). Each video was assigned an identification number and listed in unique, random order for each rater to control for order effects.

Rater Training. Rater training was organized in the same manner for phases one and two. Raters were provided with an overview of the RESET project goals, a description of how the explicit instruction rubric was developed, and a description of the meaning and intent of each item. Research project staff then explained each item of the EI rubric and clarified any questions the raters had about the items. Then, raters watched and scored a video that had been scored by project staff. The scores were reviewed and discussed to include the rationale for the score that each item received.

Raters then scored two videos independently, and these scores were reconciled with a master coded rubric for each video. Any disagreements in scores were reviewed and discussed. To determine rater agreement, Kendall's coefficient of concordance, *W*, was selected to allow for ordinal data with multiple raters. For the first video in phase one, the four raters statistically significantly agreed in their ratings, $W = .552$, $p < .001$, indicating that agreement between the raters can explain 55.2% of the variability that would come with perfect agreement. For the second video in phase one, the four raters statistically significantly agreed in their ratings, $W = .596$, $p < .001$. For the first video in phase two, the four raters statistically significantly agreed in their ratings, $W = .478$, $p < .005$, and for the second video, $W = .544$, $p < .001$. This level of exact agreement is consistent with that reported by other teacher observation studies (Cash, Hamre, & Pianta, 2012; Kane & Staiger, 2012).

Raters were then assigned a randomly ordered list of videos to reduce teacher and order effects, and were asked to evaluate the videos following the assigned order, to score each item, to provide time stamped evidence that they used as a basis for the score, and to provide a brief explanation of the rationale for their score. Raters were also provided with a training manual that included descriptions of each item, along with examples for each item across each level of performance. In each phase, raters were given a period of six weeks to complete their ratings.

Data Analysis

Generalizability theory (Brennan, 2001; Shavelson & Webb, 1991) was used to examine the sources of variance and measurement error in both versions of the rubric. Using EduG v. 6.1, we employed a four facet, fully crossed, mixed-model design with teachers, lessons, raters and items (T x L x R x I) to estimate the variance components. In this analysis, teachers represent the object of measurement—the facet across which the instrument is intended to differentiate. Lessons, raters, and items represent the facets related to instrumentation, across which one wishes to minimize variance. Items were identified as a fixed facet because they do not represent a sample from a larger pool. A D-study was also conducted to identify the number of lessons and raters that would be needed to optimize score

reliability with each rubric. Although the data collected from the rubric are ordinal, the sample size is too small to apply ordinal G-theory (Ark, personal communication, January 12, 2018). Therefore, the data were analyzed as though they were continuous, resulting in coefficients that represent their lower-bound estimates (Ark, 2015).

All items were scored using a three-point numeric scale with an additional option to indicate and item as 'not applicable' (N/A). Scores of N/A were handled in the same way as missing data. Missing data and N/A scores were imputed using the mode on the item for that teacher by that rater across the three other videos (Shavelson, personal communication, November 29, 2016). In phase one, 17 (0.39%) scores were imputed in this manner, and in phase two, 48 (1.2%) scores were imputed.

Results

Results of the analysis of variance components for phase one and phase two are presented in Table 4. For each facet and interaction, the table provides the estimated variance components, standard error (SE), and percentage contribution to the total variance (%). The variance component for teachers (T) shows the amount of systematic variance in their ability to implement explicit instruction. Because teachers represent the object of measurement, ideally this component would have the highest variance. As shown, several other sources of variation are greater than T, with the residual (TLRI, error) as the highest. The higher variance associated with these other facets, interactions, and error may be indicative of a lack of precision in the rubrics or the inconsistency of raters. Though items are a component of instrumentation and not the object of measurement, variance related to items is acceptable as one would expect some items to be more difficult than others. The percentage of variance attributable to teachers, lessons, and error was similar for both versions of the rubric, with error accounting for the largest percentage of variance and lessons among the smallest contributions. The low variance attributable to lessons suggests that both rubrics function consistently across lesson content, context, and occasion. The high variance attributable to error indicates a considerable amount of "noise" present with both rubrics.

The percentage of variance attributable to the rater facet (R) and some rater interactions (TR and TLR) decreased in phase two, while the percentage of variance attributable to the item facet (I) and some item-related interactions increased (TI and LI). The decline in rater-related variance in phase two indicates that inter-rater and intra-rater scores were more consistent. The overall increase in variance attributable to item-related facets suggests better differentiation across items. Together, this may indicate that the specific item-level performance descriptors led to a decrease of rater-related factors such as halo-effects or drift. However, there was an increase in the teacher-lesson-item (TLI) interaction and the teacher-rater-item (TRI) in phase two and there remains a high portion of variability due to the undifferentiated TLRI interaction and error. Possible causes include imprecision in scoring criteria or rater bias. These interactions represent sources of problematic variance that should be addressed in future applications of the rubric.

Indices of Generalizability and Dependability

The G-study computes reliability as the ratio of differentiation variance (the object of measurement, in this case T) to the instrumentation variance (L,R, and interactions). Items were considered a fixed facet, and as such they do not contribute to this ratio. The reliability is expressed in two coefficients--a relative coefficient (generalizability, addressing rank order) and an absolute coefficient (dependability, position relative to a criterion). In terms of providing feedback to teachers on their implementation of explicit instruction, the generalizability coefficient would be acceptable. If the rubric is used as an observation instrument that describes proficiency, the dependability coefficient is more appropriate. With general descriptors of performance levels (phase 1), the relative coefficient was 0.61 (SE 0.18) and the absolute coefficient was 0.52 (SD 0.21). With item-specific descriptors (phase 2), the relative coefficient waas 0.74 (SE 0.14) and the absolute coefficient was 0.66 (SE 0.16). Because the ordinal data were analyzed as though continuous, these calculations are lower-bound estimates of reliability and may be attenuated (Ark, 2015).

It is important to note that in G-theory, coefficients are not precisely equivalent to reliability statistics from classical test theory. Because these coefficients consider multiple sources of variance (whereas reliability statistics only consider one), these coefficients are generally lower than reliability statistics. Therefore, it is more appropriate to compare them to each other than to standards that are typical for other measures of reliability (Mashburn, Downer, Rivers, Brackett, & Martinez, 2014). The guidance in the literature suggests coefficients > .70 an acceptable reliability estimate for observation instruments (Nunnally & Bernstein, 1994; Shavelson & Webb, 1991; Erlich & Shavelson, 1976, 1978). Particularly when adjusting for attenuation, the rubric with item-specific descriptors had better reliability and more closely approaches these thresholds.

Decision (D) Study

We used the information generated by the G-study to further investigate the lesson and rater facets, and specifically, to examine the number of raters needed to score each lesson and the number of lessons for each teacher needed to achieve acceptable reliability. Figure 3 shows the relative and absolute G coefficients for both phase one and two rubrics as the number of raters are adjusted under conditions with a fully-crossed design with four lessons. For phase one, the G coefficients range from 0.39 to 0.63. For phase two, the G coefficients range from 0.52 to 0.82. This finding suggests that the use of specific item-level performance descriptors has the potential to provide greater reliability than rubrics that use more general performance descriptors across a variety of rater designs. Figure 4 shows the relative and absolute G coefficients for both phase one and two rubrics as the number of lessons are adjusted under conditions with a fully-crossed design with four raters. For phase one, the G coefficients range from 0.38 to 0.54. For phase two, the G coefficients range from 0.55 to 0.67. This finding suggests that use of the specific item-level performance descriptors results in greater reliability across designs with different numbers of lessons observed.

Discussion

The purpose of this study was to compare an explicit instruction observation rubric with general descriptors to one with item-specific descriptors. It has been argued that observation instruments must be context specific and detailed to provide actionable feedback to teachers on how to improve instructional practice (Hill & Grossman, 2013). However, creating and validating instruments with this level of detail is time-consuming. The results of this study are consistent with those reported across other contexts (Knoch, 2009; Norris & Borst, 2007), and suggest that the additional resources to create specific, detailed performance descriptors are warranted.

Major Findings and Implications for Practice

The first research question addressed the sources of variance within the observation instrument. In the case of observation instruments, the rater facet can be unduly influential. Specifically, raters constitute an important source of variation in observed scores that is not desirable because it threatens the validity of the inferences that may be drawn from the assessment results (Eckes, 2011). This is particularly the case when raters evaluate performances using high inference instruments that require expertise in the observed practice (Baker et al., 2006; Nelson-Walker et al., 2013, Smolkowski & Gunn, 2012), as is the case with the RESET explicit instruction rubric.

The strength of the g study approach to examining observation instruments is in the information about sources of variance, allowing for improvements in the instrument and measurement design aimed at minimizing bias (Cardinet, et al, 2010; Cronbach, et al., 1972). Evaluating SETs ability to implement explicit instruction with item-specific performance descriptors led to less unwanted error associated with raters. The increase in item and item-related variance also suggests that item-specific performance descriptors support the aim of differentiating performance across teachers (Kraft & Gilmour, 2017).

The second research question examined indices of generalizability and dependability, important considerations for making inferences about an observed score to the universe score of a teacher's ability to effectively implement explicit instruction. The reliability indices achieved with the more detailed rubric at 0.74 (relative) and 0.66 (absolute), are promising, especially considering the possible attenuation of our G coefficients due to ordinal data (Ark, 2015). However, there is still too much variance attributable to instrument or measurement error that requires further steps to address. For example, in addition to providing a more rigorous rater training, to provide raters with more support during the rating period, we have developed a detailed training manual that includes both an explanation and examples of items across performance levels.

The d study provides important considerations for the application of RESET in practice. School systems will likely find it not feasible to employ multiple raters across multiple observations for each special education teacher. Working to minimize the variance attributable to facets of the rubric that should not unduly influence the evaluation of instruction and the feedback provided to teachers will be critical for the successful and fair use of RESET to make high stakes decisions.

Limitations and Implications for Future Research

There are a number of limitations to this research that must be addressed. First, this research was conducted with a limited number of participants. The small sample prevented us from employing methods to analyze ordinal data, and therefore the coefficients reported likely reflect lower bound estimates (Ark, 2015). Additionally, the number of teachers and raters used in this study limits the ability to generalize to a larger population. Research should include a greater number of SETs, ensuring diversity across demographics, school contexts, and career stages. SETs are also likely to be observed by individuals without extensive knowledge of best practices for SWD. Therefore, more research is needed on the implementation of these rubrics with raters with diverse backgrounds. Second, by using different raters in the two phases of the study, the rater effect is confounded with the rubric. We made this choice to address issues of validity and the likelihood of bias and fatigue effects if raters were to score the same video twice with different rubrics. This limitation can likely only be overcome with very large samples of raters scoring the videos in a counter-balanced design, which then poses a different set of constraints when using generalizability theory (e.g. limitations of missing data when designs are not fully crossed). We chose to accept the limitation and base our interpretations on the ability of generalizability theory to provide statistics that allow inferences to be generalized to a larger sample. Third, this process describes the development of a single rubric on a single instructional model. Further research is needed to evaluate empirically developed rubrics for other instructional practices and content areas.

Despite these limitations, this study contributes to the research on performance measurement and evaluation of special education teacher practice by providing evidence that item-specific descriptors of performance levels offer greater reliability than do general descriptors. Future research is needed to ensure that the item-specific descriptors of performance levels facilitate the provision of feedback that is actionable for teachers and results in the improved implementation of evidence-based instructional practices.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990's: The communicative legacy* (pp. 71-86). Hemel Hempstead, UK: Modern English Publications.

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

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework* (Doctoral dissertation, University of British Columbia).

Baker, S. K., Gersten, R., Haager, D., & Dingle, M. (2006). Teaching practice and the reading growth of first-grade English learners: Validation of an observation instrument. *The Elementary School Journal, 107* (2), 199-219.

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2009). Reading comprehension instruction for students with learning disabilities, 1995–2006: A meta-analysis. *Remedial and Special Education*. *31*(6), 423-436.

Brennan, R L. (2001). *Generalizability theory.* New York, NY: Springer.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542.

Cohen, J. & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher 45*(6), 378-387.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: John Wiley & Sons.

Dennis, M. S., Sharp, E., Chovanes, J., Thomas, A., Burns, R. M., Custer, B., & Park, J. (2016). A Meta-Analysis of Empirical Research on Teaching Students with Mathematics Learning Difficulties. *Learning Disabilities Research & Practice, 31*(3), 156-168.

Dexter, D. D., Park, Y. J., & Hughes, C. A. (2011). A Meta-Item-specific performance level review of graphic organizers and science instruction for adolescents with learning disabilities: Implications for the intermediate and secondary science classroom. *Learning Disabilities Research & Practice, 26*(4), 204-213.

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

Erlich, O. & Shavelson, R. (1976). Generalizability of measures: A computer for two- and three-facet designs. Behavior Research Methods and Instrumentation, 8(4), 407-408.

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(2), 77-89.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

Garcia Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on scale-anchoring of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417-444.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting Students Struggling with Mathematics: Response to Intervention (RtI) for Elementary and Middle Schools. NCEE 2009-4060. *What Works Clearinghouse*.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*(3), 1202-1242.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2008). Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide.(NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education. *National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from http://ies. ed. gov/ncee/wwc/publications/practiceguides*.

Gillespie, A., & Graham, S. (2014). A meta-analysis of writing interventions for students with learning disabilities. *Exceptional Children*, *80*(4), 454-473.

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.

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. C., & Grossman, P. (2013.) Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371-401.

Jitendra, A. K., Lein, A. E., Im, S. H., Alghamdi, A. A., Hefte, S. B., & Mouanoutoua, J. (2018). Mathematical Interventions for Secondary Students With Learning Disabilities and Mathematics Difficulties: A Meta-Analysis. *Exceptional Children*, *84*(2), 177-196.

Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (in press). Using evidence-centered design to create a special educator observation system. *Educational Measurement: Issues and Practice.*

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.

Kraft, M. A. & Gilmour, A. F. (2017). Revisiting The Widget Effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249.

Mashburn, A., Downer, J., Rivers, S., Brackett, M. & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science 15*(2), 146-155. Doi: 10.1007/s11121-012-0357-3

McLeskey, J., Barringer, M-D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M., Lewis, T., Maheady, L., Rodriguez, J., Scheeler, M. C., Winn, J., & Ziegler, D. (2017, January). *High-leverage practices in special education.* Arlington, VA: Council for Exceptional Children & CEEDAR Center.

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

Nelson-Walker, N. J., Fien, H., Kosty, D. B., Smolkowski, K., Smith, J. L. M., & Baker, S. K. (2013). Evaluating the effects of a systematic intervention on first-grade teachers' explicit reading instruction. *Learning Disability Quarterly*, *36*, 215–230.

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*(3), 237-251.

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language testing, 15*(2), 217-262.

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

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language assessment quarterly, 12*(2), 153-177.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183-212.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *In M. Milanovic & N. Saville (Eds.), Performance testing, cognition and assessment. (pp. 74–91).* Cambridge, UK: Cambridge University Press.

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

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

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of Direct Instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research, ## (X),* 1-29.

Swanson, 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*(3), 129-140.

Wong, C., Odom, S. L., Hume, K. A., Cox, C. W., Fettig, A., Kurcharczyk, S., et al. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. Journal of Autism and Developmental Disorders. doi: 10.1007/s10803-014-2351-z

Table 1

Organization and Structure of RESET

Subscale	Content	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

Special education teacher participant teaching context and demographics

Teacher	Content	Grade	Context	Student: Teacher ratio	Years teaching	Highest degree
1	Math	4th	RR	5:1	18	MA
2	Math	3rd	ERR	1:1	10	MA
3	Math	4th	ERR	3:1	27	MA
4	Math	4th	RR	3:1	5	BA
5	Math	8th	RR	14:1	8.5	BS
6	Reading	2nd	RR	5:1	1.5	BA
7	Reading	6th	RR	6:1	20	BA
8	Reading	4th	RR	6:1	16.5	MA
9	Reading	4th	RR	4:1	7	MA
10	Reading	5th	RR	4:1	2	BA

Note. RR=resource room; ERR=extended resource room; SLD=specific learning disability; CI=cognitive impairment; LI=language impairment; DD=developmental delay; ADHD=attention-deficit hyperactivity disorder; OHI=other health impairment; ASD=autism spectrum disorder; EBD=emotional-behavioral disorder.

Table 3

Rater demographics

Rater	Gender	Position	Years Experience	Highest degree
		Phase 1		
1	Female	Teacher	10	B.A.
2	Male	Administrator	44	M.Ed.
3	Female	Post-doc researcher	9	Ed.D.
4	Female	Teacher, RtI lead	15	M.Ed.
		Phase 2		
5	Female	Teacher	3	B.A.
6	Female	RtI coordinator	29	Psy.S.
7	Female	Post-doc researcher	12	Ph.D.
8	Male	University faculty	40	Ph.D.

Table 4

Variance components across the two phases of the rubric*

Phase 1 Rubric				Phase 2 Rubric		
Variance	SE	%	Source	Variance	SE	%
0.044	0.031	7.0	*Teachers (T)*	0.044	0.026	7.6
-0.002	0.003	0.0	*Lessons (L)*	0.003	0.004	0.5
0.047	0.035	7.5	*Raters (R)*	0.026	0.019	4.5
0.065	0.020	10.0	*Items (I)*	0.074	0.025	12.2
0.043	0.016	6.9	*TL*	0.016	0.007	2.8
0.054	0.018	8.6	*TR*	0.036	0.012	6.1
0.041	0.006	6.5	*TI*	0.045	0.007	7.8
0.002	0.003	0.3	*LR*	-0.001	0.002	0.0
0.000	0.001	0.0	*LI*	0.001	0.001	0.2
0.017	0.004	2.7	*RI*	0.042	0.008	7.2
0.062	0.010	9.8	*TLR*	0.042	0.007	7.3
0.009	0.004	1.5	*TLI*	0.027	0.005	4.7
0.026	0.005	4.1	*TRI*	0.032	0.005	5.5
-0.001	0.002	0.0	*LRI*	-0.002	0.002	0.0
0.218	0.007	34.9	*TLRI*	0.196	0.006	33.8

**Note.* Phase 1 rubric used general performance descriptors, phase 2 rubric used item specific descriptors.