

Variance and Reliability in Special Educator Observation Rubrics

Assessment for Effective Intervention
2019, Vol. 45(1) 27–37
© Hammill Institute on Disabilities 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1534508418781010
aei.sagepub.com



Angela R. Crawford, EdD¹ , Evelyn S. Johnson, EdD¹,
Laura A. Moylan, MA¹, and Yuzhu Zheng, PhD¹

Abstract

This study describes the development and initial psychometric evaluation of the Recognizing Effective Special Education Teachers (RESET) observation instrument. 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, for evaluating special education teacher implementation of explicit instruction. Eight raters (four for each version of the rubric) viewed and scored videos of explicit instruction in intervention settings. The data from each rubric were analyzed with a four facet, crossed, mixed-model design to estimate the variance components and reliability indices. Results show lower unwanted sources of variance and higher reliability indices with the rubric with item-specific descriptors of performance levels. Contributions to the fields of intervention and teacher evaluation are discussed.

Keywords

special education teacher evaluation, explicit instruction, observation systems, generalizability theory

Teacher observation systems (OSs) are seen as an important component of education reform because they offer the opportunity to evaluate teaching practice, facilitate accountability, and support professional growth (Cohen & Goldhaber, 2016; Hill & Grossman, 2013). To accomplish this, OSs must meet two criteria: (a) they must be context specific to provide concrete guidance for improving practice and (b) they must provide accurate and consistent evaluations of a teacher's ability to implement the desired instructional practices (Hill & Grossman, 2013). Many OSs, however, are not designed to measure implementation of evidence-based instructional practices (EBPs) within a particular context, limiting the quality of the feedback provided to teachers through this mechanism (Grossman et al., 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).

Often, observation instruments used by districts and states for accountability are designed for broad application across contexts (Cohen & Goldhaber, 2016; Goe, Bell, & Little, 2008), which may limit the degree to which they can provide specific feedback to teachers working with SWD. For example, two commonly used instruments, framework for teaching (FFT; Danielson, 2007) and the classroom assessment scoring system (CLASS; Pianta, Hamre, Haynes, Mintz, & La Paro, 2006), measure implementation of practices across a general instructional settings, rather than focusing on practices that best support students in interventions.

Several observation instruments have been designed for special education, intervention contexts, or for special populations, including the Reading Instruction in Special Education (RISE; Klingner, Urbach, Golos, Brownell, & Menon, 2010), the Classroom Observations of Student–Teacher Interactions (COSTI; Smolkowski & Gunn, 2012), and the English Learner Classroom Observation Instrument (ELCOI; Baker, Gersten, Haager, & Dingle, 2006). However, these instruments were designed to characterize the nature of instructional interactions, address multiple aspects of classroom interactions, and/or provide feedback specific to grade bands or content areas. Design of an instrument with one purpose in mind does not necessarily support its application for another purpose (Alderson, 1991; Pollitt & Murray, 1996). Therefore, there is a need for observation instruments that (a) are designed to measure implementation of EBPs that are effective for SWD, (b) provide specific and actionable feedback to teachers who work with SWD, and (c) can be used in interventions where these instructional practices are crucial.

In addition to providing concrete guidance, observation instruments also must be accurate and reliable. Previous studies of observation instruments have indicated that many

¹Boise State University, ID, USA

Corresponding Author:

Angela R. Crawford, College of Education, Boise State University, 1910 University Dr., MS-1725, Boise, ID 83725, USA.
Email: angelacrawford1@boisestate.edu

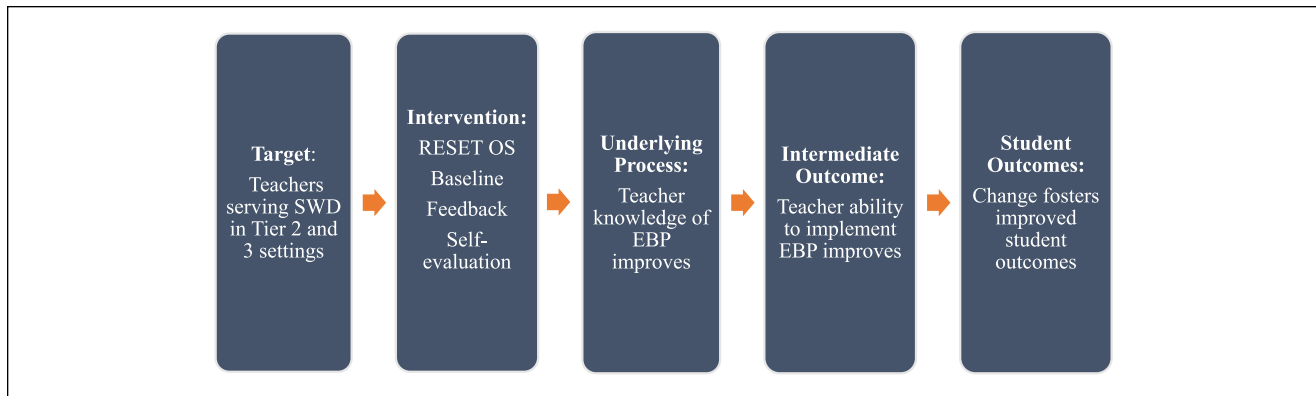


Figure 1. Theory of change for the RESET teacher OS.

Note. RESET = Recognizing Effective Special Education Teachers; SWD = students with disabilities; EBP = evidence-based instructional practices; OS = observation system.

factors contribute to unwanted variance in scores, suggesting that multiple facets (e.g., raters, occasions, scoring designs) of OSs should be investigated (Hill, Charalambous, & Kraft, 2012; Kane & Staiger, 2012). Also, studies of OSs have indicated a propensity for bias, leading to concerns about restriction of range in the scores—a lack of desirable variance (Cohen & Goldhaber, 2016; Kane & Staiger, 2012). Furthermore, research suggests that significant variance around theoretically meaningful cut scores indicates a lack of clarity about quality instruction and what it looks like in practice (Cohen & Goldhaber, 2016; Polikoff, 2015).

The study described here uses generalizability theory (G-theory; Brennan, 2001; Cardinet, Johnson, & Pini, 2010; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) to investigate these issues related to quality of observation in the Recognizing Effective Special Education Teachers (RESET) Explicit Instruction rubric. With G-theory, we investigate sources of variance attributable to multiple facets and compare that variance across two versions of the rubric that describe the quality of implementation of explicit instruction at different levels of specificity. We first provide an overview of the RESET OS and methods commonly used for developing rubrics. We then describe the application of G-theory to compare the sources of variance and reliability indices between two versions of the 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 will provide special education teachers (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 (2018), for a detailed description).

Several sources informed the starting points for developing this OS, including the Council for Exceptional Children and CEEDAR Center's High-Leverage Practices (McLeskey et al., 2017), IES practice guides (Gersten, Beckmann, et al., 2009; Gersten et al., 2008), meta-analyses of instructional practice for SWD (e.g., Berkeley, Scruggs, & Mastropieri, 2010; Dennis et al., 2016; Dexter, Park, & Hughes, 2011; Gersten, Chard, et al., 2009; Gillespie & Graham, 2014; et al., 2018; Stockard, Wood, Coughlin, & Rasplia 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: (a) instructional methods, (b) content organization and delivery, and (c) individualization. Within each domain, we outlined the rubrics to create an overall blueprint for RESET. The list of rubrics is included in Table 1.

To create individual items for each rubric, we first extracted the critical components specific to a particular 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 made appropriate revisions to create a set of items that described proficient implementation.

Table 1. Organization and Structure of RESET.

Subscale	Content	Rubrics
Instructional methods	NA	Explicit instruction Cognitive strategy instruction Peer-mediated learning
Content organization and delivery	Reading	Letter sound correspondence Multisyllabic words and advanced decoding Vocabulary Reading for meaning Comprehension strategy instruction Comprehensive reading lesson
	Math	Problem solving Conceptual understanding of number sense and place value, operations, fractions, algebra Procedural understanding of number sense & place value, operations, fractions, algebra
Individualization	Writing	Automaticity Spelling Sentence construction Self-regulated strategy development Conventions
		Executive function/self-regulation Cognitive processing accommodations Assistive technology Duration/frequency/intensity

Note. RESET = Recognizing Effective Special Education Teachers.

Development of Rubric Rating Scales

The process just described was followed for all RESET rubrics. For the remainder of this article, 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 of explicit instruction 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 of rating scales is limited, there are studies that report on development 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 research reports on comparisons of general versus specific descriptors, arguing that specific descriptors (a) enable test users to more readily interpret 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. Although it is likely that instruments with context-specific descriptors 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). 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 (Hill et al., 2012; Papageorgiou et al., 2015).

Therefore, the purpose of the current study is to compare a rating scale with general descriptors of performance levels to a rating scale with item-specific descriptors of performance levels. The goal is to develop a rating scale that

Table 2. Special Education Teacher Participant Teaching Context and Demographics.

Teacher	Gender	Race/Ethnicity	Content	Grade	Context	Student–Teacher Ratio	Years of Teaching	Highest Degree
1	Female	White	Math	4th	RR	5:1	18	MA
2	Female	White	Math	3rd	ERR	1:1	10	MA
3	Female	Asian	Math	4th	ERR	3:1	27	MA
4	Female	White	Math	4th	RR	3:1	5	BA
5	Female	White	Math	8th	RR	14:1	8.5	BS
6	Female	White	Reading	2nd	RR	5:1	1.5	BA
7	Female	White	Reading	6th	RR	6:1	20	BA
8	Female	White	Reading	4th	RR	6:1	16.5	MA
9	Female	White	Reading	4th	RR	4:1	7	MA
10	Female	White	Reading	5th	RR	4:1	2	BA

Note. RR = resource room; ERR = extended resource room.

Table 3. Rater Demographics.

Rater	Gender	Race/Ethnicity	Position	Years of Experience	Highest Degree
Phase 1					
1	Female	White	Teacher	10	BA
2	Male	White	Administrator	44	MEd
3	Female	White	Postdoc researcher	9	EdD
4	Female	White	Teacher, Rtl lead	15	MEd
Phase 2					
5	Female	White	Teacher	3	BA
6	Female	White	Rtl coordinator	29	PsyS
7	Female	White	Postdoc researcher	12	PhD
8	Male	White	University faculty	40	PhD

provides SETs with reliable and actionable feedback that will, ultimately, positively affect student outcomes. Through the use of G-theory, we examined the following research questions:

Research Question 1: Do the ratings produced with the two versions of the rubric differ in terms of the relative contribution of sources of variance?

Research Question 2: Do the ratings produced with the two versions of the rubric differ in terms of their indices of generalizability and dependability?

Research Question 3: How many raters and lessons are needed to achieve strong levels of dependability with the two versions of the rubric?

Method

Participants

SETs. A sample of 10 SETs were recruited from across three states to participate in this study. Table 2 provides information about teaching context and demographics for the teacher participants. Teacher participants were paid a

US\$500 stipend for providing 20 videos across the 2015–2016 school year. Teachers were asked to record lessons that reflected their use of explicit instruction. No further instruction regarding what practices constitute explicit instruction were given. Four teachers used district-provided programs a majority of the time. Six teachers used district-provided materials on occasion.

Raters. A total of eight raters participated, with different raters assigned to Phase 1 ($n = 4$) or Phase 2 ($n = 4$) to control for bias. Table 3 provides demographic information for the raters. Raters were recruited and selected on the basis of experience with instruction for SWDs, explicit instruction, and teacher observation. Although the use of 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 1 rubric to influence interpretations using the Phase 2 rubric and to limit the possibility of rater fatigue. In addition, in G-theory, reliability is understood as the degree to which we can generalize from one observation to a universe of observations (Cronbach et al., 1972). Hence, G-theory supports the inference that the observed score is a *universe*

Phase One Explicit Instruction Rubric (Sample of Items)					
Components	Item	Implemented Descriptor	Score	Evidence	Explanation
Identifying and Communicating Goals	1	The goals of the lesson are clearly communicated to the students			
	2	The stated goal(s) is specific			
	3	The teacher clearly explains the relevance of the stated goal to the students			
Alignment	4	Instruction is completely aligned to the stated or implied goal			
	5	All of the examples or materials are aligned to the stated or implied goal			
	6	Examples or materials selected are aligned to the instructional level of most or all of the students			

Phase Two Explicit Instruction Rubric (Sample of Items)						
Components	Item	3 - Implemented	2 - Partially Implemented	1 - Not Implemented	Score	Explanation
Identifying and Communicating Goals	1	The goals of the lesson are clearly communicated to the students.	The goals of the lesson are not clearly communicated to the students.	The goals of the lesson are not communicated to the students.		
	2	The stated goal(s) is/are specific.	The stated goal(s) is/are broad or vague.	There is no stated goal.		
	3	The teacher clearly explains the relevance of the stated goal to the students.	The teacher tries to explain the relevance of the stated goal to the students, but the explanation is unclear or lacks detail.	The teacher does not explain the relevance of the stated goal to the students.		
Alignment	4	Instruction is completely aligned to the stated or implied goal.	Instruction is partially or loosely aligned to the stated or implied goal.	Instruction is not aligned to the stated or implied goal.		
	5	All of the examples or materials selected are aligned to the stated or implied goal.	Some of the examples or materials are aligned to the stated or implied goal; OR examples and materials are somewhat aligned to the stated or implied goal.	Examples or materials selected are not aligned to the stated or implied goal.		
	6	Examples or materials selected are aligned to the instructional level of most or all of the students.	Examples or materials selected are aligned to the instructional level of some of the students.	Examples or materials selected are not aligned to the instructional level of most students.		

Figure 2. Sample of items on the Phase 1 and Phase 2 explicit instruction rubrics.
Note. OR = odds ratio.

score and permits generalizing from a specific sample to the universe of interest (Shavelson & Webb, 1991).

Measures. In both phases of the study, we used the RESET Explicit Instruction rubric. In Phase 1, the rubric contained items with descriptions of proficient implementation. Each item is scored on a 3-point scale where a score of 3 is *implemented*, a 2 is *partially implemented*, and a 1 is *not implemented* with an option to indicate an item as *not applicable* (NA). In Phase 2, 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., 2018). Figure 2 contains a sample of items to demonstrate the differences in the two versions of the rubric.

Procedures

Video collection. All SET participants were asked to video-record their instruction with a consistent group of students

using the Swivl® video capture system. Each teacher contributed a total of 20 videos over the 2015–2016 year. Videos are used by the RESET research team to test and refine the rubrics that comprise the RESET OS. 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 min. The videos were edited to remove any time at the beginning or end that did not reflect instruction (e.g., recording a few minutes before students entered the classroom) and assigned random identification codes.

Rater training. Rater training was organized in the same manner for Phases 1 and 2. Training occurred over the course of 1 week and consisted of approximately 12 hours of training and 6 hours of practice scoring. Research project staff provided raters with an overview of the project goals, a description of how the rubric was developed, and a description of the meaning and intent of each item. Project staff then answered any questions the raters had. Next, raters watched a video and were provided with master scores and rationales to serve as a model. These were reviewed and discussed in depth. On each of the next two days, raters scored a video independently, and these scores were reconciled with master scores. Any disagreements were reviewed and discussed in depth. Again, raters were provided with a copy of the master scores and rationales to serve as models.

To determine rater agreement, Kendall's coefficient of concordance, W , was used to allow for ordinal data with multiple raters. In Phase 1, the raters significantly agreed in their ratings of the first video, $W = .552$, $p < .001$, indicating that agreement between raters can explain 55.2% of the variability that would come with perfect agreement. For the second video, raters significantly agreed in their ratings, $W = .596$, $p < .001$. In Phase 2, raters significantly agreed in their ratings of the first video, $W = .478$, $p < .005$, and 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, & Myers, 2012; Kane & Staiger, 2012).

Raters were assigned a randomly ordered list of videos to reduce teacher and order effects. They were asked to evaluate the videos following the assigned order, score each item, provide time-stamped evidence used as a basis for the score, and provide a brief explanation of the rationale for their score. In each phase, raters were given a period of six weeks to complete their ratings.

Data Analysis

A generalizability study (G-study) was used to compare the sources of variance and reliability indices that occurred with the rubrics in Phases 1 and 2. Using EduG v. 6.1, we

employed a four-facet, fully crossed mixed-model design with teachers, lessons, raters, and items ($T \times L \times R \times I$) with both sets of data. 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. Because items are not sampled, they are identified as a fixed facet. A decision study (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 (T. 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).

Scores of NA were handled in the same way as missing data. Missing data and NA scores were imputed using the mode on the item for that teacher by that rater across the three other videos (R. J. Shavelson, personal communication, November 29, 2016). In Phase 1, 17 (0.39%) scores were imputed in this manner, and in Phase 2, 48 (1.2%) scores were imputed.

Results

G-Study: Sources of Variance

Results of the analysis of variance for Phase 1 and Phase 2 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 percentage of variance attributable to teachers (T), lessons (L), and the residual (TLRI, error) was similar for both versions of the rubric, with residual accounting for the largest percentage and lessons among the smallest. High variance attributable to TLRI, error indicates a considerable amount of “noise” present with both rubrics. Low variance attributable to L suggests that both rubrics function consistently across lesson content, context, and occasion. The variance for T shows the amount of systematic variance in teachers' implementation of explicit instruction; ideally, this component would have the highest variance. As shown, several other sources of variation are greater than T. The higher variance associated with these other facets, interactions, and residual may be indicative of a lack of precision in the rubrics or the inconsistency of raters. Though the facet items (I) is a component of instrumentation, variance related to I is acceptable as one would expect some items to be more difficult than others.

The percentage of variance attributable to the rater facet (R) and some related interactions (TR and TLR) decreased in Phase 2, whereas the percentage of variance attributable to I and some related interactions (TI and LI) increased. The

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.

decline in rater-related variance in Phase 2 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 item-specific 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) and teacher–rater–item (TRI) interactions in Phase 2. 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.

G-Study: 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, as a fixed facet, do not contribute to this ratio. The reliability is expressed in two coefficients—a generalizability coefficient (relative, addressing rank order) and a dependability coefficient (absolute, position relative to a criterion). For the purpose of providing feedback to teachers, the generalizability coefficient would be sufficient. For the purpose of providing a criterion-based evaluation, the dependability coefficient is more appropriate. In Phase 1, the generalizability coefficient was 0.61 ($SE = 0.18$) and the dependability coefficient was 0.52 ($SD = 0.21$). In Phase 2, the generalizability coefficient was 0.74 ($SE = 0.14$) and the dependability coefficient was 0.66 ($SE = 0.16$). Because the ordinal data were analyzed as though continuous, these cal-

culations are attenuated and represent lower-bound estimates of reliability (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 with each other than with standards that are typical for other measures of reliability (Mashburn, Downer, Rivers, Brackett, & Martinez, 2014). The guidance in the literature suggests coefficients > 0.70 are acceptable reliability estimates for observation instruments (Erlach & Shavelson, 1976, 1978; Nunnally & Bernstein, 1994; Shavelson & Webb, 1991). Particularly considering attenuation, the rubric with item-specific descriptors more closely approaches this threshold.

D-study: Raters and Lessons

We conducted a D-study to investigate the number of raters and lessons per teacher that would be needed to achieve acceptable reliability. Given the criterion-focused descriptors of *implemented*, *partially implemented*, and *not implemented*, we investigated designs that would result in a stronger dependability coefficient.

Figure 3 shows the results of the D-study. The graph on the left shows dependability coefficients as the number of raters are adjusted under conditions with a fully crossed design with four lessons. For Phase 1, the dependability coefficients range from 0.33 to 0.53. For Phase 2, the coefficients range from 0.46 to 0.69. Therefore, only

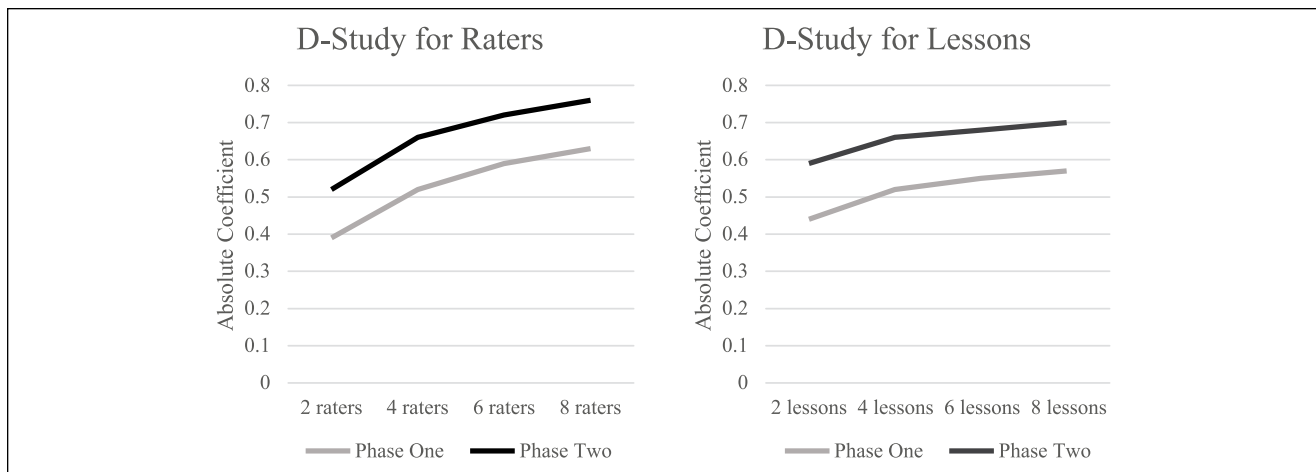


Figure 3. D-study results for raters viewing four lessons and D-study results for lessons observed by four raters.

item-specific descriptors approached the 0.70 reliability threshold in a design with eight raters.

The graph on the right of Figure 3 shows the dependability coefficients as the number of lessons are adjusted under conditions with a fully crossed design with four raters. For Phase 1, the coefficients range from 0.44 to 0.57. For Phase 2, the coefficients range from 0.59 to 0.70. These nonoverlapping data suggest that item-specific descriptors can result in greater reliability across designs, even when fewer lessons are 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 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 to provide teachers with feedback that will support improved outcomes for students in interventions.

Major Findings and Implications for Practice

The first research question addressed the sources of variance within the observation instruments. With 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 G-theory in examining observation instruments is in the information about sources of variance, allowing for improvements in the instrument and measurement design (Cardinet et al., 2010; Cronbach et al., 1972). Recently, two studies have applied G-theory to evaluate observation instruments for general education instruction, the Mathematical Quality of Instruction (MQI) instrument (Hill et al., 2012), and the FFT as part of the Measures of Effective Teaching (MET) project (Ho & Kane, 2013). Results of the study described here are consistent with these two previous studies. The rubric with item-specific performance descriptors demonstrated less unwanted error associated with raters, at approximately 18%. The studies of the MQI and FFT reported rater-related variance to range between 13% and 38%, depending upon the research design (Hill et al., 2012; Ho & Kane, 2013). Also, the 35% residual variance found in Phase 2 is similar to that found in with the MQI and FFT, where residual ranged from 22% to 46%, depending upon the research design (Hill et al., 2012; Ho & Kane, 2013).

Also promising at this stage of development is variance from Phase 2 that supports the aim of differentiating performance across teachers (Kraft & Gilmour, 2017), such as greater variance for teachers and items and lower variance related to lessons. Again consistent with MQI and FFT reporting teacher-related variance between 30% and 45% (Hill et al., 2012; Ho & Kane, 2013), the rubric used in Phase 2 showed approximately 36% of variance was teacher or item related. Lesson-related variance in Phase 2 was approximately 4%, whereas lesson variance ranged from 3% to 28% with MQI and FFT (Hill et al., 2012; Ho &

Kane, 2013). This suggests that the rubric with item-specific descriptors shows potential for stability across lessons.

The second research question examined indices of generalizability and dependability, important considerations for making inferences about a teacher's ability to effectively implement explicit instruction. Again, these results are similar to those reported with MQI and FFT, which report dependability (absolute) coefficients ranging from 0.60 to 0.73 for designs similar to those reported here (Hill et al., 2012; Ho & Kane, 2013). The dependability coefficient of 0.66 achieved with Phase 2 rubric is promising, especially considering the possible attenuation due to ordinal data (Ark, 2015).

Taken together, the results regarding variance and reliability show that desirable and undesirable variance in the RESET Explicit Instruction rubric with item-specific descriptors is similar to that of other instruments and that further development is warranted. Further development will aim to reduce the number of raters needed for acceptable levels of dependability. For example, in addition to planning for more rigorous rater training, we have developed a detailed training manual based on the item-specific descriptors that includes explanations of items and examples across performance levels.

This study reveals important considerations for the application of the rubric in practice. It is likely that school systems will not find it feasible to employ multiple raters across multiple observations for each SET. However, the D-study indicates that school systems may be able to conduct four or fewer observations; increasing the number of observations beyond four has minimal impact on the dependability coefficient. For practical application, therefore, it will be critical to minimize the variance attributable to raters and error, further reducing the number of observations and raters needed.

Overall, these findings suggest the RESET Explicit Instruction instrument can provide a valuable support to teachers for improving instruction in interventions. Although other instruments may also support instruction in interventions, such as RISE (Klingner et al., 2010), the COSTI (Smolkowski & Gunn, 2012), and the ELCOI (Baker et al., 2006), these instruments are limited to either lower grades and/or specific content. In contrast, the RESET Explicit Instruction instrument shows promise for evaluating the implementation of explicit instruction in interventions across content areas and grade levels.

Limitations and Implications for Future Research

There are a number of limitations to this research that must be addressed. First, the small sample size prevented us from employing methods to analyze ordinal data, and, therefore,

the coefficients reported reflect lower bound estimates (Ark, 2015). In addition, the number of teachers used in this study limits the ability to generalize to a larger population. Further research should include larger samples, ensuring diversity across demographics, school contexts, and career stages. SETs are also likely to be observed by individuals representing more diversity, including those 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 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 counterbalanced design, which then poses a different set of constraints when using G-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 G-theory to 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 SET 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 EBPs. Also future research is needed to link the implementation of EBPs in interventions to student outcomes. This work has the potential to provide quality assessments of instruction in interventions and provide teachers with feedback that can have maximum impact on the achievement of students.

Authors' Note

The opinions expressed are solely those of the authors.


Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Institute of Education Sciences, award number R324A150152 to Boise State University.

ORCID iD

Angela R. Crawford  <https://orcid.org/0000-0003-3646-0335>

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). Retrieved from <https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0166304>
- 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. *Elementary School Journal, 107*, 199–220. doi:10.1086/510655
- 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*, 423–436. doi:10.1177/0741932509355988
- 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 large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*, 529–542. doi:10.1016/j.ecresq.2011.12.006
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*, 378–387. doi:10.3102/0013189X16659442
- 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.
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- 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*, 156–168. doi:10.1111/ldrp.12107
- Dexter, D. D., Park, Y. J., & Hughes, C. A. (2011). A meta-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*, 204–213. doi:10.1111/j.1540-5826.2011.00341.x
- Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt, Germany: Peter Lang.
- Erlich, O., & Shavelson, R. (1976). Generalizability of measures: A computer program for two- and three-facet designs. *Behavior Research Methods and Instrumentation, 8*, 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*, 77–89.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*, 5–29. doi:10.1177/0265532209359514
- 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). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/rti_math_pg_042109.pdf
- 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*, 1202–1242. doi:10.3102/0034654309334431
- 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. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/rti_reading_pg_021809.pdf
- Gillespie, A., & Graham, S. (2014). A meta-analysis of writing interventions for students with learning disabilities. *Exceptional Children, 80*, 454–473. doi:10.1177/0014402914527238
- Goe, L., Bell, C., & 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/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf>
- Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*, 2055–2100. Retrieved from <https://tedd.org/wp-content/uploads/2014/03/Grossman-et-al-Teaching-Practice-A-Cross-Professional-Perspective-copy.pdf>
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*, 56–64. doi:10.3102/0013189X12437203
- Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*, 371–384. doi:10.17763/haer.83.2.d11511403715u376
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Retrieved from <https://files.eric.ed.gov/fulltext/ED540957.pdf>
- 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*, 177–196. doi:10.1177/0014402917737467

- Johnson, E. S., Crawford, A. R., Moylan, L. A., & Zheng, Y. (2018). Using Evidence-Centered Design to Create a Special Educator Observation System. *Educational Measurement: Issues and Practice*. <https://doi.org/10.1111/emip.12182>
- Johnson, E. S., & Semmelroth, C. L. (2014). Special education teacher evaluation: Why it matters, what makes it challenging, and how to address these challenges. *Assessment for Effective Intervention*, *39*, 71–82. doi:10.1177/1534508413513315
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Available from <http://www.metproject.org>
- Klingner, J. K., Urbach, J., Golos, D., Brownell, M., & Menon, S. (2010). Teaching reading in the 21st century: A glimpse at how special education teachers promote reading comprehension. *Learning Disability Quarterly*, *33*, 59–74. doi:10.1177/073194871003300201
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*, 275–304. doi:10.1177/0265532208101008
- Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*, 234–249. doi:10.3102/0013189X17718797
- 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*, 146–155.
- McLeskey, J., Barringer, M.-D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M., . . . Ziegler, D. (2017). *High-leverage practices in special education*. Arlington, VA: Council for Exceptional Children & CEEDAR Center.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67. doi:10.1207/S15366359MEA0101_02
- Nelson-Walker, N. J., Fien, H., Kosty, D. B., Smolkowski, K., Smith, J. L. M., & Baker, S. K. (2013). Evaluating the effects of a systemic intervention on first-grade teachers' explicit reading instruction. *Learning Disability Quarterly*, *36*, 215–230. doi:10.1177/0731948712472186
- 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. doi:10.1177/002242940705500305
- North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*, 217–262. doi:10.1177/026553229801500204
- 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*, 153–177. doi:10.1080/15434303.2015.1008480
- Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S., & La Paro, K. M. (2006). *CLASS Classroom Assessment Scoring System: Manual middle secondary version pilot*. Charlottesville, VA: Teachstone.
- Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *12*, 183–212. doi:10.1086/679390
- 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*, *27*, 316–328. doi:10.1016/j.ecresq.2011.09.004
- Stockard, J., Wood, T. W., Coughlin, C., & Rasplika Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*. Advance online publication. doi:10.3102/0034654317751919
- 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. doi:10.1207/sldrp1403_1
- Wong, C., Odom, S. L., Hume, K. A., Cox, C. W., Fetting, 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.