10-1-2015

Validating an Observation Protocol to Measure Special Education Teacher Effectiveness

Evelyn S. Johnson
Boise State University

Carrie L. Semmelroth
Boise State University

Validating an Observation Protocol to Measure Special Education Teacher Effectiveness

Dr. Evelyn S. Johnson

Dr. Carrie L. Semmelroth

Boise State University

Abstract

This study used Kane's (2013) Interpretation/Use Argument (IUA) to measure validity on the Recognizing Effective Special Education Teachers (RESET) observation tool. The RESET observation tool is designed to evaluate special education teacher effectiveness using evidence-based instructional practices as the basis for evaluation. In alignment with other studies (Bell et al., 2012), we applied and interpreted Kane's (2006) four inferences for trait observation: scoring, generalization, extrapolation, and decision rules. Results from this study show that acceptable levels of validity are promising for the RESET observation tool. Because the RESET observation tool is premised on the idea that by increasing the use of evidence-based practices, student achievement will also increase, further investigations into the relationship between fidelity of implementation of instruction and student achievement will be critical for moving project work forward.

Validating an Observation Protocol to Measure Special Education Teacher Effectiveness

Special education teacher evaluation systems are of high interest nationally because they will be used to support judgments about the quality of teaching that students with disabilities receive. With this purpose in mind, it follows that a special education teacher evaluation system should be based on a conceptual framework that defines an effective special education teacher as one who employs evidence-based practices to improve student outcomes (Johnson & Semmelroth, 2014a). Through an evaluation system that emphasizes the use of evidence-based instruction, special education teachers will prioritize the use of practices that are most likely to positively impact student learning, and ultimately, outcomes for students with disabilities will improve (Cook, Tankersley, & Landrum, 2009).

An evaluation system designed on this framework requires the use of an observation protocol that captures the trait of effective special education teaching. A trait, as defined by Kane (2006), is a disposition to behave or perform in some way under a range of circumstances. To capture the trait of effective special education teaching, an observation protocol should specify the components of evidence-based practices (EBP) to be incorporated into practice, and should provide a mechanism through which special education teachers receive feedback on their observed instruction relative to the desired EBP. An examination of the effects of instruction on outcomes should demonstrate a high correlation between the use of EBP and student growth. It is upon this basic connection between effective instructional practice and student outcome data that

the Recognizing Effective Special Education Teachers (RESET) observation system was developed (Johnson & Semmelroth, 2014; Semmelroth, 2013; Semmelroth & Johnson, 2014).

There are two main purposes of the RESET observation system: 1) to evaluate special education teacher effectiveness and 2) to improve special education teacher instruction in the classroom. The focus of RESET differs markedly from the current emphasis on value-added models (VAM), which have as their primary purpose determining teacher attribution to student outcomes as measured by performance on state standardized assessments. The distinction is important, because special education has been and continues to be a high demand field, with high turnover and attrition rates (Connelly & Graham, 2009), and with high percentages of emergency or alternate-route certified teachers (McLeskey, Tyler, & Flippin, 2004). Models of teacher evaluation that attempt only to differentiate between effective and ineffective teachers based primarily on student test scores (e.g. VAM) may not be easily applied to special education for a variety of measurement reasons (Buzick & Laitusis, 2010; Holdheide, Browder, Warren, Buzick, & Jones, 2012) and will do little to address the critical shortages in special education. Therefore, we argue that a special education teacher evaluation system should draw on the extensive research base in special education specifying evidence based practices and the resulting effects on student growth.

Establishing Validity of a Special Education Teacher Evaluation System

Prior to adopting a high stakes teacher evaluation system, it is critical to evaluate its psychometric defensibility to ensure that the system will accomplish what it purports to accomplish while limiting the unintended, negative consequences (Herlihy et al., 2014). In a recent review of state teacher evaluation systems, Herlihy et al. (2014) note that few states had specified programs of research to examine the effects of implementing their teacher evaluation system. Considering the stakes attached, we argue, as others have (e.g., Bell et al., 2012; Herlihy et al., 2014), that regardless of the model of evaluation system adopted, it is imperative to apply the same assessment standards to teacher evaluation systems as have been applied to other areas of educational assessment.

However, given the challenges of measuring a construct as complex as effective special education teaching, traditional approaches to establishing psychometric soundness may be insufficient. Because observation-based measures require a series of inferences to be made about a small number of performances to the universe of possible performances of a complex construct, they are arguably best validated through more comprehensive approaches. One such approach to validity is Kane's (2006) argument-based approach. In Kane's model, there are two types of arguments to be specified in a validation effort: the interpretive/use argument (IUA) and the validity argument (Kane, 2013). The IUA presents "the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2013, p. 23). The validity argument evaluates those inferences and assumptions using empirical data and analytic reasoning. Kane summarized the argument-based approach to validity succinctly, "The approach is quite simple: state what is being claimed and evaluate the claims being made" (Kane, 2006, p. 451).

Specific IUA and validity arguments must be viewed in light of the proposed uses of the resulting scores on a measure. This involves a straight-forward, two-step process: 1) the development of a clear statement of the claims inherent in the proposed interpretations and uses of the measure and resulting scores, and 2) a critical evaluation of these claims. Bell et al. (2012) outline an application of Kane's argument based approach to validity in a recent article on observing and evaluating algebra teachers. Following Kane's approach and Bell et al.'s application of his approach, the following inferences should comprise the IUA for a special education teacher observation system: 1) scoring, 2) generalization, 3) extrapolation, and 4) decision rules (Bell et al., 2012; Kane, 2013). Each of these inferences will be explained within the context of the RESET special education teacher evaluation system, following a brief description of RESET.

Recognizing Effective Special Education Teachers (RESET) Observation Protocol. RESET is a state-sponsored project that was developed to meaningfully include special education teachers into a state's teacher evaluation system. The RESET project was designed to 1) define what an effective special education teacher is, and 2) to create an evaluation system that reliably identifies effective special education teachers and provides them with feedback to improve their practice. To develop the definition of an effective special education teacher, we first considered the complexity of the role. Special education teachers work under a variety of conditions, with a heterogeneous population, and support student progress towards individualized goals. In addition to providing individualized instruction, special education teachers manage caseloads, coordinate related services, and provide consultation in the general education classroom. However, when developing the definition of effective special education teaching however, we focused on instructional practice because it is the single component of a special education teacher's responsibility that has a documented, direct, and substantial impact on student outcomes.

Based on this rationale, the following definition was created to guide the conceptual framework of the RESET observation protocol: effective special education teachers are able to identify a student's strengths and needs, implement evidence-based instructional practices and demonstrate student growth (Johnson & Semmelroth, 2012; Semmelroth, Johnson, & Allred, 2013). Therefore, the RESET observation protocol was designed to collect observations of special education teacher's instructional practice and to evaluate these observations according to specifications developed from the research explaining the critical components of a variety of evidence-based practices for students with disabilities. A significant body of research has established a number of effective instructional practices to meet the needs of students with disabilities (see for example, Browder, Ahlgrim-Delzell, Spooner, Mims, & Baker, 2009; Browder & Cooper-Duffy, 2003; Chard, Ketterlin-Geller, Baker, Doabler, & Apichatabutra, 2009; Cook & Odom, 2013; Fuchs & Fuchs, 2005; Gersten et al., 2009; National Autism Center, 2009; Odom, 2009; Spooner, Knight, Browder, & Smith, 2012). This body of research guided the development of detailed rubrics that are the primary component of the RESET observation protocol and are used to evaluate a special education teacher's instructional practice. RESET is flexible enough to be used across special education settings because it includes rubrics for a substantial and growing number of evidence-based practices, and specific enough on its focus on EBP to provide targeted, individualized feedback for special education teachers.

To evaluate a teacher using RESET, special education teachers are video-taped across multiple lessons using the Teachscape video capture system. Then, trained raters use the associated rubrics to assign a score and provide feedback on a teacher's instructional practice. In its current design, RESET relies on the use of a four-point scale to align with Danielson's Framework for Teaching (FFT; Danielson, 2013), because RESET was developed in a state that adopted FFT for general education teachers. More detailed information about the scoring process is included in the methods section. To date, over 4,000 minutes of special education instruction across a variety of settings have been used to inform the continued development of RESET. Current research studies have focused largely on establishing reliability and determining the optimal number of raters and observations to ensure reliable results. Using generalizability theory to examine data, initial research suggests that optimal results are reached when evaluations are based on four observations and four raters (Semmelroth, 2013; Semmelroth & Johnson, 2014). These results are consistent with those reported in large-scale studies of teacher observation systems (Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013). Given that this may not be feasible for implementation, current studies are underway to determine whether more rigorous rater training efforts could reduce the number of raters required to obtain acceptable thresholds of reliability.

Interpretive/Use Argument (IUA) Inferences. To examine the validity of RESET in accomplishing its dual purposes of identifying effective special education teachers and providing feedback on instructional practice, we applied Kane's IUA. As outlined by Kane (2006, 2013) the IUA for observation protocols includes four major inferences: 1) scoring, 2) generalization, 3) extrapolation, and 4) decision rules. Each inference relies on several assumptions explained below (Kane, 2006).

Scoring: Test performances are communicated through scoring systems that assign scores to observed performances. Assumptions about scoring include that 1) the scoring rule is appropriate; 2) the scoring rule is applied accurately and consistently; 3) the scoring is bias free; and 4) the data fit the scoring model. Multiple sources of data can provide evidence for the scoring inference, including reviewing scoring distributions of scored samples, conducting reliability studies, and confirmatory factor-analysis. For the RESET observation protocol, this would include engaging in activities such as reviewing rater consistency with master coded observations, and examining score distributions across samples – especially in light of evidence suggesting that many teacher observation systems result in "the widget effects" (i.e., by evaluating all teachers as above average, all teacher performances lose variation and become interchangeable) (Mead, Rotherham, & Brown, 2012; Weisberg, Sexton, Mulhern, & Keeling, 2009).

Generalization: Observation systems require users to be able to generalize from a limited domain of observations to the universe of all possible observations of that teacher's practice. Therefore, assumptions about generalization include that 1) the sample adequately represents the universe of all possible observations, and 2) unexpected error is accounted for. Generalizability studies and reliability indices provide empirical support for the generalizability of scores across raters and over samples of test items (Kane, 2013). Most observation systems will include multiple sources of variance, and G-studies can provide estimates of variance components associated with the universe of possible observations (Brennan, 2001). In an ideal scenario, the main source of variance in an observation system would be the different teachers being observed.

However, variance from raters, lessons, items and interactions among these factors can account for variance in the observed scores, and it is critical to understand the contextual factors that influence performance. If we do not understand the degree to which contextual factors shape the scores received during observation, it will be difficult to justify resulting decisions (Bell et al., 2012).

Extrapolation: Tests are used to evaluate how well people can perform certain activities over some range of conditions. For practical reasons, it is generally not feasible to employ samples of the performance of interest under all possible conditions, or when measuring more complex constructs, to represent the full range of tasks that comprise the larger construct. In the case of special education teaching for example, it is extremely challenging to consistently quantify student outcomes, yet improved outcomes for students with disabilities is a reasonable indication of teacher effectiveness. RESET is premised on the idea that if special education teachers can reliably implement EBP, they should realize gains in student outcomes consistent with effect sizes reported in the research. Therefore, an extrapolation inference could be examined through data examining student outcomes (in terms of effect sizes to account for the variety of individualized student goals and measures) achieved when EBP are employed.

Another indicator of effective special education teaching is the development of Individualized Education Plans that outline the instructional practices and individual goals for students. It would be time consuming to include a comprehensive review of Individualized Education Plans (IEPs) within an evaluation system. The assumption of RESET is that if a special education teacher is effectively delivering EBP, the IEPs of the students will reflect goals and methods consistent with relevant EBPs. In this case, the extrapolation inference could be validated through a review of IEPs and analysis of the relationship between IEPs and implementation of EBP. The specific assumptions underlying the extrapolation inference include that 1) the score on all lessons is related to the teaching quality special education teachers are able to enact, and 2) there are not systematic errors that undermine the extrapolation to overall teaching quality (Kane, 2013).

Decision rules: The IUA for decisions will involve a chain of inferences that begins with the observed performances which lead to statements about a trait, and then to decisions based on the estimated values (scores) of that trait. In other words, based on the score assigned to the special education teacher observation, we are making claims about a teacher's effectiveness. Decision rules related to special education teacher evaluation might include decisions about promoting, retaining or mentoring a special education teacher. Decision rules are therefore, typically evaluated in terms of both their expected and unexpected consequences. As Herlihy et al. (2014) note however, this is the least well-developed aspect of most teacher evaluation systems nationally. Assumptions underlying the decision rules include that the decisions associated with the observed scores and performances are appropriate, and that the properties of the observed scores support the implications associated with the judgments of teaching performance.

The Validity Argument

Once the IUA for an observation protocol is developed, the validity argument evaluates those inferences and assumptions using empirical data and analytic reasoning. Different kinds of inferences will require different kinds of analysis for their evaluation. The IUA provides a framework for identifying the empirical studies to be included in the validation process, and then

the validation process provides a critical appraisal of the IUA's coherence and plausibility, with the understanding that new evidence could lead to a reconsideration of this conclusion (Kane, 2013).

Observation systems for special education teacher evaluation hold significant promise as a means for identifying special education teachers who are effective, and for improving the instructional practice of special education teachers. However, prior to their implementation, it is critical to understand whether the inferences made from the scores assigned to an observation constitute a valid statement about the quality of a special education teacher. Drawing on Kane's argument-based approach to validity, and its application to observation protocols (Bell et al., 2012), the purpose of the current study was to establish the IUA and evaluate initial empirical evidence to examine the scoring and generalization inferences for the RESET observation protocol. In our application of Kane's argument based approach to RESET, we viewed establishing validity of the scoring and generalization inferences as prerequisites to examining extrapolation and decision inferences. Without consistent scoring procedures that generalize across observations, it would be premature to examine evidence regarding RESETs extrapolation to other areas of teacher quality or to examine the consequences of the decisions and resulting implications. After describing and reporting the results of our examination of evidence of the scoring and generalization inferences, we outline next steps based on those findings, as well as describe next steps for examining the extrapolation and decision inferences.

Methods

Validation of the IUA requires multiple approaches to data collection and evaluation. All of the analyses conducted in this study were based on the video observations of special education teachers, and the ratings assigned to the video observations by trained mentor special education teachers. In this section, we describe the special education teacher participants who provided the video taped lessons, the special education teacher raters who evaluated the lessons, and the RESET observation protocol. Because each of the IUA inferences requires multiple types of analysis, the analyses used for the validity argument are included in the presentation of results.

Participants

Special Education Teacher Participants. A total of 19 special education teachers from five districts contributed a total of 4,082 minutes of video taped-lessons across a variety of special education settings. To recruit special education teacher participants, we contacted the special education directors of five of the larger school districts in the state. In our request for participants, we asked special education directors to help identify special education teachers who were highly skilled, as well as novice special education teachers, so that we would be sure to have exemplar instructional videos across the range of possible scores on RESET. Once a special education teacher agreed to participate, we also received consent from the parents of students in the classroom to video record lessons. All of the participating special education teachers were Caucasian females, ranging in experience in teaching from 1 to 15 years, with a mean of 9.5 years of experience. 28% held graduate degrees in education. Each teacher had a minimum of five lessons captured.

Raters. Five special education teachers were invited to participate as raters in two sessions to evaluate the videos of special education classroom instruction collected from the 2011-12 and 2012-13 school years via the Teachscape 360-degree video system. Raters were selected through communication with special education directors. Predetermined criteria were observed to ensure that invited raters represented a balanced sample of the range of content, placement and grade level found in special education, and that the invited raters had a minimum of five years of teaching experience. Table 1 provides rater demographics, including current teaching assignment, total years teaching and highest level of education completed. Additionally, the lead author scored each of the video recorded lessons individually to develop a master coded set of scores. The master coding served as a benchmark against which consistency and inter-rater reliability were evaluated for participating raters.

Table 1

Rater Teaching Background and Experience

Rater	Teaching Assignment	Years of Experience	Education
1	Secondary Resource	10+	Master's
2	Elementary EBD/Self-contained	30+	Master's
3	Secondary Resource	3	Master's
4	Elementary Resource	15+	Bachelor's
5	University Teacher Supervisor	3	Master's

Measures

RESET Observation Protocol. The RESET observation protocol is a special education teacher evaluation system guided by the idea that the increased use of evidence-based instructional practices will lead to increases in student outcomes. RESET is comprised of three subscales: 1) Lesson Objective (LO), 2) EBP Implementation (EBP), and 3) Whole Lesson Summary (WL). Each item for each subscale is scored on a 1-4 scale in order to align with the Danielson (2013) scoring system. The LO subscale determines the clarity of the lesson objective, and consists of three items. The EBP subscale consists of rubrics that were developed using the criteria and key characteristics of various evidence-based instructional practices identified in the existing special education literature (Cook & Odom, 2013; Gersten et al., 2005; Horner et al., 2005; Odom et al., 2005). Most of the EBP rubrics contain between 4-6 criteria, each assigned a score on the 1-4 scale. The WL subscale consists of three items designed to provide a broad evaluative score of the special education teacher's performance throughout the lesson. Inter-rater agreement for RESET has ranged from .72 to .95, with a median agreement of .85 (Semmelroth & Johnson, 2014). Generalizability studies examining sources of variance for RESET have resulted in promising G-coefficients, ranging from .79 - .86 (Johnson & Semmelroth, 2014b).

Procedures

Video recorded lessons were collected of the 19 participating special education teachers during the 2011-12 and 2012-13 school years. Using the Teachscape video capture system, a total of 4,082 minutes of instruction was captured from the participating teachers over a minimum of

five lessons. Each rater scored each lesson. This was done to mitigate potential bias by assigning specific raters to particular teachers, and in order to help identify the optimal number of raters to receive acceptable levels of reliability. Raters attended a one-day training on scoring RESET. Training consisted of explaining the purpose and design of RESET, and orienting raters to the 45-page user manual that explains the structure and scoring procedures for RESET. Then, raters individually evaluated two training videos. The results of the training videos were compiled, and disagreements with the master-coded scores were discussed as a group to reach consensus on scores. Interrater agreement achieved during the training sessions ranged from .72 to .95 across subscales.

Raters then evaluated each video in random order. The order of videos was randomly presented to mitigate the possibility of a teacher x rater interaction (e.g. from viewing the same teacher five times in a row), and to reduce the possibility of an order effect. Raters completed their evaluations over a three-day period in a designated coding area. The authors were available to answer questions and to help resolve any technical issues. Scoring was input using the Qualtrics data system. Raters entered their scores for each video, and then the data was exported from Qualtrics to a database for analysis. Each video recorded lesson was assigned a unique identification number that allowed us to connect observations of the same teacher. Each rater's score for each item for each observation were collected in the database for analysis, along with scores for each lesson assigned by a master coder. A variety of analyses were performed for each inference of the IUA, and are discussed in the following section.

Results

Scoring Inference

Observation systems rely on the assignment of scores as an indication of the quality of the observed performance. The assumptions about the scoring system of RESET include that: 1) the scoring rule is appropriate; 2) the scoring rule is applied accurately and consistently; 3) the scoring is bias free; and 4) the data fit the scoring model. Below we discuss the analysis and the subsequent results of examining each of these assumptions using this data set.

Appropriateness of scoring rules. To determine the appropriateness of the scoring rules, we examined the score distribution across all evaluated observations (n = 216). Figure 1 displays the scoring distribution for each item of RESET, which depicts a high number of items clustered around the lowest scores, and in some cases, none of the items receiving a score above 2. Upon first review, this was considered problematic because of the lack of representation of all possible points across the scoring rubric. This does not necessarily undermine the appropriateness of the scoring inference because it might be reasonable for actual practice to be clustered around particular score points (Bell et al., 2012). Converging evidence to ensure that the skewed scores reflect actual instructional quality and not scoring error was found by comparing the raters' distribution of scores to the distribution of scores obtained by the master rater (Figure 2), which also clustered around these two scores. This suggests that the quality of observed performances of the sample was generally low overall.

Figure 1. Scoring distribution for RESET items, n = 216 evaluations

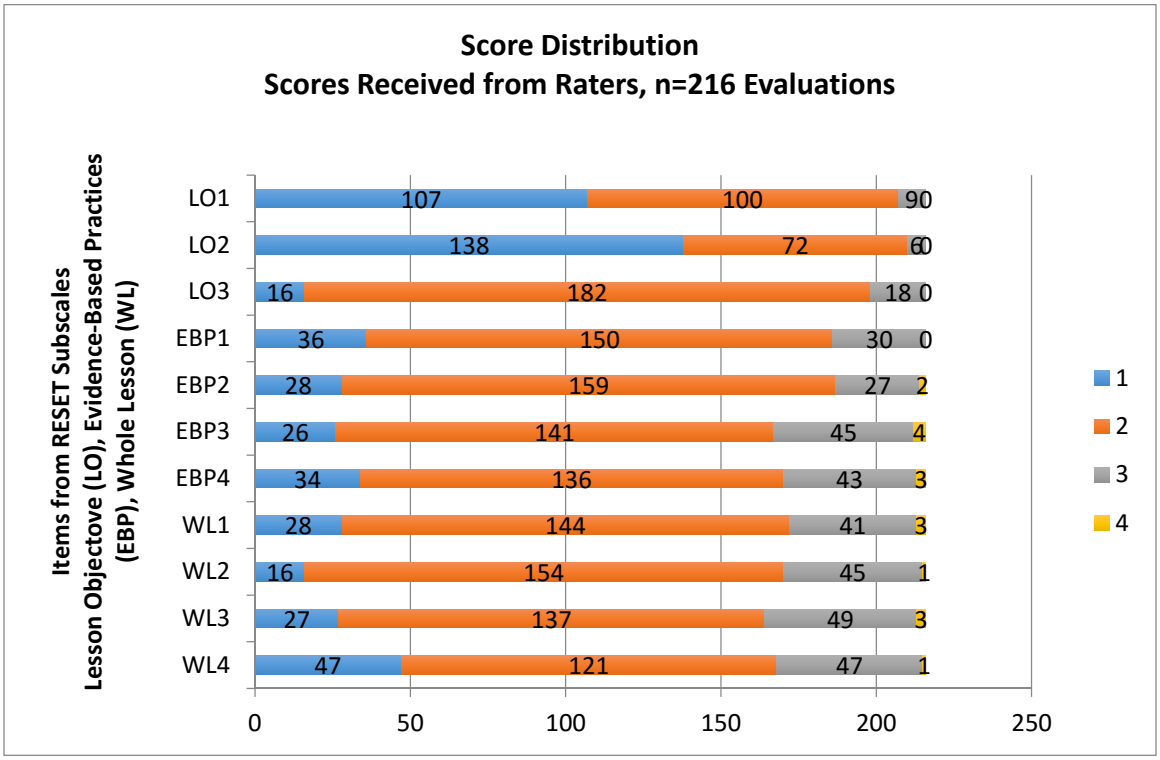
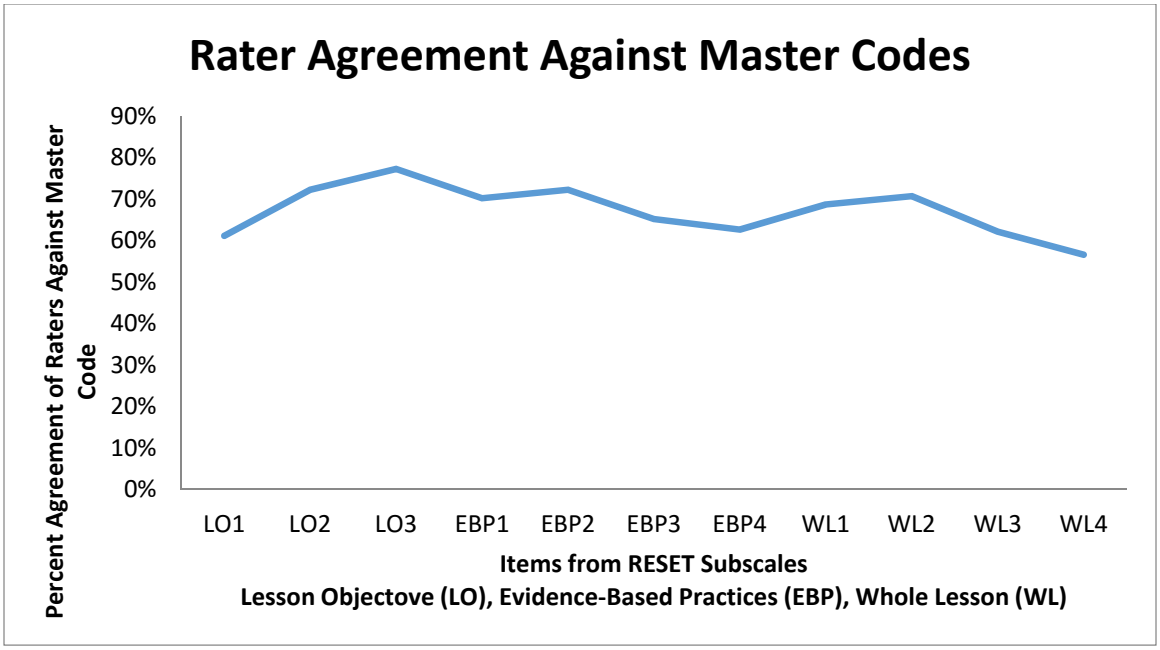


Figure 2. Rater agreement against master codes, n = 216 evaluations



To further assess the appropriateness of the scoring rules, we ran a correlation analysis to examine correlations among items within subscales. These results are presented in Table 2. In general, items belonging to the same subscale were more correlated to one another than they were to items belonging to other subscales. Correlations of items within subscales were statistically significant and moderate, ranging from .42 - .75, with the strongest correlations in subscale 3, Whole Lesson Summary.

Table 2

Correlations of Items from RESET Observation Tool Subscale

	LO1	LO2	LO3	EBP1	EBP2	EBP3	EBP4	WL1	WL2	WL3	WL4
LO1	1	.49	.42	.28	.23	.38	.22	.19	.21	.28	.26
LO2		1	.46	.35	.25	.34	.27	.29	.27	.26	.32
LO3			1	.38	.39	.39	.35	.38	.35	.39	.35
EBP1				1	.59	.51	.53	.56	.51	.59	.61
EBP2					1	.52	.64	.62	.46	.60	.52
EBP3						1	.55	.52	.54	.49	.56
EBP4							1	.58	.55	.60	.66
WL1								1	.75	.70	.67
WL2									1	.66	.61
WL3										1	.71
WL4											1

All correlations were significant at the $p < .005$ level (2-tailed)
Note: Lesson Objective (LO), Evidence-Based Practices (EBP), Whole Lesson (WL)

Accurate and consistent application of scoring rules. To examine the accuracy of raters, we compared their scores against a master code (Figure 2). Using the master code, agreement rates varied between 57-77% for all scores, with the highest level of agreement for the LO subscale. To measure consistency, we reviewed the results of a generalizability theory analysis to examine the sources of variance for the scores obtained on the observations (Brennan, 2001; Shavelson & Webb, 1991). In our g-study, potential sources of variance included teachers, occasions within teachers, raters, items and interactions among these variables. To evaluate the consistency of application of scoring rules, we examined the variance attributed to raters, and found that it was low, accounting from between 2.5 – 8.2% of the variance. These results suggest that the scoring rule is applied accurately and consistently across raters.

The scoring is bias-free. There are two main sources of potential bias for the scoring of RESET: 1) the assignment of raters to teachers, and 2) the rater assignment of scores. All raters evaluated all teachers and videos, thus addressing the first potential source of bias. Our G-study provides some data to make a determination regarding the second source of bias, rater assignment of scores (Table 3). First, the amount of variance accounted for by Item x Rater interaction was very low across subscales. Additionally, there was a low percentage of variance accounted for by teacher x item x rater interactions. However, the amount of variance accounted for by a teacher x rater interaction was of some concern, ranging from 8.2 – 15.5% across the three subscales.

Table 3

Describing Variance Components for a I x R x (O:T) Generalizability Study

Source	Description	Lesson Objective	Evidence-Based Practice	Whole Lesson Summary
Teacher (T)	"True score" variance	9.5%	16.7%	15.8%
Item (I)	Some items are more difficult than others.	17.2%	0.1%	0%
Rater (R)	Some raters score more critically than others.	2.6%	5.6%	8.2%
Occasions (O:T)	Confounded with teacher score dependence on lessons.	4.1%	6%	1.7%
T x I	Some teachers score higher on certain items.	1.4%	4.6%	1.4%
T x R	Some raters score certain teachers higher.	8.2%	10.9%	15.5%
I x R	Some raters score certain items higher.	4.8%	0%	0.6%
T x I x R	Some raters score higher certain teachers on certain items.	2.4%	2.9%	0.9%
I x (O:T)	Some items receive higher scores on certain lessons. Confounded with teacher score dependence.	1.3%	0.9%	0.8%

(O:T) x R	Some raters score certain lessons higher. Confounded with teacher score dependence.	20.2%	21.3%	18.9%
Residual (O:T) x I x R, e	Error variance	28.2%	31%	36.3%
G-Coefficients	Relative (Absolute)	.79 (.77)	.84 (.81)	.86 (.82)

The data fit the scoring model. Confirmatory factor analyses (Table 4) were conducted on scores within each of the subscales. The results showed a two-factor solution best fit the scoring model (chi squared = 1268.93, $p < .000$), where the LO subscale constituted one factor, and the WL and EBP subscales loaded on the other. This suggests that the extent to which the teacher implements EBP is strongly related to the overall evaluation of the whole lesson, which is a reasonable finding, and consistent with the conceptual framework used to develop RESET.

Table 4

Results of a Confirmatory Factor Analysis of the Three Subscales of RESET

	Factor 1	*Factor 2*
W1	.85	.14
W3	.83	.18
W4	.81	.20
EBP4	.79	.16
W2	.79	.15
EBP2	.75	.19
EBP1	.71	.29
EBP3	.63	.39
LO1		.83
LO2	.16	.80
LO3	.32	.68

chi squared = 1268.93, $p < .000$

Note: Lesson Objective (LO), Evidence-Based Practices (EBP), Whole Lesson (WL)

Generalization Inference

For the generalization inference, it is important to determine the extent to which the observed performance is representative of all areas to which we wish to generalize. Two main assumptions about the generalization inference must be tested: 1) the sample adequately represents the universe of all possible observations, and 2) unexpected error is accounted for. To examine evidence for the generalization inference, we used a G-study approach (Brennan, 2001; Shavelson & Webb, 1991) to estimate the sources of variance in RESET. Table 4 reports the results of our G-study and the relative g-coefficient, which range from .79 - .86 across the three subscales. The g-coefficient is an indication to which we can conclude that the results are generalizable to the population of all elements that could have been used to develop the measurement instrument. This conclusion is generally determined reasonable if the reliability coefficient is at least .80 (Cardinet, Johnson, & Pini, 2010; Shavelson & Webb, 1991).

The largest sources of variance included the teacher being observed, the interaction of occasion (or lesson) with the rater, and the residual or error variance. Residual error is unexplained error, which in our results was the largest source of variation across all three subscales. Substantial residual error suggest that multiple observations will be needed to generalize with any degree of accuracy from the observations to general statements about a special education teacher's effectiveness. Interactions of teachers with raters constituted the next highest source of variance, which suggests that either more training, more calibration or more precise scoring rules or a

combination of the three are needed to reduce the interaction. Finally, the observed teachers were the third largest source of variance. Ideally, they would constitute the largest source of variance. Variance due to lessons was low, ranging from 1.7 to 6%, indicating that scores assigned to teachers generalize across lessons. All of these results however, must be interpreted cautiously, given the restriction of range of scores in the data set.

Discussion

Given the rapid pace at which teacher evaluation systems are being adopted and used to make high stakes decisions, it is critical to ensure that these measures are psychometrically defensible. Comprehensive approaches to establishing validity are needed, especially if observation scores are to be used for high-stakes decisions regarding special education teachers' promotion, retention and professional development. Following Bell et al.'s (2012) application of Kane's (2006) validity argument approach to observation protocols, the purpose of this study was to examine initial evidence regarding the scoring and generalization inferences of a special education observation protocol grounded in the evaluation of teachers' use of evidence-based instructional practices. Empirical examination of the scoring and generalization inferences was viewed as the first step in the process of validation of a special education teacher observation system. Our rationale for beginning with these two inferences is that a consistent scoring system that generalizes to the universe of observations is a necessary precursor to examining broader implications of a measure's use. In this section, we discuss the findings from our analysis in the context of current research on teacher observation systems, and outline potential next steps in collecting validity evidence for the extrapolation and decision IUAs for RESET.

Application of IUA Inferences

The RESET observation protocol was developed to provide a means by which special education teachers would receive feedback on their use of EBP within an evaluation system aligned with Danielson's FFT. To accomplish this, we focused on the instructional domain of the FFT framework, developed rubrics that explicate the components of a variety of EBP, and developed a scoring scheme consistent with the Danielson model. We also relied on the use of special education teachers as raters following the recommendations of Holdheide et al. (2012), to integrate the use of peer reviewers who have the appropriate qualifications and experience to make accurate judgments about teacher performance. Initial evidence shows that we can draw on what we know about EBP to develop an observation protocol with sufficient flexibility to evaluate special education teachers across a variety of contexts. The audit of the scores assigned to this sample indicated that we were able to achieve acceptable levels of consistency and agreement with master coded evaluations. This was an encouraging finding because one of the concerns about developing special education teacher evaluation systems is that they should not be separate or parallel to the general education teacher evaluation system, but rather, should be relevant for special education teachers but consistent with the overall framework in use for general education teachers (Holdheide, Hayes, & Goe, 2013).

However, our review of the scoring distribution was less encouraging. In this sample, the majority of scores received were on the low end of the scoring distribution. This finding was of particular concern because in our recruitment of participating special education teachers, we worked with special education directors to try to recruit a broad sampling of special education

teachers to include both highly skilled and novice teachers. The sample of observations that comprised our data set reflected an overall low quality of special education consistent with findings reported in classroom observation studies indicating that children do not always receive special education services that can reasonably be expected to mitigate the effects of their disabilities (Morgan, Frisco, Farkas, & Hibel, 2008). We interpret this finding in two ways. First, this finding suggests that future validation efforts will need to include a broader sampling process to help ensure a representation of all possible score values. Alternatively, the low distribution of scores validates one of the primary purposes of RESET, which was to design a system that would improve the instructional practice of special education teachers by drawing attention to the use of EBP and providing specific feedback to teachers. That so many special education teachers scored so poorly suggests that this evaluation system is needed. If it is the case that many special education teachers are performing on the lower end of the spectrum, we may need to develop a scoring scale that is more sensitive to differentiating across the lower levels of performance in order to help document growth and to provide finer-grain feedback to teachers for improvement.

The results of our CFA also provided useful insights for the continued development of RESET. RESET includes three subscales: a) lesson objective (LO), b) evidence-based practice (EBP), and c) whole lesson (WL) evaluation. The purpose of the LO subscale is to evaluate the extent to which special education teachers are able to communicate to their students the goal of the lesson, and to align instructional practice with that objective. The EBP subscale provides the evaluation of the use of evidence-based instructional practices. The WL subscale was developed to determine whether an overall rating of teacher performance during the lesson would correlate with the finer-grain scores on the other subscales. We wanted to test if higher levels of reliability on evaluations of teacher performance are achieved when based on overall evaluative judgments of teacher practice rather than scoring the components of instructional practice.

The results of our CFA indicated a two-factor solution best fit the data, with the EBP and WL subscales loading on a common factor. Initially, we were disappointed in this finding, and questioned whether the WL subscale might be redundant. However, as we examined our findings further in the context of next steps for implementation, we believe that the common loading of these two subscales may help to solve an important concern of implementing RESET in practice. Our analyses to date have been conducted using experienced special education teachers as raters. In the early stages of development, we felt this was important because we did not want to introduce scoring error due to a rater's lack of knowledge about EBP in special education. Evaluations conducted by trained peer evaluators are believed to enhance the credibility of the evaluation process and to provide valuable feedback to improve performance (Holdheide et al., 2013), yet school administrators often want to evaluate staff for whom they have responsibility of supervising. Future studies that examine the consistency of scoring on the WL subscale when evaluated by administrators compared to scoring when evaluated by experienced special education teachers may inform how to include administrators in the evaluation process while ensuring that special education teachers receive expert feedback on their instructional practice. This strategy can leverage the expertise of special education raters while creating a culture of continued learning and collaboration between administrators and district special education staff (Holdheide et al., 2013). If results on the WL subscale are consistent across raters, administrators, who have limited time and expertise to do in-depth evaluations of special

education teacher instruction, could evaluate teachers using the WL subscale only. Then, mentor teachers or special education directors could provide the more detailed evaluation and feedback on the components of instructional practice to special education teachers.

Generalization Inference and RESET Project Work. A critical assumption of any observation protocol is that we can generalize inferences about performance from a small set of observations to the universe of all possible observations (Kane, 2006). Although the g coefficients achieved in this study were in the acceptable range of > .80 (Shavelson & Webb, 1991), the distribution of variance was of some concern. In an ideal observation system, the primary source of variance would be the teachers being observed, with limited variance due to raters, observations, interactions of these factors, and residual error. In this data set, variance due to teachers was the third largest source of variance, but only accounted for between 10 to 17%. Upon further examination of the score distribution however, this may be due to the limited distribution of scores. Further studies that include a broader distribution of scores are needed. The residual, or error, in this study indicated that between 28 to 36% of the variance comes from contextual factors currently unaccounted for. This finding is consistent with results of observation protocol analyses reported elsewhere and leads us to draw conclusions similar to those reported by Bell et al. (2012). Further investigation of the contextual factors that affect teacher effectiveness is needed to ensure that we can justify the decisions made as a result of our observation protocols.

In earlier work on RESET, we found that four observations evaluated by four raters is optimal for achieving acceptable levels of reliability (Johnson & Semmelroth, 2013; Semmelroth & Johnson, 2014). This is consistent with results reported by other teacher evaluation systems (Bell et al., 2012; Hill, Charalambous, & Kraft, 2012; Ho & Kane, 2013; Kane & Staiger, 2012) that multiple observations are needed to achieve acceptable levels of reliability about a teacher's performance. Given the logistical challenges of implementing an evaluation system that requires multiple observations per teacher however, the appeal of more direct models, such as VAM are clear. It is significantly more expedient to determine teacher effectiveness based on student performance on standardized assessments. However, the challenges of applying VAM models not only in general, but especially to special education teachers are well documented (Baker et al., 2010; Buzick & Laitusis, 2010; Holdheide et al., 2012; Johnson & Semmelroth, 2014a; Jones & Brownell, 2014). And, as we have argued elsewhere, given the challenges faced in the special education teaching profession, a focus on improving instructional practice is absolutely necessary if we are to improve outcomes for students with disabilities (Johnson & Semmelroth, 2012, 2014a; Semmelroth, Johnson, & Allred, 2013).

Implications of Extrapolation and Decision Inferences. The underlying conceptual framework of RESET is that if special education teachers routinely employ evidence-based instructional practices in their teaching, outcomes for students with disabilities will improve.
Teacher evaluation models based primarily on student outcomes are insufficient for the evaluation of special education teachers because of the complexities of determining the percentage of their contribution to student outcomes, and because the performance of students with disabilities is typically assessed using a variety of outcomes (Johnson & Semmelroth, 2014a). Additionally, special education teachers enter the field without adequate preparation to implement evidence-based practice with fidelity therefore, an evaluation tool will need to focus on evidence-based instructional practice as a way to increase the effectiveness of special

education teachers (Johnson & Semmelroth, 2014a). Finally, measures of instructional practice should correlate highly with measures of student growth, and this is the main premise upon which RESET is based. Through the use of research-based instructional practice, students with disabilities should realize levels of growth consistent with those reported in the research. In other words, high levels of fidelity of implementation of an instructional practice should correspond with levels of student growth commensurate with those reported in the research. Next steps for validating the use of RESET include testing these underlying assumptions of the conceptual framework. This will be accomplished by collecting data on student performance and linking measures of growth to the evaluation of the teachers' use of EBP. Our hypothesis is that teachers who are more adept at implementing relevant instructional practices will help their students realize growth in performance consistent with the effect sizes reported in the research.

Although the correlation between fidelity of implementation and student growth seems intuitive, Kane and Staiger (2012) reported small correlations (.19) between VAM and FFT. There are several plausible explanations for the low correlations, one of which is the restriction of range in Kane and Staiger's (2012) data set because of the distribution of scores on FFT. Disproportionate numbers of teachers were rated as proficient or distinguished on most items on FFT, suggesting that evaluations of teacher practice did not reliably discriminate among those who were skilled versus those who were not. Another plausible explanation is that the two measures tap such different elements of teaching, that measures of both are needed. Although RESET is designed to have a more direct alignment between instructional practice and student outcome, the results reported by Kane and Staiger (2012) suggest that further investigation of the relationship between instructional implementation and student growth is critical. In continuing the validation process, studies that examine the relationship of high levels of implementation of EBP and the growth that students with disabilities are able to achieve will be a significant component of establishing the psychometric defensibility of RESET. Finally, the underlying assumption of RESET is that attention must be drawn to the use of EBP in order for teachers to improve practice. Validation studies that examine the effect on teacher practice over time will establish whether the decisions made based on RESET support this assumption.

Conclusions

Teacher evaluation systems are being used to make high-stakes decisions about teacher performance, retention and pay, yet few systems have been examined to determine their psychometric defensibility to warrant these decisions (Herlihy et al., 2014) especially those developed for special education teachers. This paper described one model of special education teacher evaluation and examined initial evidence to determine its reliability and validity. The results are promising, but significantly more work is needed to develop a system that is both useful and fair. If we are to be successful in improving the practice of special education teachers, we will need to ensure that our evaluation systems:1) reliably discriminate between effective and ineffective special education teachers, 2) measure and provide targeted, specific, corrective feedback for teacher instructional practice, and 3) include the use of individualized student growth rates to define teacher effectiveness. Most importantly, we must ensure that our evaluation system leads to sound decisions regarding instructional practice, and ultimately services provided to students with disabilities.

References

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., … Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. In *Economic Policy* (Vol. 278, pp. 1–29). Economic Policy Institute. Retrieved from http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62–87. doi:10.1080/10627197.2012.715014

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

Browder, D., Ahlgrim-Delzell, L., Spooner, F., Mims, P. J., & Baker, J. N. (2009). Using time delay to teach literacy to students with severe developmental disabilities. *Exceptional Children*, *75*(3), 343–364.

Browder, D. M., & Cooper-Duffy, K. (2003). Evidence-based practices for students with severe disabilities and the requirement for accountability in "no child left behind." *The Journal of Special Education*, *37*(3), 157–163.

Buzick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher*, *39*(7), 537–544. doi:10.3102/0013189X10383560

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

Chard, D. J., Ketterlin-Geller, L. R., Baker, S. K., Doabler, C., & Apichatabutra, C. (2009). Repeated reading interventions for students with learning disabilities: Status of the evidence. *Exceptional Children*, *75*(3), 263–281.

Connelly, V., & Graham, S. (2009). Student teaching and teacher attrition in special education. *Teacher Education and Special Education*, *32*(3), 257–269. doi:10.1177/0888406409339472

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, *79*(2), 135–144.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, *75*(3), 365–383.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 Edition* (2nd ed.). Princeton, NJ: Danielson Group.

Fuchs, L. S., & Fuchs, D. (2005). Enhancing mathematical problem solving for students with disabilities. *The Journal of Special Education*, *39*(1), 45–57.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Murphy, 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., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*(2), 149–164.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). *State and Local Efforts to Investigate the Validity and Reliability of Scores from Teacher Evaluation Systems*.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. doi:10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Retrieved from http://www.metproject.org/downloads/MET_Reliability of Classroom Observations_Research Paper.pdf

Holdheide, L., Browder, D., Warren, S., Buzick, H., & Jones, N. (2012). *Summary of "using student growth to evaluate educators of students with disabilities: Issues, challenges, and next steps"* (pp. 1–36). Retrieved from http://www.gtlcenter.org/sites/default/files/docs/TQ_Forum_SummaryUsing_Student_Growth.pdf

Holdheide, L., Hayes, L., & Goe, L. (2013). *Evaluating specialized instructional support personnel supplement to the practical guide to designing comprehensive teacher evaluation systems*. This needs the GTL info, so It hink something like, Great Teachers and Leaders Center: City, ST (same with the one above)Retrieved from http://www.gtlcenter.org/tools-publications/publications

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–180.

Johnson, E. S., & Semmelroth, C. L. (2012). Examining interrater agreement analyses of a pilot special education observation tool. *Journal of Special Education Apprenticeship*, *1*(4). Retrieved from http://josea.info/index.php?page=vol1no2

Johnson, E. S., & Semmelroth, C. L. (2013). Sources of Variance in a Special Education Observation Tool. In *Pacific Coast Research Conference*. Coronado, CA this isn't the right way to cite a poster - please check the APA manual .

Johnson, E. S., & Semmelroth, C. L. (2014a). Special education teacher evaluation: Why it matters and what makes it challenging. *Assessment for Effective Intervention*, *39*(2) need page numbers.

Johnson, E. S., & Semmelroth, C. L. (2014b). Validating an Observation Tool to Measure Teacher Effectiveness. In *Pacific Coast Research Conference*. this isn't the right way to cite thisCoronado, CA.

Jones, N. D., & Brownell, M. T. (2014). Examining the Use of Classroom Observations in the Evaluation of Special Education Teachers. *Assessment for Effective Intervention*, *39*(2), 112–124. doi:10.1177/1534508413514103

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M. T. (2013). The Argument-Based Approach to Validation. *Social Psychology Review*, *42*(4), 448–457.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (pp. 1–68). Need project name & City, State Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

McLeskey, J., Tyler, N. C., & Flippin, S. S. (2004). The supply of and demand for special education teachers : A review of research regarding the chronic shortage of special education teachers. *The Journal of Special Education*, *38*(1), 5–21.

Mead, S., Rotherham, A., & Brown, R. (2012). *The hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge*.need a publication source

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2008). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*(4), 236–254. doi:10.1177/0022466908323007

National Autism Center. (2009). *National standards report*. Randolph, Massachusetts. Retrieved from http://www.nationalautismcenter.org/nsp/reports.php

Odom, S. L. (2009). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education*, *29*(1), 53–61.

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

Semmelroth, C. L. (2013). *Using generalizability theory to measure sources of variance on a special education teacher observation tool*. Boise State University is this the right way to cite a dissertation.

Semmelroth, C. L., & Johnson, E. S. (2014). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention*, *39*(2) need page numbers.

Semmelroth, C. L., Johnson, E. S., & Allred, K. (2013). Special educator evaluation: Cautions, concerns and considerations. *Journal of the American Academy of Special Education Professionals need page numbers or indication that it is online*.

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

Spooner, F., Knight, V. F., Browder, D. M., & Smith, B. R. (2012). Evidence-based practice for teaching academics to students with severe developmental disabilities. *Remedial and Special Education*, *33*(6), 374–387. doi:10.1177/0741932511421634

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. publication source? Retrieved from widgeteffect.org/downloads/TheWidgetEffect.pdf

About the Authors

Dr. Johnson is a professor of Special Education at Boise State University and the Executive director of Lee Pesky Learning Center, a non-profit organization whose mission is to improve the lives of people with learning disabilities. Dr. Johnson's research focuses on special education teacher evaluation, interventions for students with learning disabilities and improving the way we identify students with learning disabilities.

Dr. Semmelroth is a lecturer at Boise State University. Her research has focused on special education teacher evaluation.