JOSEA
JOURNAL OF SPECIAL EDUCATION APPRENTICESHIP

Vol. 1. No. 2 Nov 20. 2012 ISSN 2167-3454

Examining Interrater Agreement Analyses of a Pilot Special Education Observation Tool

Evelyn S. Johnson and Carrie L. Semmelroth, Boise State University

This paper reports the results of interrater agreement analyses on a pilot special education teacher evaluation instrument, the Recognizing Effective Special Education Teachers (RESET) Observation Tool (OT). Using evidence-based instructional practices as the basis for the evaluation, the RESET OT is designed for the spectrum of different instructional needs found within special education classrooms. The RESET OT informs what Danielson (2011) maintains are the two features of a teacher evaluation system 1) ensuring teacher quality and 2) promoting professional development. In June 2012, six special education teachers participated in a data coding session using the pilot RESET OT to evaluate video observations of special education instructional practice from the 2011-2012 school year. The teacher coders received an introductory training session to the RESET OT, and participated in two whole-group coding sessions before completing individual coding assignments. The results of the interrater agreement analysis report weak to no agreement within specific instructional practices, indicating the need for 1) additional research and development on the RESET OT 2) repeating the data coding session using a group of teacher coders who have received in-depth training on the RESET OT and evidence-based instructional practices, and 3) further investigation into the specific components of evidence-based instructional practice and how these might be applied across settings.

Keywords: special education teacher effectiveness, special education teacher evaluation, value-added model, evidence-based instructional practice, interrater reliability

Although teacher evaluation systems can be designed to improve practice, increase teacher capacity, and to identify teacher effectiveness, the National Comprehensive Center for Teacher Quality (TQ Center) suggests the ultimate goal of any teacher evaluation system is simply to improve teaching and learning (Holdheide, 2012). Similarly, Danielson (2011) maintains that the two primary features of an effective teacher evaluation system are to 1) ensure teacher quality and 2) promote professional development. However, within the past three years, 32 states have changed their policies regarding teacher evaluation, and of those, 20 states have focused heavily on using student achievement as a primary component of the teacher evaluation system (National Council on Teacher Quality, 2011). These states are now faced with resolving the purposes of teacher evaluation systems with the end goals of improving student achievement scores. States are now faced with the methodological, measurement and implementation challenges related to building these new teacher evaluation systems, and arguably, no other content area exemplifies the difficulties of designing a comprehensive, reliable, fair and efficient teacher evaluation system more than special education (Holdheide, Goe, Croft, & Reschly, 2010).

The most well-known approach for incorporating student outcomes as a primary feature of teacher effectiveness is the value-added model (VAM) (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010). VAMs define a relationship between teacher effectiveness and student academic achievement through weighted statistical formulas that incorporate values from a variety of measurements including teacher observation scores, student achievement scores, student/parent surveys, and other factors (Kane & Staiger, 2012). VAMs attempt to account for the multiple factors that may impact student achievement (Scherer, 2012), and are thought to help answer the question of how effective an individual teacher is at promoting student growth. Critics argue that VAMs suffer from numerous methodological and philosophical flaws, and can be influenced by variables outside of the teacher's abilities and control. For example, Newton, Darling-Hammond, Haertel, & Thomas (2010) found that ratings of teacher effectiveness varied significantly based on student demographics, and that teachers of students who are disadvantaged or in an at-risk category tended to have systematically lower ratings. This is presumably due to the emphasis on student outcomes in these models. Newton et al. (2010) argue as a result of the inconsistency in various models to rate teachers, VAMs should not be tied to high-stakes decisions about teacher performance.

The use of VAMs for special education teachers is especially concerning because 1) the small number of students in many special education teachers' classrooms can result in less reliable estimates of teachers' effects on student performance, 2) alternate assessments can preclude the use of value-added modeling for some teachers, 3) the inconsistent use of accommodations across years on state standardized tests can impact the measurement of growth and, consequently, the accuracy and meaning of value-added scores if they are not accounted for in the model, and 4) the mobility of some students with disabilities—and the subsequent omission of their test scores in value-added models—may preclude efforts to provide value-added scores for some teachers (Holdheide, Browder, Warren, Buzick, & Jones, 2012). Additionally, VAMs rely on the use of student outcomes and achievement scores to evaluate teacher efficacy, but standardizing these scores can be problematic for special education. Growth rates for students with disabilities are typically not consistent, and there is evidence that suggests students with very low initial performances often experience the least growth even when exposed to evidence-based instruction (Coyne et al., 2010; Wei, Blackorby, & Schiller, 2011).

Using student outcomes to define special education teacher effectiveness requires first being able to identify 1) what kind of student growth measure to use and 2) how much student growth to expect. There are clear measurement challenges to addressing both of these issues. The first challenge, defining what kind of student growth to use, is confounded because of the heterogeneous populations typically served in special education (Holdheide et al., 2012). Even small groups of students typically present a varying spectrum of academic, social, and behavioral needs (Stough & Palmer, 2003; Tyler, Yzquierdo, Lopez-Reyna, & Flippin, 2004). For example, an extended resource room might serve students representing a range of disabilities including cognitive impairment, autism, behavioral disorders, and other health impairments. Two students might be placed in the classroom with the same exceptionality, e.g. cognitive impairment, but might vary widely in their academic, functional, communicative, and social interaction skills. This variation in student needs makes it difficult to select one student outcome measure that best “fits” a particular exceptionality, student group, or even classroom. Even if one student outcome could be identified as addressing the needs of all students in a special education classroom, the next step is to define how much academic growth is considered adequate. Leaving this complex determination of measurement left to districts to solve alone is concerning given the high-stakes nature of this type of assessment.

In addition to the challenges presented by trying to obtain a consistent measure of student growth, there are other challenges to designing a one size fits all approach and applying it to special education. Special education teachers work under a variety of contexts, sometimes providing a student’s entire instructional plan, and in other cases providing consultation to the student’s general education teacher. Some students with disabilities do not participate in the general assessment that other students do, but take an alternate or accommodated version. Given the flaws with VAM approaches for general education teachers compounded with the challenges of special education, these models seem highly questionable as an effective means of teacher evaluation. In its current design, using VAMs to identify and evaluate special education teacher efficacy is especially perplexing due to the unique roles and responsibilities within the profession. This suggests that models of teacher evaluation that rely heavily on student outcome measures (i.e. Value Added Model) may not be valid for special education.

Students with disabilities typically have academic performances that are significantly below their grade level peers, which systematically disadvantage special education teachers under a VAM model. This does not mean that special education teachers should not anticipate seeing growth in their students. A large body of research in special education supports the use of effective instructional practices to help students with disabilities realize significant academic improvements. Yet, in Vannest & Hagan-Burke's (2009) year-long study that carefully examined the way EBD special education teachers spend their time at school, results indicated that only 16% of the day is spent on providing direct academic instruction and that a significant portion of the day is spent on other, related tasks. What this suggests is that while instructional time is critical for the achievement of students in special education, and although states are moving to teacher evaluation systems that heavily weights student academic achievement, the current system does not support instructional time as a valued part of the special educator’s role. Thus, both systemic and inherent qualities work against special education teachers in a VAM-based teacher evaluation system.

There are several constraints that further complicate the development of a special education teacher evaluation. Holdheide, Goe, Croft, and Reschly (2010) identified larger challenges uniquely associated with special education teachers and evaluation systems including: a) special education is a high demand field, with many positions either vacant or filled with unqualified personnel (Billingsley, Fall & Williams, 2006; Boe & Cook, 2006; McLeskey, Tyler & Flippin, 2004); b) special education teachers are typically not highly qualified in the core content areas they teach (McLeskey & Billingsley, 2008); c) special education teacher preparation programs do not often integrate the use of evidence-based practices, thus leaving new special education teachers ill-prepared to meet the challenges of the special education classroom (Reschly, Holdheide, Smart & Oliver, 2007: Walsh, Glaser & Wilson, 2007). At the same time, special education remains a high demand field, with states reporting critical shortages of special education teachers, e.g. at the beginning of the 1999-2000 school year, almost 97% of all U.S. school districts reported at least one teaching vacancy in the field of special education (Connelly & Graham, 2009). Previous studies have found a positive relationship between levels of teacher knowledge and the quality of teacher instruction (Hill et al., 2008), and within a field defined by shortage and lack of qualified teachers, it seems that special education would greatly benefit from a teacher evaluation system that meets Danielson's features of ensuring teacher quality and promoting professional development.

Developing observational tools within observational systems

Research and policy interest continues to emphasize the influence of teacher effectiveness on student achievement. After student demographics, teacher effects have been shown to explain the most variance in student achievement, with some cumulative effects as high as 34% (Kyriakides & Creemers, 2008). However, this approach assumes that only true teacher quality comes from the measurement of teacher effectiveness, without taking account for multiple sources of variances in observational scores including the sampling of lessons, differences among evaluators, and unattended characteristics of the observation instrument itself (Hill, Charalambous, & Kraft, 2012). Differences among evaluators can be nuanced through professional roles (e.g. principals, consulting teachers, etc.), and accompanying levels of knowledge and awareness of specific instructional practices. These evaluator differences influence the overall reliability of a given teacher's rated effectiveness, and suggest that until all components of an observation system can be adequately defined, caution should be used when relying on observation scores in a high-stakes evaluation system (Hill et al., 2012).

Development of an observational tool

Given the current state of special education, it seems prudent that an evaluation system for special education teachers should have the systematic goal of increasing attention on improving the quality and quantity of instructional services provided to students with disabilities. It is with these purposes that we developed the Recognizing Effective Special Education Teachers (RESET) Observation Tool (OT). The RESET OT is based on a theory of effective special education teaching that *an effective special education teacher is able to identify a student's needs, implement evidence-based instructional practices and interventions, and demonstrate student growth*. To measure teaching effectiveness, the RESET OT provides an evaluation of a teacher's ability to deliver evidence-based instructional practices that align with the content, the nature of the disability and the grade level of their students. The RESET OT relies on observable, measurable criteria, and is aligned with the research base on effective instructional practices for students with disabilities.

Our approach to evaluating special education teachers is based on an observation of the special educator's use of evidence-based instructional practices, and the resulting student outcomes reported through effect sizes on measures aligned with relevant student goals. The observation of instructional practice constitutes the main focus of the evaluation system. Special education teacher observations are captured through video. This allows for multiple viewings to ensure reliability and provides objective data that allows the special education teacher to reflect on their performance and receive feedback from a skilled evaluator, which has been demonstrated to be extremely effective in supporting the implementation of research-based practices (Greene, 2009; Kane & Staiger, 2012).

Once the focus of observing and evaluating instruction was decided, we set out to produce a more systematic approach to instructional observation that was aligned with evidence-based practices for students with disabilities. We conducted a meta-review of the literature to identify instructional practices for students with disabilities, and then used the research-based descriptions of the salient features of that instructional practice in order to develop scoring criteria. Our goal with this review was to construct a list of the salient characteristics of those instructional practices with a strong evidence base to develop an observational system that is flexible enough to be used across multiple special education settings, but specific enough to provide reliable evaluations. Our initial work has reviewed several instructional practices, identified the salient characteristics of each and identified the range of effect sizes reported in the research when these practices are implemented with fidelity (see Appendix A for an example from the shortened version of this matrix). Work on developing this matrix is ongoing to ensure a complete listing of evidence-based practices included in the literature.

To collect observations of instruction, we used the Teachscape video capture system to record 12 special education teachers providing instruction to students across the state. The demographics and characteristics of the special education teachers captured on video are included in Table 1.

Table 1.

Special Education Teachers in Observation Dataset, n=12

	Resource			**Extended Resource**	**Autism**	**Early Childhood**
	Resource	Co-Teaching	Tier 2&3			
Elementary	2	1	1	1	2	1
Junior High	1			2		1

To construct the evaluation tool, we developed scoring criteria using a four-point Likert scale. After reviewing the initial set of video recorded instructional lessons, we ensured that the practices included in the data set had corresponding evaluation criteria.

The initial version of the RESET OT consists between 28-67 items depending on the number of instructional practices being observed. The tool is web-based and operates on a direct-logic system, (i.e. some questions only appear if previous questions have been answered a pre-defined way). The RESET OT is still in its early stages of development and numerous other studies and stages of development will be required before it is ready to be used in practice. In this paper, we report on a pilot study investigating whether we could obtain high levels of interrater agreement.

Interrater agreement is defined as the degree to which two or more raters achieve identical results under similar assessment conditions. Interrater agreement is a critical component of establishing an instrument's overall consistency, and the specific procedures used are described in more detail in the methods section. One of the critical concerns for observation systems is that they have high levels of interrater reliability – scores should not vary based on who is assigning the judgments of performance. Therefore, we began our validation and development studies of the RESET OT with this initial look at whether master special education teachers could achieve high levels of interrater reliability when evaluating a special education instructional lesson.

Method

Participants

Six special education teachers were asked to participate as data coders in June 2012. The teachers were selected from the state's existing special education mentor network, and identified by their district's special education directors as exemplary. One elementary and one secondary teacher representing the range of special education instructional settings, were recruited to serve as data coders. Three of the six participating teachers were also part of the video capture phase of the study, i.e. they were both the observed and the participants, however no teacher observed herself at anytime during the coding sessions. All teacher coders were female.

Table 2.

Teacher Coder Demographics

	Years Teaching (Total, Special Ed)	Highest Level of Education Completed	Grade Level	Special Education Instructional Context
Teacher 1	20, 12	Bachelors	Elementary	Resource
Teacher 2	13	Masters	Elementary	Extended Resource
Teacher 3	37, 32	Masters	Elementary	Self-contained, EBD
Teacher 4	11	Masters	Secondary	Resource
Teacher 5	19, 9	Masters	Secondary	Extended Resource/Severe
Teacher 6	5	Masters	Secondary	Self-contained, EBD

Setting

The three-day coding session was carefully designed to protect the confidentiality and anonymity of the videos and the teacher coders. Evaluation sessions were held in a conference room on a university campus that was only open to those participating in the project. Teacher coders were placed far enough away from one another so that they could not see what a neighboring teacher was coding, and all teacher coders were given headphones to wear throughout the three-day session. At any given time, a teacher coder had two university-owned laptops in use: one to watch the assigned Teachscape video, and one to complete the observation tool. Both the RESET project director and project coordinator were available throughout the three-day coding session to answer questions and provide assistance.

Procedures

Video data. Videos were recorded and accessed through the online Teachscape Reflect system, the same technology used by the Measures of Effective Teaching (MET) study funded by the Bill and Melinda Gates Foundation (Kane & Staiger, 2012). The Teachscape video capture system consists of two cameras: 1) a 360-degree camera which allows the observer to pan and zoom on various components of the classroom environment, and 2) a fixed position camera, also referred to as a "board cam" because it is usually focused on a classroom board. The RESET project coordinator collected the videos from 12 special education teachers across three school districts (two large and one rural), beginning in October 2011 to March 2012.

After the videos were initially captured, they were processed via the online Teachscape Reflect video system. The videos were uploaded to the RESET administration account, which only the RESET project director, project coordinator, and program support person had access to. The project coordinator assigned these videos to Teachscape user accounts individually created for the teacher data coders.

Assignment to videos. The project coordinator created Teachscape Reflect user accounts for each teacher coder, and pre-assigned videos were shared with the assigned teacher coder. These accounts were only made available during the scheduled data coding sessions on campus, and the teacher coder user accounts were deleted at the end of the three-day session.

Over 1,800 instructional minutes were originally captured from October 2011-March 2012. This original dataset consisted of observations from the moment the cameras began and ended recording (based on scheduled times decided between the RESET project and the teachers involved in the study). For the purposes of the June 2012 data coding sessions, the original observations were reduced by removing any time that did not capture instruction. As a result, the original 1800 minutes was reduced to 1,311 instructional minutes which comprised the data set for the coding session described in this study. Guided by recommendations from the MET study (Kane & Staiger, 2012), the following criteria were used to assign videos from the 1,311 instructional minutes dataset to teacher coders:

1. All videos must be coded at least twice.
2. All teacher coders were assigned at least two videos of any one teacher being coded.
3. All teachers being coded were assigned a teacher coder pair that met the following two preceding criteria.

The video assignment list created from the criteria was reviewed for equitable distribution of minutes (shortest assignment

= 421 minutes, longest assignment = 440 minutes), and to ensure that no teacher was coding herself. Each coder watched at least 15 videos, and no more than 16 videos. The total paired analysis dataset includes 86 videos, or 43 pairs.

Training. A teacher manual to accompany the RESET OT was developed. The manual explained the structure of the RESET OT, and provided definitions and descriptions for the evaluation criteria as well as for the instructional practices included on the evaluation tool. On the first day of the data coding session, teacher coders were oriented through the manual and a blank observation tool. Next, a training video was observed as a whole group activity, and teachers evaluated the training video using the RESET OT. The training video scores across the six raters were reviewed and then compared for agreement. Differences in scores were discussed until consensus was reached. Finally, the first calibration video was individually evaluated using the RESET OT, and the scores across the six raters were again reviewed and compared for agreement in a whole group activity. On the second day of the data coding session, the second calibration video was evaluated using the RESET OT, and the scores across the six raters were reviewed and compared in a whole group activity. Data coding officially began in the afternoon of the second day of the session, and continued until the end of the third day.

Measures

The RESET OT was designed to be responsive to special education instructional practices, and as a result, adjusts to different placements, classrooms, grades and exceptionalities. The tool consists of three main parts: the Lesson Overview (similar to an introduction), the specific Lesson Components (focus on instructional practices), and the Lesson Summary (similar to a conclusion). Each lesson component also includes its own overview and summary.

The RESET OT operates on a direct logic system, (i.e. certain follow up questions only appear if previous questions have been selected). This is how questions related to specific, evidence-based instructional practices are addressed in the tool—only when a specific instructional practice is indicated will its components be presented for evaluation. For example, if direct, explicit instruction is indicated as the instructional practice being used in the lesson component, then only the scoring criteria related to direct instruction are revealed to the observer. However, if promoting self-determination (Wehmeyer & Field, 2007) is selected, then only the criteria for evaluating self-determination are available to the observer.

During the time of this study, the RESET OT evaluation rubric of the effective use and implementation of evidence-based instructional practices was aligned with Danielson's (2007) four-point scale of observed behavior: Unsatisfactory, Basic, Proficient and Distinguished. This evaluation scale is the primary rubric used to evaluate observed special education instruction in the RESET OT, in order to align the system with the state’s larger teacher evaluation model.

Data Analysis

Once all videos had been evaluated and scored, a data base consisting of both sets of scores for each video was created. Interrater reliability was then conducted on the data set using both perfect agreement and kappa. Each of these approaches is described below.

Perfect agreement. Scores on an item were counted as being in perfect agreement only when the scores were identical. Perfect agreement was then calculated by dividing the number of items with perfect agreement by the total number of items scored.

Kappa. To control for perfect agreement obtained by chance, we also used the kappa statistic to analyze data. Kappa was calculated in SPSS using the formula: observed percentage of agreement-expected percentage of agreement/1-expected percentage of agreement. Weighted kappa analyses were used for ordinal items, and unweighted kappa analyses were used for nominal ones.

Results

Tables 1-3 include the item agreement and Kappa analyses, and each table is defined by the instructional practice selected. Perfect agreement results are generally interpreted "intuitively", that is, the closer to 1, the stronger the agreement (Perreault & Leigh, 1989). However, this calculation can be misleading because the percentage agreement can be influenced heavily by the number of coding categories, i.e., the smaller the number of categories, the greater the likelihood of higher agreement due to chance alone (Cohen, 1960). The kappa statistic takes chance agreement into consideration, making it a more "well-behaved" index, and is generally found to be lower than the percentage of perfect agreement (Cohen, 1960; Perreault & Leigh, 1989). Landis & Koch (1977) have characterized different ranges of values for kappa with respect to the degree of agreement they suggest. For most purposes, values greater than 0.75 may be taken to represent excellent agreement beyond chance, values below 0.40 may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance (Landis & Koch, 1977).

Table 3 reports the item agreement for the lesson objective questions from the Lesson Overview and Component #1 (LO1-LO6), student engagement and instructional implementation questions from Component #1 (COMP1-COMP6), and overall lesson instructional practice questions from the Lesson Summary (LS1-LS4). The item agreement in Table 3 includes all paired analyses from the dataset that includes at least one component (Component #1) in the lesson, n=43. The LO1-LO6 kappa scores indicate fair to good agreement with scores ranging from .52 to .93. The COMP1-COMP6 kappa scores indicate no agreement, except for COMP3 with .57, which suggests fair to good agreement. The LS1-LS4 Kappa scores indicate fair agreement with scores ranging from .45 to .55, except LS2 with no agreement.

Table 3.

Lesson Overview (LO), Component #1 (COMP), Lesson Summary (LS), n=43

Item	Grand Total	Total Agreement	% Perfect Agreement	Kappa
LO1 Lesson objective evident to students	43	19	44%	.60*
LO2 Classroom routine(s) evident to students	43	36	84%	.86*
LO3 Level of instructional intensity (student to teacher or instruction professional ratio)	43	39	91%	.93*
LO4 Component #1 objective evident to students	43	22	51%	.65*
LO5 Component #1 objective aligned with the larger lesson objective?	43	20	47%	.61*

Table 3.

Lesson Overview (LO), Component #1 (COMP), Lesson Summary (LS), n=43

Item	Grand Total	Total Agreement	% Perfect Agreement	Kappa
LO6 Knowledge of curriculum being used in component #1	43	14	33%	.52
COMP1 Student Engagement-Student participation in tasks and activities	43	12	28%	.00
COMP2 Student Engagement-Students ask and answer questions (OTR)	43	13	30%	.00
COMP3 Student Engagement-Students use of learning strategies	43	18	42%	.57
COMP4 Student Engagement-Students show interest and enthusiasm for certain topics	43	22	41%	.00
COMP5 Implementation-Does the teacher adjust the lesson according to student response?	43	18	41%	.00
COMP6 Implementation-Is the lesson pacing appropriate to student abilities?	43	16	37%	.54
LS1 Is the use of time effective for lesson's learning objective	43	15	35%	.54
LS2 Does the teacher appear to have a solid understanding of the content/curriculum?	43	16	37%	.00
LS3 Does the teacher implement effective instructional practices?	43	10	23%	.45
LS4 Does the teacher effectively respond to student needs?	43	16	37%	.55

*indicates unweighted kappa analyses

Table 4 reports the item agreement for the lesson objective questions from the Lesson Overview and Component #1 (LO1-LO6), student engagement and instructional implementation questions from Component #1 (COMP1-COMP6), specific questions related to the “direct, explicit instruction” parts (DI1-DI5) and overall lesson instructional practice questions from the Lesson Summary (LS1-LS4). The item agreement in Table 4 includes all paired analyses from the dataset that indicated “direct, explicit instruction” as the primary form of instruction for Component #1 in the lesson, n=20. The LO1-LO6 kappa scores indicate good agreement with scores ranging from .60 to 1, except LO4 with no agreement. The DI1-DI6 kappa scores indicate no agreement, except DI5 with fair agreement at .52. The COMP1-COMP6 kappa scores indicate good to strong agreement with scores ranging from .71 to .83, except for COMP2 and COMP5 with no agreement. The LS1-LS4 kappa scores indicate no agreement except LS1 with good agreement at .73.

Table 4.

Lesson Overview (LO), Component #1 Direct Instruction (DI), Component #1 (COMP), Lesson Summary (LS), n=20

Item	Grand Total	Total Agreement	% Perfect Agreement	Kappa
LO1 Lesson objective evident to students	20	9	45%	.73*
LO2 Classroom routine(s) evident to students	20	18	90%	.95*
LO3 Level of instructional intensity (student to teacher or instruction professional ratio)	20	20	100%	1*
LO4 Component #1 objective evident to students	20	11	55%	.00*
LO5 Component #1 objective aligned with the larger lesson objective?	20	11	55%	.78*
LO6 Knowledge of curriculum being used in component #1	20	4	20%	.60
DI1 Component #1 instructional strategies -Direct, explicit Instruction	20	10	50%	.67
DI2 Organized Instruction	20	7	35%	.00
DI3 Sequenced Instruction	20	6	30%	.00
DI4 Student Participation	20	7	35%	.00
DI5 Scaffolding	20	5	25%	.52
DI6 Assessment	20	6	30%	.00
COMP1 Student Engagement-Student participation in tasks and activities	20	11	55%	.78
COMP2 Student Engagement-Students ask and answer questions (OTR)	20	6	30%	.00
COMP3 Student Engagement-Students use of learning strategies	20	9	30%	.71
COMP4 Student Engagement-Students show interest and enthusiasm for certain topics	20	13	35%	.83
COMP5 Implementation-Does the teacher adjust the lesson according to student response?	20	8	40%	.00
COMP6 Implementation-Is the lesson pacing appropriate to student abilities?	20	11	55%	.78
LS1 Is the use of time effective for lesson's learning objective	20	9	45%	.73
LS2 Does the teacher appear to have a solid understanding of the content/curriculum?	20	9	45%	.00
LS3 Does the teacher implement effective instructional practices?	20	8	40%	.00
LS4 Does the teacher effectively respond to student needs?	20	9	45%	.00

*indicates unweighted kappa analyses

Table 5 reports the item agreement for the lesson objective questions from the Lesson Overview and Component #1 (LO1-LO6), student engagement and instructional implementation questions from Component #1 (COMP1-COMP6), specific questions related to the “other, instruction” parts (OTH-OTH6) and overall lesson instructional practice questions from the Lesson Summary (LS1-LS4). The item agreement in Table 5 includes all paired analyses from the dataset that indicated “other, instruction” as the primary form of instruction for Component #1 in the lesson, n=31.

The LO1-LO6 kappa scores indicate good agreement with scores ranging from .52 to .90. The DI1-DI6 kappa scores indicate no agreement, except OTH1 and OTH2 with fair agreement at .52 and .56, respectively. The COMP1-COMP6 kappa scores indicate no agreement, except for COMP3 and COMP4 at .52 and .20, respectively. The LS1 and LS4 kappa scores fair agreement at .51 and .52, respectively, and LS2 and LS3 indicate no agreement.

Table 5.

Lesson Overview (LO), Component #1 Other (OTH), Component #1 (COMP), Lesson Summary (LS), n=31

Item	Grand Total	Total Agreement	% Perfect Agreement	Kappa
LO1 Lesson objective evident to students	31	13	42%	.59*
LO2 Is classroom routine(s) evident to students	31	24	77%	.81*
LO3 Level of instructional intensity (student to teacher or instruction professional ratio)	31	27	94%	.90*
LO4 Component #1 objective evident to students	31	14	48%	.63*
LO5 Component #1 objective aligned with the larger lesson objective?	31	16	52%	.65*
LO6 Knowledge of curriculum being used in component #1	31	10	32%	.52
OTH1 Component #1 instructional strategies Other	31	17	55%	.56
OTH2 Academic focus	31	12	39%	.49
OTH3 Precise sequencing of content	31	6	19%	.00
OTH4 High student engagement (ability appropriate)	31	7	23%	.00
OTH5 Careful teacher monitoring of student progress	31	6	19%	.00
OTH6 Specific corrective feedback to students	31	6	19%	.00
COMP1 Student Engagement-Student participation in tasks and activities	31	14	45%	.00
COMP2 Student Engagement-Students ask and answer questions (OTR)	31	14	45%	.00
COMP3 Student Engagement-Students use of learning strategies	31	10	32%	.52
COMP4 Student Engagement-Students show interest and enthusiasm for certain topics	31	16	52%	.20
COMP5 Implementation-Does the teacher adjust the lesson according to student response?	31	14	45%	.00
COMP6 Implementation-Is the lesson pacing appropriate to student abilities?	31	11	35%	.00
LS1 Is the use of time effective for lesson's learning objective	31	9	29%	.51
LS2 Does the teacher appear to have a solid understanding of the content/curriculum?	31	12	39%	.00
LS3 Does the teacher implement effective instructional practices?	31	5	16%	-.13
LS4 Does the teacher effectively respond to student needs?	31	10	32%	.52

*indicates unweighted kappa analyses

Discussion

Tables 3-5 report similarites and differences in kappa scores that support generalizations of the findings. The first similarity is the overall level of fair to good agreement in the Lesson Overview sections for the three tables. This indicates that raters generally agree upon the major components of the lesson. Contrasting two of the strongest consistent agreements amongst the three tables, LO2, “Classroom routine(s) evident to students” and COMP3 “Student Engagement-Students use of learning strategies”, to one with a consistently weaker agreement LO1 “Lesson objective evident to students”, indicates that raters tended to find more agreement distinguishing between repetitive versus instructional classroom processes. This pattern is in alignment with the MET study that found those characteristics related to observing student behavior and class management leaned towards higher levels of rater reliability than items related to instruction like questioning and communicating with students (Kane & Staiger, 2012). Another similarity is the consistent lack of agreement for LS2 “Does the teacher appear to have a solid understanding of the content/curriculum?” across all tables. This finding suggests there is a lack of agreement as to what “content/curriculum” might look like, or there might be disagreement between coders on a given teacher’s ability to demonstrate this understanding. The last, and most pronounced similarity among the kappa scores in Tables 4 and 5 is the consistent lack of agreement within the specific components of instructional practices (DI1-DI6 and OTH1-OTH6). This finding suggests among other things, there is a deep lack of understanding about what these instructional practices are and what different levels of proficiency look like. This suggests the importance of training evaluators when using a tool specific to evidence-based instructional practices.

The differences in kappa scores, both between and within Tables 3-5, are a reflection of a lack of agreement among raters either within a specific teacher’s observed instructional practice, the disagreement among raters on the evaluation criteria and terms of the RESET OT itself. The sporadic levels of disagreements within kappa scores could be a reflection of the prevalence of the behavior being observed, and how this affects the kappa calculation (Banerjee & Fielding, 1997; Feinstein & Cicchetti, 1990; Feuerman & Miller, 2008). The true prevalence of a target behavior is defined by the relative probability of “Yes” and “No” in the population. If the “true” prevalence of an observed behavior is high, then proportion of agreement expected by chance is enlarged, and thus lowers the value of kappa (Banerjee & Fielding, 1997; Feuerman & Miller, 2008). In the case of the RESET OT, if an evaluator (teacher coder) is untrained on the specificity and sensitivity of the tool, and is unable to distinguish between classroom instruction and routine, then the “true” prevalence of instruction can become erroneously high, inaccurately enlarging the proportion of agreement expected by chance, and lowering the value of kappa. Another way to state this is to say that if a teacher coder does not fully understand the specifics of instructional practice, but indicates that “proficient” or “distinguished” levels of instruction is being observed, when it is not actually present, it can lead to a distorted level of “true” prevalence, ultimately lowering kappa scores.

Another important difference in the results are the discrepancies between the reported perfect agreement percentages and kappa scores. For example, in Table 3 COMP4 “Student Engagement-Students show interest and enthusiasm for certain topic" and COMP5 "Implementation-Does the teacher adjust the lesson according to student response?" both report perfect agreement scores of 40%, but a kappa score of .00. This kappa score would seem to indicate that the level of agreement is non-existent. However, the reason for this paradox between the perfect agreement and kappa might be due to the fact that the presence of high student engagement and effective instructional implementation are rare findings in the dataset, suggesting that kappa may not be reliable (Feinstein & Cicchetti, 1990; Viera & Garrett, 2005). Kappa is affected by the prevalence of the finding (presence of effective instruction) under consideration (Viera & Garrett, 2005), and one method to account for this paradox is to distinguish between agreement on the two levels of the finding, e.g. agreement on positive ratings compared to agreement on negative ratings (Feinstein & Cicchetti, 1990). For the purposes of this pilot study, this type of analysis was not conducted, but it might be useful to consider in future studies.

Furthermore, because rater agreement rates attend to only one source of variation—the rater— it leaves out other sources of variation that affect the consistency within teacher scores. This emphasis on this one specific type of agreement fails to address interactions between raters, teachers, and lessons (Hill et al., 2012). As a result, there might be many other ways to interpret the coded data that is narrowly restricted through the lens of a singular measurement and variation. Thus, this narrow interpretation of data through rater agreement might not necessarily provide a comprehensive picture of the reliability of scores generated from an observational system, especially when that system is within its infancy stages. The interpretation of the data produced by teacher coders might better lend itself to generalizability theory analyses, which will be considered in future projects (Brennan, 1992; Hill et al., 2012).

The differences in kappa scores between and within the tables might also be a result of the more practical reason of the initial disagreement of what type of instructional practice was being utilized. In the case of Tables 4 and 5, which were organized by "direct, explicit instruction" and "other" instructional practices, perfect agreement and kappa scores were reported simply by the presence of agreement. However, if a pair of teacher coders selected a different type of instructional practice, e.g. in the case of Table 5 "other", one teacher coder selected "other", but the other selected "self-determination", then all of the following ratings for that video will report only the scores for the teacher that selected "other". Because the RESET OT only presents specific instructional components for an instructional practice when it is selected, the teacher coder that selected "self-determination" would not be given the opportunity to evaluate any of the components for "other". As a result, this creates a dataset with missing values for a given pair, potentially distorting the reported rater agreement values.

Lastly, on a more subjective note, the differences in agreement and kappa scores might be a reflection of our special education system's lack of focus on instructional practice. While current educators and policy makers claim to value instructional practice, sufficient evidence in recent years describes a special education system that is burdened by administrative requirements and that increasingly moves away from the focus on providing individualized instruction. Additionally, there is strong evidence suggesting that teachers enter the field inadequately prepared because universities do not universally focus on teaching evidence-based practices to special education teacher candidates, and because special education is a profession plagued by attrition and high turnover, which makes it a persistent high demand field (Gersten, Vaughn, Deshler, & Schiller, 1997; Greenberg, Pomerance, & Walsh, 2011; Greenberg & Walsh, 2012). These factors might combine to create a current status quo of the majority of a profession, while dedicated and working hard to serve children with disabilities, unable to implement and unable to recognize effective instruction.

Limitations

This paper reports the results of the interrater reliablity study using the pilot RESET OT, and there are several limitations that warrant caution in generalizing the results. The first limitation is that the reported study reflects the initial attempt at collecting psychometric evidence of the RESET OT. Although the observation tool has been rewritten and revised through multiple versions, the version used in this study was the first used by members outside of the research setting, and it will undergo numerous changes as we continue with development, validation and field studies. Second, the level of training the participating teacher coders received on specific components of evidence-based instructional practices was minimal—it was assumed that the teachers selected to participate in the study came with a strong background in evidence-based instructional practices. Future interrater reliability studies will take measures to address this issue by 1) providing a lengthy, in-depth training on the specific components of evidence-based instructional practices, and 2) conducting pre and post tests of teacher coder knowledge of evidence-based instructional practices before and after the training session. Third, a pre-determined level of interrater reliability was not required before allowing teacher coders to evaluate instruction. Large-scale, but similar studies require coders to pass an interrater reliability threshhold before being allowed to code, and future studies on the RESET OT will maintain similar requirements (Kane & Staiger, 2012). Fourth, using kappa as a measure of interrater reliability might have limited the generizability of the findings. Banerjee & Fielding (1997) point out that the use of kappa as a measure sometimes can be misleading because of the limitations of the statistic, and given the potential biases and influences on the prevalence of the presence of effective instructional practices, the kappa scores should be interpreted with caution.

Conclusion

In this study, we examined the interrater reliability scores of six, selected teacher coders who used the pilot RESET OT to evaluate special education instruction. The results of this study indicate that while there are some areas that maintained fair to good agreement, there are other areas that consistently reported no agreement and signal the need for more research and development. Based on the findings from this study, future directions for this research include 1) deeper development of the RESET OT and the specification of evidence-based instructional practices 2) a repeat of the interrater reliability study using a group of teacher coders who have received in-depth training on both the revised RESET OT as well as evidence-based instructional practices, and 3) further investigation into the specific components of evidence-based instructional practice and how these might present similarly or differently across special education settings.

References

Banerjee, M., & Fielding, J. (1997). Interpreting kappa values for two-observer nursing diagnosis data. *Research in Nursing and Health*, *20*, 465–470.

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practices*, *11*, 27–34.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

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

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., Ruby, M., Crevecoeur, Y. C., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness*, *3*, 93–120. doi:10.1080/19345741003592410

Danielson, C. (2007). *Enhancing professional practice□: A framework for teaching*. Alexandria, Va.: Association for Supervision and Curriculum Development.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership*, *68*, 35–39.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549.

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: Sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice*, *14*, 930–933. doi:10.1111/j.1365-2753.2008.00984.x

Gersten, R., Vaughn, S., Deshler, D., & Schiller, E. (1997). What we know about using research findings: Implications for improving special education practice. *Journal of Learning Disabilities*, *30*, 466–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9293227

Greenberg, J., Pomerance, L., & Walsh, K. (2011). *Student teaching in the United States*. Retrieved from http://www.nctq.org/edschoolreports/studentteaching/docs/nctq_str_full_report_final.pdf

Greenberg, J., & Walsh, K. (2012). *What teacher preparation programs teach about K-12 assessment: A review*. Retrieved from http://www.nctq.org/p/publications/docs/assessment_report.pdf

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511. doi:10.1080/07370000802177235

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

Holdheide, L. (2012). State considerations in designing and implementing evaluation systems that include teachers of students with disabilities. *Office of Special Education Programs Project Director's Conference*. Washington DC.

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.tqsource.org/pdfs/TQ_Forum_SummaryUsing_Student_Growth.pdf

Holdheide, L., Goe, L., Croft, A., & Reschly, D. J. (2010). *Challenges in evaluating special education teachers and english language learner specialists* (pp. 1–40). Washington DC.

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

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford Review of Education*, *34*, 521–545. doi:10.1080/03054980701782064

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*, 363–374.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101. doi:10.3102/10769986029001067

National Council on Teacher Quality. (2011). *State of the states: Trends and early lessons on teacher evaluation and effectiveness policies*. Washington DC.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23).

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, *26*, 135–148.

Stough, L. M., & Palmer, D. J. (2003). Special thinking in special settings: A qualitative study of expert special educators. *The Journal of Special Education*, *36*, 206–222.

Tyler, N. C., Yzquierdo, Z., Lopez-Reyna, N., & Flippin, S. S. (2004). Cultural and linguistic diversity and the special education workforce: A critical overview. *The Journal of Special Education*, *38*, 22–38.

Vannest, K. J., & Hagan-Burke, S. (2009). Teacher time use in special education. *Remedial and Special Education*, *31*, 126–142. doi:10.1177/0741932508327459

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*, 360–363. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15883903

Wehmeyer, M. L., & Field, S. L. (2007). *Self-determination: Instruction and assessment strategies*. Thousand Oaks, CA: Corwin Press.

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement of students with disabilities, ages 7 to 17. *Exceptional Children*, *78*, 89–106.

Author Note

This research was supported by a grant from the Idaho State Department of Education. Correspondence concerning this article should be addressed to Carrie Semmelroth, Department of Special Education, Boise State University, Boise, ID 83725. Email: carriesemmelroth@boisestate.edu.

Appendix A: Excerpt from Evidence-Based Instructional Practice Matrix (shortened form)

Authors	Instructional Practice	Description of Instructional Practice	Exceptionality	Grade/Age	Academic Achievement Area	Effect Size	Standardized Assessments to Measure Student Performance
Torgesen, Alexander, Wagner, Rashotte, Voeller, & Conway, 2001	Direct Instruction	Instructional programs incorporating effective instruction in phonemic awareness and phonemic decoding skills. Students randomly assigned to 2 groups, 1 receiving Auditory Discrimination in Depth Program (ADD), the other received Embedded Phonics (ED). (No definite "control group" identified in the study, though the regular resource room intervention is assumed as the control group.) After intensive intervention provided, children received generalization training within general education classroom for one 50-minute session for application skills across environments.	LD, ADHD	8 - 10 years	Basic Reading	1.38	Woodcock Reading Mastery Test - Revised Word Identification measure (ADD)
					Reading Comprehension	0.56	Woodcock Reading Mastery Test - Revised Passage Comprehension measure (ADD)
					Basic Reading	1.61	Woodcock Reading Mastery Test - Revised Phoneme Decoding Efficacy measure (ADD)
					Basic Reading	1.47	Gray Accuracy (ADD)
					Reading Fluency	0.58	Gray Rate (ADD)
					Reading Comprehension	1.18	Gray Comprehension (ADD)
					Basic Reading	2.31	Woodcock Reading Mastery Test - Revised Word Attack measure (EP)
					Basic Reading	1.54	Woodcock Reading Mastery Test - Revised Word Identification measure (EP)
					Reading Comprehension	0.64	Woodcock Reading Mastery Test - Revised Passage Comprehension measure (EP)
					Basic Reading	1.34	Woodcock Reading Mastery Test - Revised Phoneme Decoding Efficacy measure (EP)
					Basic Reading	0.84	Gray Accuracy (EP)
					Reading Fluency	0.07	Gray Rate (EP)
					Reading Comprehension	0.58	Gray Comprehension (EP)
					Basic Reading	0.30	Woodcock Reading Mastery Test - Word Identification (WRMT-WI): Phonemic Decoding Efficiency (Exp)

Appendix A: Excerpt from Evidence-Based Instructional Practice Matrix (shortened form)

Authors	Instructional Practice	Description of Instructional Practice	Exceptionality	Grade/Age	Academic Achievement Area	Effect Size	Standardized Assessments to Measure Student Performance
Spencer & Manis, 2010	Direct Instruction	A randomized experimental design to measure/test the efficacy of a fluency intervention program on the word-identification and reading-comprehension outcomes of students with severe reading disabilities. Both experimental and control groups were provided 10 minutes a day of 1-to-1 instruction with a trained paraprofessional, the control group receiving study skills instruction and experimental group receiving fluency-based instruction in areas of sounds/individual words, short sight phrases, and connected text.	Self-Contained: LD, MR, Lang Imp, ASD	6th - 8th Grade (10 - 15 years old)	Basic Reading	0.32	WRMT-WI: Sight Word Efficiency (Exp)
					Basic Reading	-0.09	WRMT-WI: Word Attack (Exp)
					Basic Reading	-0.11	WRMT-WI: Word Identification (Exp)
					Reading Fluency	0.60	GORT-III: Rate (Exp)
					Reading Fluency	0.71	GORT-III: Accuracy (Exp)
					Reading Comprehension	0.72	GORT-III: Passage (Exp)
					Reading Comprehension	0.05	WRMT-R/NU: Comprehension (Exp)
Coyne, McCoach, Loftus, Zipoli, Ruby, Crevecoeur & Kapp, 2010	Direct Instruction; Peer Support	Quasi-experimental design investigating efficacy of direct vocabulary instruction for schools serving academically at-risk populations. Program looked specifically at target word knowledge and generalized language ad literacy. Separate classes at Schools A and B were assigned to either a treatment or control condition, whereas all individual students at School C were randomly assigned to experimental or control groups. 36 half-hour lessons total, 2 lessons per week over 16 weeks	GenEd/Collaborative: ELL, LD, EBD, DD, ADHD, Lang Imp	Kindergarten	Vocabulary	0.22	Peabody Picture Vocabulary Test (PPVT-III): Control
					Vocabulary	0.33	PPVT-III: Experimental