Interrater Agreement of the Individualized Behavior Rating Scale Tool

Rose Iovannone, Paul E. Greenbaum, Wei Wang, Glen Dunlap and Don Kincaid

Assessment for Effective Intervention published online 16 May 2013
DOI: 10.1177/1534508413488414

Hammill Institute on Disabilities

Assessment for Effective Intervention
XX(X) 1–13

DOI: 10.1177/1534508413488414
aei.sagepub.com
SAGE

Interrater Agreement of the Individualized Behavior Rating Scale Tool

Rose Iovannone, PhD[1], Paul E. Greenbaum, PhD[1], Wei Wang, PhD[1], Glen Dunlap, PhD[1], and Don Kincaid, EdD[1]

Abstract

Data assessment is critical for determining student behavior change in response to individualized behavior interventions in schools. This study examined the interrater agreement of the Individualized Behavior Rating Scale Tool (IBRST), a perceptual direct behavior rating tool that was used by typical school personnel to record behavior occurrence in students requiring individualized interventions. Two independent observers (teacher and data collector) used the IBRST to rate student-specific problem and appropriate behaviors during specified observation times. Data were collected across 19 students and agreement between raters was compared. Resulting linear- and quadratic-weighted kappa coefficients indicated generally adequate agreement between raters on problem behaviors and appropriate behaviors. When ratings were categorized into more or less salient behaviors, less than adequate agreement (<.60) was found for some behaviors that were less salient. Agreement remained stable from baseline to intervention. Implications for practice, limitations of the study, and directions for future research are discussed.

Keywords

behavior measures, behavior outcomes, direct behavior ratings

During the last decade, there has been an increased focus on collecting data on student educational outcomes in response to academic and behavior interventions and using the data within a problem-solving framework for making decisions about intervention (Batsche et al., 2005; Deno, 1989, 1995). The problem-solving framework, initially conceptualized by Deno (1989, 1995), consists of four major steps, including (a) identifying the problem, (b) analyzing the problem, (c) developing and implementing interventions to address the problem, and (d) evaluating the student response to the interventions and determining next steps. The crux of the problem-solving framework is that academic and behavior intervention decisions are based on the student outcome data (Batsche et al., 2005). In the academic world, curriculum-based measurement (CBM) procedures have established a wealth of empirical support for their use as a standardized, repeated-measurement method for evaluating student outcomes in response to academic interventions (Batsche et al., 2005; Gresham, 2003; Shapiro & Eckert, 1994; Shinn, 1989; Vaughn, Linan-Thompson, & Hickman, 2003).

The need for a similar psychometrically sound yet feasible method of standardized repeated-measurement methods used in the context of typical schools for evaluating behavior outcomes of individual students has eluded the field, specifically for those working with students having the most serious problem behaviors requiring intensive individualized levels of behavioral supports (Chafouleas, Volpe, Gresham, & Cook, 2010). The problem in creating such a data tool that emulates CBM has been twofold. First, grade-level behavior standards, benchmarks, or outcomes are not universally predetermined for behavior as they are for reading fluency or mathematical operations (Chafouleas et al., 2010). Rather, behavior expectations are more typically defined contextually at multiple levels (e.g., teacher, classroom, school, district, community) being influenced by philosophies, tolerance levels, culture, and values (Gresham, 2004; Jones, Caravaca, Cizek, Horner, & Vincent, 2006). The absence of universal behavior outcomes presents challenges in developing standardized repeated-measure methods of behavior assessment. Second, most of the school-based behavioral assessment measures in existence have been developed primarily for determining eligibility for special education disability categories (e.g., emotional/behavioral disorders) and are not intended or advised to be used for continuous assessment of behavioral

[1]University of South Florida, Tampa, USA

Corresponding Author:
Rose Iovannone, University of South Florida, 13301 Bruce B. Downs Blvd., MHC 2113A, Tampa, FL 33612, USA.
Email: iovannone@usf.edu

change (Crone & Horner, 2003; Repp & Horner, 1999). Therefore, there is a need for a standardized behavior data tool that can be used by school personnel to determine student outcomes in response to intensive, individualized behavior interventions.

The gold standard for individualized behavior assessment traditionally has been systematic direct observations (SDO; Shapiro & Heick, 2004). SDO procedures include precise definitions of target behaviors; identification of contexts, routines, or time periods of equal length in which repeated observations will occur; specific recording methods; and consistent checks of interobserver agreement (Gall, Gall, & Borg, 2007; Salvia, Ysseldyke, & Bolt, 2007). Although SDO techniques have been the core method of evaluating single-subject research and have been used specifically for examining the impact of interventions on individual student behavior change (see Cooper, Heron, & Heward, 2007), they have drawbacks when used in applied school settings. The foremost of these drawbacks is the resources required (i.e., time and skill level) to conduct SDOs in the classroom setting (Chafouleas, McDougal, Riley-Tillman, Panahon, & Hilt, 2005; Hintze & Matthews, 2004). Thus, the impracticality of SDOs makes them less useful as behavior outcome data tools within a problem-solving process (Christ, Riley-Tillman, Chafouleas, & Jaffery, 2011).

More recently, an emerging research base has explored the use of direct behavior ratings (DBR) as a viable means for efficient evaluation of student outcomes that, in specific settings, can serve as a practical alternative for SDO. DBRs are not meant to supplant SDO; rather, they can be considered as a set of strategies that produce a reasonable estimate of direct observations that are more efficient and practical for teacher use in evaluating student outcomes in applied settings (Chafouleas et al., 2005). DBRs combine the SDO characteristic of repeated measurement to record behavior occurrence within a specified routine or time frame with the efficiency of the rating scale characteristic of using rankings to represent the degree of behavior occurrence (Chafouleas, Riley-Tillman, & Christ, 2009). An example of a common DBR is the Daily Behavior Report Card (DBRC) used in the Behavior Education Program (BEP) and other similar small group or supplemental behavior interventions (Crone, Hawken, & Horner, 2010; Riley-Tillman, Chafouleas, & Briesch, 2007). The expanding literature base to examine the use of DBRs as evaluation tools for social behaviors has evaluated their efficacy for school-based progress monitoring (Riley-Tillman, Methe, & Weegar, 2009), explored their psychometric properties (Burke, Vannest, Davis, Davis, & Parker, 2009; Chafouleas et al., 2005), assessed variables that impact accuracy of ratings (Riley-Tillman, Chafouleas, Briesch, & Eckert, 2008; Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009), and examined acceptability of use (Chafouleas, Riley-Tillman, & McDougal, 2002; Chafouleas, Riley-Tillman, & Sassu, 2006). Although the research has shown great promise for use of DBRs as efficient and accurate behavior outcome tools, their primary use has been on evaluating supplemental (i.e., Tier 2), or classroom-wide behavioral interventions. Thus, there is a need to examine the use of DBR as a sound and feasible method for teacher evaluation of behavioral outcomes of students having the most serious problem behaviors.

One potential DBR strategy was used in a recent study (Iovannone et al., 2009) that used a randomized controlled trial to explore the effectiveness of the Prevent-Teach-Reinforce (PTR) model of individualized behavior intervention. In developing the PTR model, it was important to make the process simple enough to be feasible for use in typical school settings. Therefore, when determining the daily data gathering method that would be used by teachers, it was decided that the tool needed to be (a) efficient, so that behavior recording would be quick and nonintrusive, and (b) functional for teachers in interpreting the student outcomes for each targeted behavior. These key considerations resulted in the Individualized Behavior Rating Scale Tool (IBRST). The IBRST uses a 5-point Likert-type scale that teachers use daily across baseline through intervention to rate their perception of the student's performance of the target behaviors including, at a minimum, one problem behavior and one appropriate replacement behavior. While the IBRST has features that fall under the umbrella of DBRs (e.g., person who directly observes the behavior rates its occurrence close to time of performance), it has unique features that differentiate it from other DBRs (e.g., DBRC, DBR-Single Item Scale). Rather than using general behavior categories and global scales of ratings (e.g., 100-mm line divided into 10 intervals with anchor integers between 0 and 10; see Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007), each teacher defined the individual student behaviors of concern targeted for intervention and determined the measurement characteristics to be used for recording including the measurement metric (i.e., frequency, duration, intensity) and the estimates of the targeted behavior being measured for each rating point integer embedded within the 5-point scale.

Given the practical drawbacks of SDO and the increasing need to use sound behavioral assessment methods for monitoring outcomes of students with serious problem behaviors within a problem-solving framework, the purpose of this study was to evaluate the interrater agreement of a data tool that balances the need for repeated measures of behavior outcomes with the limited resources of typical classroom teachers and schools (Steege, Davin, & Hathaway, 2001). This study examined the interrater agreement of the IBRST as recorded by two independent raters in the context of typical school settings and its potential use as a daily behavior outcome measurement tool for students

receiving the most intensive levels of behavior support. In addition, the study examined behavioral dimensions such as measurement scaling type and behavior salience that may impact agreement as well as behaviors for which the IBRST may or may not be a functional tool. In this study, the following research questions were explored:

Research Question 1: To what extent are the IBRST ratings of problem and appropriate behaviors for two independent observers similar?
Research Question 2: How do the dimensions of behavior salience and measurement scaling type impact the interrater agreement of the ratings?
Research Question 3: Does interrater agreement of the ratings remain stable over time?

Method

The unit of analysis in this interrater agreement study was a set of two IBRST data points recorded by two independent raters following an observation of an identified student during a specific routine or period of time. One rater was always one of the student's school-based instructional staff (i.e., teacher, paraeducator), and the other rater was always a data collector from the university research project. Observations were conducted on the behaviors of 19 students enrolled in 13 schools in central Florida. All of the students were participants in a larger study of the efficacy of a behavior intervention model, PTR (Iovannone et al., 2009). During every observational period, each rater provided ratings on individualized scales of the student's identified problem behavior(s) and the student's identified appropriate behavior(s).

Participants and Setting

Student participants and settings. Participants in the study consisted of 19 students recruited from 13 schools (9 elementary, 1 intermediate, 1 middle school, 2 alternative schools) located in three Central Florida school districts. The inclusion criteria were as follows: Students had serious problem behaviors, these students' teachers indicated a willingness to provide additional IBRST data recordings, and the teachers were willing to have observers in their room at multiple time points. Teachers used Stage 1 of the *Systematic Screening of Behavior Disorders* (SSBD; Walker & Severson, 1992) to nominate and rank order one to three students in their classrooms who displayed externalizing problem behavior(s). Caregiver consent for the top-ranked student was sought. If the caregiver of the top-ranked student did not give permission for the student to participate, consent was sought for the second-ranked student, and subsequently the third-ranked student if the second-ranked student's caregiver refused. There were no instances in which a fourth student needed to be identified due to caregivers refusing consent for the three top-ranked students nominated.

The 19 students included in the study ranged in grade levels from kindergarten to seventh grade, with 2 in kindergarten, 2 in first grade, 7 in second grade, 4 in third, 1 in fourth, 2 in fifth, and 1 in seventh. Ages ranged from 5 years 1 month to 13 years 5 months (M = 8 years 1 month). They were predominately boys (n = 16). Twelve of the students were White, 6 were Black, and 1 was Hispanic. Demographic data indicated 14 had Individualized Education Programs (IEPs), with a variety of primary disabilities: 4 with autism spectrum disorders, 4 with emotional/behavioral disorders, 2 with learning disabilities, and 1 each with other health impairment, intellectual disability, speech/language impairment, and visual impairment. Placement data indicated that 13 received their academic reading instruction in special education settings and the remaining 6 in general education.

The time periods for the observations and ratings of each student's target behaviors were determined by each student's teacher and were contingent upon the routine in which problem behavior was highly probable. The wide range of student grade levels and educational placements affected the variation of observation time lengths. For example, some student behavior problems occurred during independent work time routines that for some of the participants (e.g., younger students, students with significant cognitive disabilities) were of 15-min duration while for others (e.g., older students in general education settings) were 60 min in duration. The time lengths of the observations ranged from 15 to 135 min, with a mean of 48.48 min (SD = 23.12). The number of overall observations for each student ranged from 2 to 20, with a mean of 10.4 observations (SD = 5.90), with 38% of the observations occurring during baseline and 62% occurring during post-intervention.

Raters. Raters consisted of 23 school-based instructional personnel and 11 university-based data collectors. Of the 23 school-based instructional personnel, 14 taught in special education classrooms, and 9 taught in general education classrooms. All were White women with instructional experience ranging from 1 year to 17 years (M = 5.67 years, SD = 5.67). Sixteen students had one teacher as their rater, whereas 3 students had multiple school personnel (2–3) who completed ratings. One student's team included a special education teacher and an instructional assistant in a self-contained elementary classroom, the second student's team included three intermediate schoolteachers who taught the student at different periods during the day, and the third included 2 fifth-grade general education teachers who cotaught the student. The 11 university-based data collectors were 10 graduate students in behavioral analysis,

public health, and school psychology and one assistant in research faculty member.

Procedures

Training data collectors. The primary author trained the data collectors. Training included procedural steps for (a) selecting targeted behaviors for intervention, (b) defining behaviors, (c) developing the IBRST scale, and (d) instructing the teacher to use the scale. The data collectors engaged in two or three role-plays with the primary author to practice and demonstrate competencies in following the IBRST procedures. In the first role-play scenario, the primary author modeled the procedural steps for setting up a sample IBRST with the data collector playing a teacher role. In the second role-playing scenario, the primary author and data collector switched roles. If the data collector omitted any procedural steps during the role-play, additional role-plays were scheduled until the data collector accurately completed 100% of the procedural steps. There were no instances in which any data collector required more than one role-play to achieve 100% procedural accuracy.

The school-based team engaged in the process of developing the IBRST as described in the next section (developing the IBRST). After the IBRST was developed, school-based instructional personnel were trained by the data collectors to use the tool for rating behaviors each day. First, the data collector reviewed the behavior definitions and the scale point descriptors with the teacher. Next, the teacher practiced use of the IBRST by circling the scale point that most accurately described the student's performance of the behavior during the previous day's specified time period and verbally explaining why they selected the rating. Once the teacher showed an understanding in using the IBRST, the date for beginning to record daily data was determined. Within a week of the teacher initiating use of the IBRST, data collectors checked in by email, phone call, or a face-to-face meeting to discuss the teacher ratings, confirm the acceptability of the tool and to gauge the functionality of the teacher-estimated scale point measures by inspecting the data point trend trajectory. For both problem and appropriate behaviors, data ratings for the 1st week of baseline were expected to be scored as a "bad day" or a "very bad day." If the teachers rated the behaviors at a less severe level (i.e., "good day," "very good day") the data collector and teacher discussed possible reasons. If it was determined that the scale estimates were over- or underestimates, necessary modifications to the scale were made. Of the 19 IBRSTs developed and used in this study, no scale required modifications.

Developing the IBRST. The DBR tool used in this study consisted of a 5-point Likert-type scale developed by Dunlap et al. (2010) that enabled raters to record their perceptions of the occurrence of specific problem and appropriate behavior targets for each student. Behaviors to be rated were identified by the teacher as the targets of highest concern and included, at a minimum, one problem behavior and one appropriate replacement behavior. A protocol for creating the IBRST was established and consisted of several phases. During Step 2 of the five-step PTR process, a data collector assigned to the team guided the teacher to identify, prioritize, and operationally define the problem behaviors and potential replacement behaviors.

For each prioritized behavior, the data collector asked the teacher a series of questions to determine the points on the IBRST scale. First, to determine the time periods of ratings, the teachers were asked whether they wanted to rate behavior throughout the entire day or during a specific routine in which the behavior was more likely to occur. Second, to determine the appropriate measurement approach, the data collector presented teacher-friendly examples of different strategies (i.e., frequency, duration, intensity) and asked teachers to identify the method that best captured their concern about each behavior's occurrence. For instance, teachers were asked whether they were most concerned about (a) how often the behavior occurred, (b) how long the behavior lasted, (c) how severely the behavior was manifested, or (d) the percentage of the day or routine in which behavior occurred. After teachers indicated the method that best reflected their concerns about the behavior occurrence, the data collector ensured that the measurement approach selected made sense for the behavior typography. For example, if a teacher targeted academic engagement as the appropriate replacement behavior and stated the concern was the number of times the student was engaged during a specific time period, the data collector would ask the teacher for further description of how this would be estimated. If through this description the teacher had difficulty coming up with a description of an appropriate estimate using the selected measurement method, the data collector would then suggest other methods that may more closely match the concern as well as being the best measurement method for the behavior. In the example, the data collector may suggest that other ways for estimation could be the amount of overall time (e.g., minutes) or the overall percentage of time the student performed engaged behaviors.

After determining the measurement approach, the teachers were asked to estimate behavior occurrences for the purpose of setting the scale points for each target behavior. First, the teacher provided an approximation of the behavior's occurrence during a *typical bad day* (i.e., Scale Point 4). Next, the *very bad day* (i.e., Scale Point 5) was defined. The teacher was then asked to identify a *very good day* (i.e., Scale Point 1), or a reasonable goal for behavior performance after intervention implementation. Finally, Points 2 and 3 were identified. This process was repeated for each prioritized behavior targeted for intervention.

Student: Tracy____ School: Maple_____
Rater;__Joe_____

Problem Behavior	Date	02/01/11	02/02/11															
Hitting	>10 times	5 (4)	(5) 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4	5 4
	7-9 times	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	5-7 times	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	3-4 times	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	2 or less																	
Engagement	>60%	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	45-60%	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
	30-44%	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	20-29%	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	<20%	(1)	(1)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 1. Sample IBRST completed for a student with one problem behavior and one appropriate behavior.
Note. Hitting definition: Touching peers or adults with an open hand, fist, foot, or object. Record estimate of number of hitting events each day—5 = very bad day; 4 = typical bad day; 3 = so-so day; 2 = good day; 1 = very good day. Engagement definition: Record estimate of percentage of time engaged during independent work time—eyes on work materials or teacher, pencil moving or hand raised to ask question—1 = very bad day; 2 = typical bad day; 3 = so-so day; 4 = good day; 5 = very good day.

Appropriate behavior scale points were reversed as described previously. Because the teachers were the drivers of the behavior definitions and scale points of the IBRST, diverse scales that included all four measurement strategies (i.e., frequency, duration, intensity, and percentage) resulted. An example of a completed IBRST is shown in Figure 1.

Upon completion of the IBRST, the data collector provided an opportunity for the teacher to practice or rehearse using the scale by asking the teacher to rate each of the behaviors based on the previous school day in which the student was in attendance. For example, the data collector would ask the teacher, "How would you rate the student's academic engagement yesterday?" Once the teacher responded with a rating, the data collector followed up by asking, "Why did you select that rating for the behavior occurrence yesterday?" If the teacher responded showing understanding of the scale (e.g., "Yesterday was a typical day, so I would rate the student's engagement as a 2."), and that the ratings made sense, the teacher then determined the initiation date for starting baseline data collection. If the teacher had any difficulties using the scale to rate the student's behavior in the practice session, the scale was revised until the teacher indicated accurate use of the ratings.

Data Collection Procedures

Data collectors were assigned to schools based on geographic proximity to their home addresses. The data collectors scheduled observation times to occur during the routines or classes in which the teacher indicated that there was a high likelihood of problem behavior taking place. Upon arriving at the classroom, the data collector checked in with the teacher, ensured the student was present that day, and gave the teacher a clean paper copy of the IBRST developed for the student which included identifying codes for the teacher and student, the date and time range of the observational period, and a blank envelope. This specific IBRST measurement was in addition to the daily IBRST data collection required by the teacher as a participant of the PTR randomized controlled trial. The teacher was instructed to complete the IBRST provided by the data collector at the end of the observational period, place the IBRST in the envelope, and seal the envelope prior to giving it to the data collector. The data collector observed the student and at the end of the observational period, rated the student's behavior using a clean copy of the IBRST that included identifying codes for the teacher, student, and the project staff along with the date and observational time period. Similar to the teachers, the project staff placed their completed IBRST

form in a separate blank envelope and sealed it. The data collector gave both sealed envelopes to the primary author for entry into PASW Statistics 18 database. For the purpose of rating consistency, PASW data entry of ratings for problem behavior was recoded to match the ratings of appropriate behaviors. That is, for purposes of statistical analyses a rating of 5 represented a very good day, a rating of 4 represented a good day, a rating of 2 represented a typical bad day, and a rating of 1 represented a very bad day for both problem and appropriate behaviors.

Data Analysis

Demographic and IBRST rating data were initially entered into PASW Statistics 18 database. Interrater agreement and related statistical tests were conducted in S-PLUS v8.10 and R v2.12.2 with the Psych package. All significance testing was conducted at the .05 probability level, and no adjustments were made for multiple comparisons as all tests were designated a priori.

To answer the first two research questions (i.e., determine the extent to which the IBRST ratings of problem and appropriate behaviors for two independent observers were similar overall and by facets of salience of behaviors and measurement type), interrater agreement between the independent observations of teachers and data collectors was calculated using Cohen's linear-weighted (LW) and quadratic-weighted (QW) kappa coefficients (K_{lw}; Cicchetti & Allison, 1971, K_{qw}; Cohen, 1968, Fleiss & Cohen, 1973). Weighted kappas have been considered a better measure of agreement than a simple percentage of agreement score as kappa statistics take into account agreement that would be expected purely by chance alone. Among the family of kappa coefficients, for ordinal scales as used in this study, weighted kappas have been preferred over unweighted kappas, as unweighted kappa treats all rater discrepancies as equal disagreements, whereas weighted kappas consider ordinality in rater disagreements. The exact weights used for weighted kappa are arbitrary, and both linear (K_{lw}; Cicchetti & Allison, 1971) and quadratic (K_{qw}; Cohen, 1968; Fleiss & Cohen, 1973) weights have been commonly used. Among weighted kappas, K_{lw} penalizes small discrepancies between raters more heavily than K_{qw} and produces more conservative estimates of agreement. However, K_{qw} has the useful property of being equivalent to the Pearson product–moment correlation (r), a measure of association or reliability; when the marginal distributions of the judges' ratings are the same, and when marginals are unequal, it produces attenuated values of kappa that provide more conservative estimates of agreement (Schuster, 2004).

In addition, with sample sizes greater than 25, as in this study, K_{qw} provides a ready comparison for rater judgments that use quantitative scales because the interclass correlation (ICC), a widely used reliability measure for interval data (Cicchetti et al., 2006), is a special case of K_{qw} when the categories are equally spaced. As the value of weighted kappa can vary considerably depending on which type of weight has been applied (Graham & Jackson, 1993; Vanbelle & Albert, 2009), in practice, standards for interpreting the strength of agreement have not made a distinction between whether K_{lw} or K_{qw} has been reported. For this reason, we report both forms of weighted kappa. All kappa coefficients are scaled from −1.00 to 1.00. A kappa value of 0.00 indicates no agreement other than what would be expected by chance alone with values of 1.00 indicating perfect agreement and −1.00 indicating perfect disagreement. Proposed standards for interpreting kappa statistics include < .01 to 0 = poor, .01 to .20 = slight, .21 to .40 = fair, .41 to .60 = moderate, .61 to .80 = substantial, and .81 to 1.00 = almost perfect (Landis & Koch, 1977). A similar standard has been proposed by Cicchetti and Sparrow (1981) with values of .75 and above considered to indicate excellent agreement and values between .60 and .74 to indicate good agreement. Horner et al. (2005) suggested .60 as a minimal standard for single-subject research.

To determine the degree that agreement remained stable over time, weighted kappas were calculated at baseline and post-intervention for each type of behavior and overall. An approximate permutation test, also known as Monte Carlo permutation tests or random permutation tests (Dwass, 1957), was conducted to assess the significance of the obtained difference between the baseline and posttest kappas. Ten thousand random permutations were generated for each test to simulate the approximate distribution of the difference of the two kappa's under the null hypothesis and thus to acquire the two-sided probability value of the obtained kappa difference.

The salience of the behavioral targets was established by having two of the authors independently read the definitions for each of the 66 behaviors included on the 19 IBRSTs and judge whether it was *more or less salient.* A behavior was judged to be more salient if it was discrete (i.e., had a clear beginning and ending) and had the ability of being counted as a discrete unit. Examples of more salient target behaviors included frequency of hitting, performing self-injurious behaviors, or shouting out cuss words. A behavior was considered less salient if it was continuous (i.e., no clear beginning and ending) and did not have the ability to be counted as a discrete unit (e.g., percentage of time, intensity). Examples of less salient behaviors included duration of academic engagement, tantrums, or intensity of self-injurious behaviors. The two authors then compared their scores and for any in which there was a disagreement (n = 18) came to a consensus.

Results

Descriptives

Table 1 shows the titles of the problem and appropriate behaviors identified and defined by the teachers of each of

Table 1. Behavior Categories, Verbatim Observed Behavior Titles, and Measurement Types.

		Scale *n*				
Category	Verbatim Title of Target Behavior	F	D	I	%	Total
Problem behaviors						
Aggressive	Aggression	5	1	0	0	6
	Inappropriate touching of others	3	0	0	0	3
	Hitting	3	0	0	0	3
	Eye poking	1	1	0	0	2
	Personal space	1	0	0	0	1
	Grabbing	0	0	1	0	1
	Spitting	0	0	0	1	1
	Self-injury	1	0	0	0	1
	Aggressive posturing	0	0	0	1	1
	Horseplay	1	0	0	0	1
	Total	15	2	1	2	20
Disruptive	Out of area/seat	4	0	0	1	5
	Profanity/inappropriate language	4	0	0	1	5
	Disruptions	1	0	0	2	3
	Behavior outbursts/tantrums	2	0	0	0	2
	Screaming	1	0	0	0	1
	Yelling out	1	0	0	0	1
	Total	13	0	0	3	16
Appropriate behaviors						
Academic	Engaged/on task	2	4	0	7	13
	Task/work completion	0	0	0	5	5
	Compliance	0	0	1	1	2
	Total	2	4	1	13	20
Social	Appropriate social interactions	3	0	0	2	5
	Appropriate use of hands	0	0	0	1	1
	Total	3	0	0	3	6
Communicative	Expresses wants, needs, frustrations	1	0	0	4	5
	Total	1	0	0	4	5
Total—All behavior titles		34	6	2	25	67

Note. F = frequency (the number of behavioral events); D = duration (length of time of behavioral event); I = intensity (the severity of the behavioral event; % = percentage of the day or percentage of opportunities behavioral event occurred. Scale *n* indicates the total number of times specific behavior appeared on student Individualized Behavior Rating Scale Tools (IBRSTs).

the 19 students, the number of times the behavior title appeared among the 19 students, and the types of measurements used for rating each of the behaviors. The Problem Behavior titles of Aggression ($n = 6$) and Out of Area ($n = 5$) were the most frequently identified behaviors of concern, while Academic Engagement ($n = 13$) was the appropriate behavior most often selected as the desired replacement skill. Frequency was the measure chosen most often for rating selected behavior titles ($n = 18$), with percentage being selected for seven behavior titles. Duration and intensity ratings were used least often ($n = 3$ for each).

Interrater agreement. K_{lw} and K_{qw} were calculated for the four types of teacher-selected behavior ratings, the primary and secondary problem behaviors, and the primary and secondary appropriate behaviors.

Problem behavior ratings. For the primary problem behavior ratings, 105 pairs of observations were available. As shown in Table 2, $K_{qw} = .82$ and $K_{lw} = .65$. Among the 105 rating pairs, 66% of the ratings were agreements. Among disagreements ($n = 36$), the vast majority (81%, $n = 29$) varied by a single category on the 5-point rating scale, with the remaining disagreements ($n = 7$) varying by only two scale points. A similar pattern of agreement was found for the second problem behavior selected. Because not all teachers selected a second problem behavior for each child, only 90 pairs of observations were available for these ratings. K_{qw} for these observations was .77 and $K_{lw} = .59$. Agreements accounted for 58% ($n = 52$) of the ratings. Among disagreements, 74% ($n = 28$) differed by a single scale point, with the remainder ($n = 10$) reflecting a difference of two scale points. Combining both the primary and secondary

Table 2. Weighted Kappa of Problem and Appropriate Behaviors Overall and Behavior Salience.

Variable	*n*	Kappa QW	95% CI	Kappa LW	95% CI	% Agreement
Problem Behavior 1	105	.82	[.75, .89]	.65	[.54, .75]	.66
More salient	70	.82	[.72, .91]	.66	[.53, .75]	
Less salient	35	.82	[.71, .93]	.63	[.47, .78]	
Problem Behavior 2	90	.77	[.68, .87]	.59	[.48, .71]	.58
More salient	65	.83	[.74, .91]	.64	[.51, .76]	
Less salient	19	.45	[.11, .78]	.32	[.01, .62]	
Problem behaviors combined	195	.80	[.74, .86]	.62	[.55, .70]	
Appropriate Behavior 1	103	.65	[.50, .79]	.53	[.41, .69]	.46
More salient	42	.68	[.50, .87]	.51	[.44, .68]	
Less salient	61	.63	[.43, .82]	.53	[.38, .69]	
Appropriate Behavior 2	56	.76	[.62, .89]	.64	[.50, .77]	.59
More salient	38	.92	[.86, .98]	.81	[.71, .91]	
Less salient	18	.43	[.14, .72]	.74	[.11, .59]	
Appropriate behaviors combined	159	.69	[.58, .79]	.56	[.47, .65]	

Note. QW = quadric-weighted; CI = confidence interval; LW = linear-weighted.

Table 3. Weighted Kappa by Type of Measurement: All Behaviors Combined.

Measurement Type	*n*	Kappa QW	95% CI	Kappa LW	95% CI
Frequency	164	.76	[.69, .83]	.59	[.50, .67]
Duration	33	.78	[.63, .92]	.62	[.46, .78]
Intensity	19	.79	[.63, .94]	.59	[.39, .78]
Percentage	132	.72	[.60, .83]	.61	[.51, .70]

Note. QW = quadric-weighted; CI = confidence interval; LW = linear-weighted.

problem behavior ratings, an overall K_{qw} = .80 and K_{lw} = .62.

Appropriate behavior ratings. For the primary appropriate behavior, 103 pairs of observations were available with K_{qw} = .65 and K_{lw} = .53. Among the ratings, 46% (n = 47) were agreements. The frequency distribution of disagreements, which for these observations were differences that ranged from one to four scale points, was as follows: one scale point, 75% (n = 42); two scale points, 14% (n = 8); three scale points, 5% (n = 3); and four scale points, 5% (n = 3). For the secondary appropriate behavior, only 56 pairs of observations were available. K_{qw} = .76 and K_{lw} = .64. Fifty-nine percent (n = 33) of the ratings were agreements. Disagreements ranged from one to three scale point differences, with 61% (n = 14) of the disagreements varying by one scale point, 30% (n = 7) by two scale points, and 9% (n = 2) by three scale points. Combining the primary and secondary appropriate behavior ratings, the overall K_{qw} = .69 and K_{lw} = .56 (see Table 2).

Behavior salience and scale type. When the overall observations were categorized into behaviors that were more (n = 215) or less (n = 133) salient, K_{qw} indicated acceptable agreement for both categories, but as might be expected was higher for the more salient behavior observations, K_{qw} = .83 versus .65. As shown in Table 2, a similar pattern was found when K_{qw} was calculated within each of the behavior types; however, K_{qw} values for the less salient Problem Behavior 2 and Appropriate Behavior 2 were below the .60 level (.45 and .43, respectively). Table 3 displays K_{qw} for the different types of measurement when levels of salience for all behaviors and by behavior type (problem and appropriate behavior) were combined. As shown in Table 3, when agreements were broken out by how the behavior was scaled, there were few differences, with K_{lw} ranging from .59 (i.e., intensity and frequency) to .62 (duration), and K_{qw} ranging from .72 for percentage ratings to .79 for intensity ratings.

Agreement across time. K_{qw} and K_{lw} were calculated at baseline and post-intervention for each type of behavior and overall. As shown in Table 4, results from the Monte Carlo permutation tests of the difference between either the mean K_{qw} or K_{lw} at the two time points indicated that none of the comparisons was significantly different (ps > .05).

Table 4. Weighted Kappa by Time of Observation: Baseline vs. Post-Test.

Variable	*n*	Kappa QW	95% CI	*p*[a]	Kappa LW	95% CI	*p*[a]
Problem Behavior 1				.39			.41
Baseline	36	.84	[.74, .95]		.68	[.53, .83]	
Posttest	69	.78	[.67, .89]		.59	[.46, .73]	
Problem Behavior 2				.60			.46
Baseline	35	.75	[.62, .89]		.59	[.41, .76]	
Posttest	55	.69	[.51, .81]		.49	[.29, .70]	
Appropriate Behavior 1				.74			.33
Baseline	39	.58	[.32, .83]		.42	[.21, .62]	
Posttest	64	.63	[.43, .84]		.54	[.38, .70]	
Appropriate Behavior 2				.26			.12
Baseline	24	.82	[.65, .99]		.74	[.56, .93]	
Posttest	32	.66	[.45, .88]		.52	[.31, .72]	
Problem Behaviors 1 and 2 (combined)				.38			.28
Baseline	71	.80	[.72, .89]		.64	[.53, .75]	
Posttest	124	.74	[.65, .84]		.55	[.44, .67]	
Appropriate Behaviors 1 and 2 (combined)				.66			.75
Baseline	63	.69	[.53, .85]		.56	[.42, .70]	
Posttest	96	.64	[.49, .80]		.53	[.40, .66]	
All behaviors (combined)				.41			.40
Baseline	134	.77	[.69 .85]		.62	[.54, .71]	
Posttest	220	.72	[.63, .81]		.58	[.49, .66]	

Note. QW = quadric-weighted; CI = confidence interval; LW = linear-weighted.
[a]We used permutation-based nonparametric method to test the significance of difference of kappa over time.

Discussion

The primary purpose of this study was to evaluate the inter-rater agreement of an individualized behavior outcome tool that can be practical for use by teachers in rating daily occurrence of target problem and appropriate behaviors of students presenting with serious problem behaviors requiring individualized interventions. The study examined inter-rater agreement of the IBRST ratings as completed by two independent raters, a classroom teacher and a data collector. This study used data collected from 19 of the 245 students who participated in the primary PTR randomized controlled trial (Iovannone et al., 2009). The results of the current study indicate that the IBRST has the potential of being a feasible and reliable repeated-measures behavior assessment outcome tool to be used by typical classroom teachers in day-to-day school settings. The tool's properties showed suitability for classroom use across time and for behaviors that are more salient. The kappa coefficients obtained across problem and appropriate behaviors on a 5-point Likert-type scale with measurement ranges defined by teachers reached moderate to substantial levels as suggested by Landis and Koch (1977) and Horner et al. (2005). Furthermore, agreement remained adequate across all four types of measurement scales. Nevertheless, some heterogeneity in rater agreement was observed. When a behavior's salience was examined, agreement was only fair for some of the less salient behaviors; however, the number of observations for these ratings were few (i.e., less than 20) with large confidence intervals that encompassed moderate to substantial agreement. The tool shows promise of stability as evidenced by the consistency of kappa coefficients from baseline to post-intervention.

The kappa coefficients for problem behaviors were slightly higher than for appropriate behaviors. One explanation may be that teachers were more accustomed and alert to the occurrence of problem behaviors that were more salient (obvious) than appropriate behaviors, which may have contributed to greater agreement in the perception of behavior occurrence (Chafouleas, Sanetti, Jaffery, & Fallon, 2012). Appropriate behavior, on the other hand, may not prompt the focused attention from teachers that problem behaviors produce resulting in inexact estimations or less accuracy in recognizing when the behavior is performed (Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012). Furthermore, appropriate behavior definitions tend to be less discrete than problem behavior definitions. That is, the majority of the problem behaviors identified by teachers in this study had clear and observable beginnings and endings (e.g., hitting, cursing, elopement) making scoring more objective while the majority of appropriate behaviors had less clear beginnings and endings (e.g., engagement,

independent work), thus lending themselves to more subjectivity in measurement. The difficulty in measuring less discrete behaviors is not unique to the IBRST.

In addition, the authors hypothesized that the behaviors of highest concern (e.g., Problem Behavior 1, Appropriate Behavior 1) would have higher coefficients than the second-ranked behaviors. While this was supported by the problem behavior coefficients, the reverse occurred for the appropriate behaviors with the second-ranked behavior having higher agreement between raters than the top-ranked behavior. Upon closer inspection of the data, three instances in which the two raters were at the extreme opposites on the scale (i.e., ratings of 1 and 5) were noted. It is unclear whether the ratings were polarized due to different perceptions of behavior occurrence or due to rater error. In teacher use of the IBRST, appropriate behavior measurement points were the reverse of problem behavior points; that is, a 5 rating represented the best day for an appropriate behavior but denoted the worst day for problem behavior. All of the IBRSTs consisted of one hard-copy paper page that included problem and appropriate behaviors sequentially listed in rows with the problem behaviors appearing on the top one or two rows and the appropriate behaviors immediately following. It is possible that in these three isolated instances, one of the raters continued to use the problem behavior rating hierarchy in estimating the appropriate behavior occurrence. Although there were very few cases in which the ratings were at the opposite ends, these three specific ratings contributed to the lower kappa coefficient for appropriate behavior one.

Another encouraging outcome is the consistency of the kappa coefficients when comparing interrater agreement by the measurement scale selected (e.g., frequency, duration, intensity, percentage) as well as over time (baseline to post-intervention). Although it was hypothesized that percentage ratings would have lower interrater agreement than frequency or duration ratings, it was not supported by the results of this study. Teachers were the main drivers of the IBRST development and by providing them with a standardized process to identify and define the top behaviors of concern and determine the most feasible and accurate measurement for daily rating of behavior, yielded a measurement tool that was functional for teacher use.

The salience of the target behaviors measured showed that behaviors that were more salient (e.g., discrete, able to be counted) had higher rates of agreement between two raters than behaviors that were less salient. These initial findings may suggest that the IBRST is a more appropriate data tool for use with discrete behaviors such as hitting, cussing, and making appropriate social comments than with behaviors that are of a continuous, less salient nature such as academic engagement, appropriate social behavior, or intensity of tantrums.

A question arises as to why in this study the kappa QW and the kappa LW differ. As demonstrated by Brenner and Kliebsch (1996), when the number of categories increases, kappa QW increases, approaching the correlation coefficient for the underlying continuous measures, and kappa LW slightly declines, thus the difference between the two increases. The size of the difference also depends on the magnitude of the correlation. For example, when there are five categories and the true correlation is .80, the difference between kappa LW and kappa QW is approximately .18, when the correlation is .50, the difference is about .08. This is in line with our findings.

Which weighting scheme is to be preferred? If the measured variable is believed to be truly continuous in nature, but the measurement scale has been ordinally scaled, then the kappa QW would be most appropriate. On the other hand, kappa LW is less affected by the number of categories and might be useful if the researcher wants to compare kappas between items with different numbers of categories or arbitrary item cutpoints. Graham and Jackson (1993) also pointed out that the choice of weighting schemes can greatly influence the estimated value of the statistic and the weighted kappa statistic is not always sensitive to differences in the observed proportion of exact agreement and that high values of weighted kappa can be observed even when the level of agreement is low. That said, presenting all three, kappa QW, kappa LW, and proportion of exact agreement, may give us a more complete picture in understanding interrater agreement of an ordinal scale.

Limitations and Future Research

Efforts were made to identify and control for potential variables that would impact the study design. However, there are several limitations that warrant mention. First, the sample size of 19 students used in this study is small. This limitation is present in previous studies exploring technical adequacy of DBR measures. For example, Burke et al. (2009) used a sample of 7 elementary-age students with emotional and behavior disorders while Chafouleas et al. (2007) conducted a study with 15 preschool students. Although the sample size of 19 students is small for traditional psychometric studies, the student participants were diverse in age, placement, and disability classification. In addition, the sample was a subset of a larger, randomly assigned sample, and the behaviors examined in this study are representative of serious problem behaviors that require the most intensive levels of support. It would behoove the field to conduct a large-scale reliability and validity study exploring the use of tools such as the IBRST and identify variables that impact their use. Many questions remain to be answered or further explored such as the behaviors for which the tool is best suited, the number of rating points

needed for sensitivity to behavior change, the impact of training on teacher reliability and validity, as well as the process teachers engage in when deciding upon the most accurate rating score of student behavior.

Second, the study did not examine concurrent validity of the IBRST. Technically adequate agreement, as shown in this study, is necessary for determining a tool's appropriateness for measuring a specific variable of interest. However, once this is established, it is essential to know whether the rating of the behavior on the IBRST correlates with data from SDOs taken during the same time periods. Future research is needed to examine whether the IBRST ratings accurately reflect actual behavior events occurring, or at a minimum, reflect an accurate trend direction of behavior change. Further research may also explore whether using the IBRST is sensitive to behavior change across different phases of intervention.

Third, the training of the data collectors to set up the IBRST did not include interrater reliability of the performance of the steps or on data collectors scoring of the IBRSTs. Although the training was standardized so that it was similarly presented to each data collector, the study would have been strengthened by reporting the percentage of steps completed correctly by the data collector as well as interrater agreement on the number of steps. While the team considered collecting interrater agreement of data collectors' IBRSTs, it was omitted to decrease the intrusive nature of having more than one additional person in the teacher's instruction of the classroom during instruction as well as the potential impact of being a contextual variable that may modify the student's performance of target behaviors. Future research should include obtaining the interrater agreement of data collectors' IBRST scores through use of videos or other nonintrusive methods.

Another limitation is the variation of the time lengths of the observations as well as the diverse number of observations per student. This study was part of a larger randomized controlled trial that tested an intervention within the context of the daily school setting in which behavior problems of students occurred in a variety of routines that were not standardized in length without an introduction of researcher control. Thus, this study represents the use of the IBRST within typical school days of typical school practitioners. However, future research efforts may want to explore how the interrater coefficients are impacted if more experimenter control is established. This could include having standardized lengths of routines for using the IBRST as well as a consistent number of observations for each student.

The IBRST was developed as an efficient, feasible, teacher-friendly method of obtaining repeated measures of target behaviors. An important data outcome that was not measured in this study was the teachers' social validity ratings of the IBRST. It would be important to determine whether the primary consumer of the tool found it to be acceptable and effective. The randomized controlled trial from which the participants were recruited did report social validity ratings of the entire PTR process from all of the teachers assigned to the intervention group. Using a modified version of the 5-point *Treatment Acceptability Rating Form* (TARF; Reimers & Wacker, 1988) a mean social validity score of 4.20 (SD = .52) was reported by 124 teachers (Iovannone et al., 2009) indicating high to very high acceptance of the PTR process. It would be import for future research, however, to evaluate the specific social validity of the IBRST and possibly compare it with acceptance of SDO or other data measurement procedures.

Implications for Practice

It is important to develop a behavioral data outcome that has technically adequate agreement between raters and that is feasible for daily (i.e., repeated measure) use by teachers in the contexts of typical school settings. This study initiated an exploration of the interrater agreement of a specific data tool with DBR features that was used by teachers within a randomized controlled trial exploring the efficacy of PTR, an individualized behavior intervention process. It is important to note, however, that interrater agreement analyses evaluates the extent of the *exact or perfect* agreement between two independent judges. Data results highlighted in this study showed that the majority of inexact agreements between the two raters were only one point suggesting that there was agreement in whether the behavior occurrence was perceived as a *bad day* or a *great day*. In only three instances were the two ratings at polar opposites (i.e., 1 and a 5).

While establishment of interrater agreement of a scale is vital, the primary use of the IBRST in collecting behavior data from baseline through intervention and making data-based decisions about intervention strategies within a problem-solving framework may not require the standard of perfect agreement to be a useful tool for teachers. For practitioners in applied settings, it may be more important to determine whether two independent observers using the IBRST agree in their perceptions of the trend or direction of behavior change. The most important feature of any data collected is whether the data are able to be used to make sound instructional decisions. Therefore, an important question for the field in using the IBRST is whether it is more important for both observer ratings to match or for both observer ratings to move in the same direction across time and whether those changes in trends result in appropriate modifications to behavior interventions so that students meet behavioral goals. The IBRST was used daily by typical teachers in daily classroom contexts. The data from this study that were the focus of the article only report the

agreements when two observers were present in the classroom. However, the teachers, as part of the PTR randomized controlled trial, used the IBRST every day (*M* length = 71 days) from baseline through intervention without the need for any researcher prompt (Iovannone et al., 2009). This indicates that the IBRST has a high potential of adoption for use by teachers in school settings.

In conclusion, the present study adds to the emerging literature base examining the use of DBR methods to efficiently and reliably evaluate the behavior outcomes of students needing individualized, intensive behavior interventions. Interrater reliability methods were used to examine the level of agreement between two independent judges, the teacher and a research data collector. Results obtained suggest that the IBRST is a tool that teachers can learn to use reliably and is feasible for use in typical school situations for evaluating student behaviors and response to receiving individualized, intensive levels of behavioral support. The outcomes support future research to further establish the IBRST and other similar instruments as potential behavior assessment data tool.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported by the U.D. Department of Education, National Center for Special Education Research (Grant No. H324P04003) and in part by OSEP TA Center on PBIS III (Grant No. H326S080003) and Institute of Education Sciences, U.S. Department of Education through Grant R324A120097 to the University of Nevada, Reno. Opinions expressed herein are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Batsche, G., Elliott, J., Graden, J., Grimes, J., Kovaleski, J. F., Prasse, D., & Tilly, D. W., III. (2005). *Response to intervention: Policy considerations and implementation*. Alexandria, VA: National Association of State Directors of Special Education.

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted Kappa coefficients on the number of categories. *Epidemiology*, *7*, 199–202.

Burke, M. D., Vannest, K., Davis, J., Davis, C., & Parker, R. (2009). Reliability of frequent retrospective behavior ratings for elementary school students with EBD. *Behavioral Disorders*, *34*, 212–222.

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A. M. (2007). Generalizability and dependability of Direct Behavior Ratings to assess social behavior of preschoolers. *School Psychology Review*, *36*, 63–79.

Chafouleas, S. M., Kilgus, S. P., Riley-Tillman, T. C., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of direct behavior ratings. *Journal of School Psychology*, *50*, 317–334.

Chafouleas, S. M., McDougal, J. L., Riley-Tillman, T. C., Panahon, C. J., & Hilt, A. M. (2005). What do daily behavior report cards (DBRCs) measure? An initial comparison of DBRCs with direct observation for off-task behavior. *Psychology in the Schools, 42*, 669–676. doi:10.1002/pits.20102

Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2009). Direct behavior rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention*, *34*, 195–200.

Chafouleas, S. M., Riley-Tillman, T. C., & McDougal, J. L. (2002). Good, bad, or in-between: How does the daily behavior report card rate? *Psychology in the Schools, 39*, 157–169. doi:10.1002/pits.10027

Chafouleas, S. M., Riley-Tillman, T. C., & Sassu, K. A. (2006). Acceptability and reported use of daily behavior report cards among teachers. *Journal of Positive Behavior Interventions*, *8*, 174–182.

Chafouleas, S. M., Sanetti, L. M. H., Jaffery, R., & Fallon, L. M. (2012). An evaluation of a classwide intervention package involving self-management and a group contingency on classroom behavior of middle school students. *Journal of Behavior Education*, *21*, 34–57.

Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. (2010). School-based behavioral assessment within problem-solving models: Current status and future directions. *School Psychology Review*, *34*, 343–349.

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S., & Jaffery, R. (2011). Direct behavior rating: An evaluation of alternate definitions to assess classroom behaviors. *School Psychology Review*, *40*, 181–199.

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., & Tyrer, P. (2006). Rating scales, scales of measurement, issues of reliability: Resolving some critical issues for clinicians and researchers. *The Journal of Nervous and Mental Disease, 194*, 557-564.

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*, 101–110.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*, 127–137.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provisions for scaled disagreement or partial credit. *Psychosocial Bulletin*, *70*, 213–220.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.

Crone, D. A., Hawken, L. S., & Horner, R. H. (2010). *Responding to problem behavior in schools: The behavior education program* (2nd ed.). New York, NY: Guilford.

Crone, D. A., & Horner, R. H. (2003). *Building positive behavior support systems in schools: Functional behavioral assessment*. New York, NY: Guilford.

Deno, S. L. (1989). Curriculum-based measurement and alternative special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based*

measurement: Assessing special children (pp. 1–17). New York, NY: Guilford.

Deno, S. L. (1995). School psychologist as problem solver. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (Vol. 3, pp. 471–484). Washington, DC: National Association of School Psychologists.

Dunlap, G., Iovannone, R., Kincaid, D., Wilson, K., Christiansen, K., Strain, P., & English, C. (2010). *Prevent-teach-reinforce: A school-based model of individualized positive behavior support*. Baltimore, MD: Paul H. Brookes.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, *28*, 181–187.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Boston, MA: Pearson.

Graham, P., & Jackson, R. (1993). The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology*, *48*, 1055–1062.

Gresham, F. M. (2003). Establishing the technical adequacy of functional behavioral assessment: Conceptual and measurement challenges. *Behavioral Disorders*, *28*, 282–298.

Gresham, F. M. (2004). Current status and future directions of school-based behavioral interventions. *School Psychology Review*, *33*, 326–343.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*, 258–270.

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*, 165–179.

Iovannone, R., Greenbaum, P. E., Wang, W., Kincaid, D., Dunlap, G., & Strain, P. (2009). Randomized controlled trial of a tertiary behavior intervention for students with problem behaviors: Preliminary outcomes. *Journal of Emotional and Behavioral Disorders*, *17*, 213–225.

Jones, C., Caravaca, L., Cizek, S., Horner, R. H., & Vincent, C. G. (2006). Culturally responsive schoolwide positive behavior support: A case study in one school with a high proportion of Native American students. *Multiple Voices*, *9*, 108–119.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Reimers, T., & Wacker, D. (1988). Parents' ratings of the acceptability of behavioral treatment recommendations made in an outpatient clinic: A preliminary analysis of the influence of treatment effectiveness. *Behavioral Disorders*, *14*, 7–15.

Repp, A. C., & Horner, R. H. (Eds.). (1999). *Functional analysis of problem behavior: From effective assessment to effective support*. Belmont, CA: Wadsworth.

Riley-Tillman, T. C., Chafouleas, S. M., & Briesch, A. M. (2007). A school practitioner's guide to using Daily Behavior Report Cards to monitor interventions. *Psychology in the Schools*, *44*, 77–89.

Riley-Tillman, T. C., Chafouleas, S. M., Briesch, A. M., & Eckert, T. L. (2008). Daily behavior report cards and systematic direct observation: An investigation of the acceptability, reported training and use, and decision reliability among school psychologists. *Journal of Behavioral Education, 17*, 313–327. doi:10.1007/s10864-008-0070-5

Riley-Tillman, T. C., Methe, S. A., & Weegar, K. (2009). Examining the use of direct behavior rating methodology on classwide formative assessment: A case study. *Assessment for Effective Intervention*, *34*, 242–250.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment* (10th ed.). Houston, TX: Houghton Mifflin.

Schlientz, M. D., Riley-Tillman, T. C., Briesch, A. M., Walcott, C. M., & Chafouleas, S. M. (2009). The impact of training on the accuracy of direct behavior ratings (DBR). *School Psychology Quarter*, *24*, 73–83.

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*, 243–253.

Shapiro, E. S., & Eckert, T. L. (1994). Acceptability of curriculum-based assessment by school psychologists. *Journal of School Psychology*, *32*, 167–183.

Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, *41*, 551–561.

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children*. New York, NY: Guilford.

Steege, M. W., Davin, T., & Hathaway, M. (2001).Reliability and accuracy of a performance-based behavioral recording procedure. *School Psychology Review*, *30*, 252–261.

Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*, 157–163.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children*, *30*, 16–19.

Walker, H. M., & Severson, H. H. (1992). *Systematic screening for behavior disorders (SSBD): User's guide and technical manual*. Longmont, CO: Sopris West.