Reliability of Direct Behavior Ratings â□□ Social Competence (DBR-SC) Data: How Many Ratings Are Necessary?

Article *in* School Psychology Quarterly · January 2015

DOI: 10.1037/spq0000128

CITATIONS
7

READS
239

5 authors, including:

Stephen Kilgus
University of Wisconsin–Madison
73 PUBLICATIONS **905** CITATIONS
SEE PROFILE

T. Chris Riley-Tillman
University of Missouri
92 PUBLICATIONS **2,090** CITATIONS
SEE PROFILE

Janine Peck Stichter
University of Missouri
114 PUBLICATIONS **1,914** CITATIONS
SEE PROFILE

Alexander M. Schoemann
East Carolina University
39 PUBLICATIONS **1,016** CITATIONS
SEE PROFILE

School Psychology Quarterly

Reliability of Direct Behavior Ratings – Social Competence (DBR-SC) Data: How Many Ratings Are Necessary?

Stephen P. Kilgus, T. Chris Riley-Tillman, Janine P. Stichter, Alexander M. Schoemann, and Katie Bellesheim

Online First Publication, November 2, 2015. http://dx.doi.org/10.1037/spq0000128

School Psychology Quarterly

1045-3830/15/$12.00 http://dx.doi.org/10.1037/spq0000128

Reliability of Direct Behavior Ratings – Social Competence (DBR-SC) Data: How Many Ratings Are Necessary?

Stephen P. Kilgus,
T. Chris Riley-Tillman,
and Janine P. Stichter
University of Missouri

Alexander M. Schoemann
East Carolina University

Katie Bellesheim
University of Missouri

The purpose of this investigation was to evaluate the reliability of Direct Behavior Ratings—Social Competence (DBR-SC) ratings. Participants included 60 students identified as possessing deficits in social competence, as well as their 23 classroom teachers. Teachers used DBR-SC to complete ratings of 5 student behaviors within the general education setting on a daily basis across approximately 5 months. During this time, each student was assigned to 1 of 2 intervention conditions, including the Social Competence Intervention–Adolescent (SCI-A) and a business-as-usual (BAU) intervention. Ratings were collected across 3 intervention phases, including pre-, mid-, and postintervention. Results suggested DBR-SC ratings were highly consistent across time within each student, with reliability coefficients predominantly falling in the .80 and .90 ranges. Findings further indicated such levels of reliability could be achieved with only a small number of ratings, with estimates varying between 2 and 10 data points. Group comparison analyses further suggested the reliability of DBR-SC ratings increased over time, such that student behavior became more consistent throughout the intervention period. Furthermore, analyses revealed that for 2 of the 5 DBR-SC behavior targets, the increase in reliability over time was moderated by intervention grouping, with students receiving SCI-A demonstrating greater increases in reliability relative to those in the BAU group. Limitations of the investigation as well as directions for future research are discussed herein.

Keywords: direct behavior rating, progress monitoring, rating scale, social competence

A common concern within many school settings pertains to student social competence, defined as the extent to which student use of social skills results in successful completion of social tasks, resulting in positive adult and peer interactions (Gresham, 1986; Gresham, Sugai, & Horner, 2001; McFall, 1982). An estimated 1 in 68 eight-year-old children is now diagnosed with Autism Spectrum Disorder (Center for Disease Control, 2014), a condition principally defined by social impairments (American Psychiatric Association, 2013). Social concerns are also prevalent within the typically developing student population, with nearly 1 in 5 school-age children and adolescents exhibiting moderate levels of social skill deficits necessitating intervention (Gresham, Elliott, & Kettler, 2010).

Recognition of such commonality in social difficulties has resulted in the development of numerous interventions intended to promote social competence. Taken together, work regard-

Editor's Note. Matthew Mayer served as the action editor for this article.—SRJ

Stephen P. Kilgus and T. Chris Riley-Tillman, Department of Educational, School, and Counseling Psychology, University of Missouri; Janine P. Stichter, Department of Special Education, University of Missouri; Alexander M. Schoemann, Department of Psychology, East Carolina University; Katie Bellesheim, Department of Psychological Sciences, University of Missouri.

Correspondence concerning this article should be addressed to Stephen P. Kilgus, Department of Educational, School, and Counseling Psychology, 16 Hill Hall, University of Missouri, Columbia, MO 65211. E-mail: kilguss@missouri.edu

ing interventions targeting student social difficulties has proliferated. Review of the What Works Clearinghouse report regarding social skills training for children with disabilities suggested that although the extent of studies meeting evidentiary standards is somewhat limited, social skills interventions have proven effective at supporting student behavior and social-emotional development (U. S. Department of Education, Institute of Education Sciences, 2013). In contrast, work is limited regarding assessment tools that might be used to evaluate student response to social competence interventions. Though social skills assessment methods have been established, each lacks the efficiency, flexibility, or defensibility considered to be a prerequisite for use on a repeated basis (e.g., once per day) to document changes in student social functioning given intervention exposure (Christ, Riley-Tillman, & Chafouleas, 2009). For instance, the Social Skills Improvement System–Rating Scale (SSIS-RS; Gresham & Elliott, 2008) serves as a gold standard within social skills assessment, allowing for the detection of acquisition and performance deficits across several social skill domains. Though reliability and validity evidence is promising (Frey, Elliott, & Gresham, 2011; Gresham, Elliott, Cook, Vance, & Kettler, 2010), the measure, which consists of approximately 80 items, requires too much time and effort to be completed on a repeated basis. In addition, the National Center for Intensive Intervention's (NCII; intensiveintervention.org) review of the SSIS-RS literature has suggested evidence of the instrument's sensitivity to change is unconvincing to date. This limitation is noteworthy given that sensitivity to change serves as the primary litmus test for an applied progress monitoring tool (Chafouleas, Sanetti, Kilgus, & Maggin, 2012). A more efficient version of the SSIS-RS, the SSIS Performance Screening Guide (Elliott & Gresham, 2008b), has been established and initially examined. Though it requires less time and effort, the instrument still lacks the sensitivity to change evidence required to justify applied use. In addition, the need to consider a student's behavior over the past several weeks when completing the SSIS Performance Screening Guide suggests it is unlikely to be sensitive to recent and potentially important changes in student behavior.

In sum, the state of the literature presents a dilemma. Researchers have identified interventions that promote social competence, but have not as of yet established the progress monitoring tools required to (a) document such promotion and (b) determine how these interventions might be modified in support of their effectiveness. This shortcoming is made clear through certain recent grant competitions, which have specifically sought to fund research related to the development of novel social competence outcome measures in support of clinical trials (e.g., Simons Foundation; www.simonsfoundation.org). It is therefore of interest to develop novel methods or examine the extent to which existing methods are appropriate for use in evaluating change in social competence. One assessment methodology with demonstrated defensibility in progress monitoring and that might be adapted to target social competence is found in *Direct Behavior Rating Single-Item Scales* (DBR-SIS).

Direct Behavior Rating Single-Item Scales

DBR-SIS represents a hybrid methodology, incorporating elements of multiple behavior assessment methodologies (Christ et al., 2009). Specifically, DBR-SIS is akin to behavior rating scales in that data collection requires a user to complete a brief rating of student behavior. DBR-SIS is also like systematic direct observation, as data are collected in the time and setting within which behavior is exhibited under prespecified circumstances (e.g., large-group science instruction, 11:00–11:50 a.m.). Individual DBR-SISs commonly correspond to broadly defined behaviors (e.g., disruption), representative of response classes incorporating topographically dissimilar behaviors that serve a common purpose (Riley-Tillman, Chafouleas, Christ, Briesch, & LeBel, 2009). Research regarding DBR-SIS has proliferated within the school-based social behavior domain, resulting in the detection of three core social behavior targets of academic engagement (AE), disruptive behavior (DB), and respectful behavior (RB; Chafouleas, 2011). Researchers have suggested these behaviors might represent "keys to success," serving as broad and general predictors of student social behavioral functioning (Chafouleas, 2011; Christ et al., 2009). DBR-SIS work has been promising, with the NCII identifying

DBR-SIS targets of AE and DB as possessing convincing evidence of validity, reliability, and sensitivity to change.

That the three core DBR-SIS targets pertain to key social behaviors might suggest their relevance to social competence progress monitoring. Yet, such applicability might be questioned given the likelihood of construct underrepresentation, defined as a narrow focus resulting in an instrument's omission of important dimensions of a construct (Messick, 1995). Although considered broad and general outcome measures of social behavior (Kilgus, Riley-Tillman, Chafouleas, Christ, & Welsh, 2014), the three DBR-SIS targets might not represent the entirety of social functionality targeted by social competence interventions. Though DB, AE, and RB influence one's *potential* to create and sustain adult and peer relations, they do not necessarily correspond to the behaviors and skills through which one actually fosters and maintains these relationships. One need only consider the aforementioned social skills/competence curricula to identify such behaviors. Classes of behaviors addressed via these interventions, and therefore viable progress monitoring targets, include communication, conflict resolution, problem solving, self-management, and adult and peer relations (Elliott & Gresham, 2008a; Knoff, 2001; Stichter et al., 2010).

DBR-Social Competence

In recognition of the potential shortcomings of the core DBR-SIS targets in terms of construct representation, efforts have recently been made to develop new DBR-SIS targets with greater relevance to social competence. These novel targets, when combined with the three core DBR-SIS targets, constitute what has been referred to as the *DBR-Social Competence* (DBR-SC) form. The new DBR-SC targets represent interactional behaviors, corresponding to skills that foster and sustain relationships with others. Targets were defined to be both positively oriented and molar (Riley-Tillman et al., 2009), representing broad response classes of behaviors known to be relevant to social competence in the school setting. The first novel DBR-SC target is *appropriate social interaction with teacher* (AT), defined as student interaction with a teacher in a socially expected manner. Examples included listening and following instructions provided by the teacher, responding/acknowledging direct teacher requests in a timely manner, using polite tone and/or body language, using appropriate eye contact and making appropriate social initiations with adult, or asking appropriate academic questions. The second novel DBR-SC target is *appropriate social interaction with peers* (AP), defined as student interaction with one or more other students in a socially expected manner. Examples included talking to other students when allowed in the setting, working collaboratively on cooperative assignments, appropriately reciprocating to interaction by another student(s), using polite tone and/or body language toward other student(s), or using appropriate eye contact and making several appropriate social initiations with peers. Both AT and AP are defined in such a way to incorporate behaviors that are displayed either (a) within a direct social interaction or (b) within a broader social context (e.g., classroom, playground) in accordance with contextual and social norms.

An initial study regarding the DBR-SC form examined the concurrent criterion-related validity of the three core and two novel targets, as compared to systematic direct observation, which targeted similarly defined behaviors (Riley-Tillman, Stichter, Kilgus, Schoemann, & Owens, 2015). DBR-SC and observational data were collected (a) within the same time and setting regarding students with social difficulties receiving intervention, and (b) at three time points corresponding to pre-, mid-, and postintervention implementation. All data were collected as part of a broader randomized control trial, wherein one group of students received a Business-as-Usual (BAU) control social skills training, and the other received the *Social Competence Intervention—Adolescent* (SCI-A), a suite of curricula targeting elementary, adolescent, and high school students with social needs (Schultz, Schmidt, & Stichter, 2011; Stichter, O'Connor, Herzog, Lierheimer, & McGhee, 2012). With data collapsed across these two groups and time, DBR-SC core targets evidenced small to moderate correlations with systematic direct observation, with AE = .45, DB = .25, and RB = .25. The novel DBR-SC behaviors evidenced lower but still small correlations (per Cohen's [1988] criteria for correlational magnitude), with coefficients equal to .25 for AT and .10 for AP. Additional analyses

indicated that correlations did not differ to a statistically significant degree across time despite general upward trends in magnitude.

This initial study was considered to provide partial but inconsistent support for DBR-SC psychometric defensibility. Although some correlations fell in or approximated the moderate range, others just met the 'small' criterion for correlational magnitude (i.e., AP). Riley-Tillman et al. (2015) proposed potential reasons for this latter finding. Specifically, the authors suggested findings may have reflected the somewhat restricted correspondence between DBR-SC and direct observation in terms of the constructs assessed. For instance, whereas the DBR-SC AP and AT targets are broad indicators representative of several interactional behaviors (e.g., play, eye contact, verbal behavior), the observational targets to which they were compared captured the comments made by students alone.

Alternative reasons for the somewhat restricted validity might be proposed, including the presence of limited DBR-SC *temporal reliability*, defined as the stability of a trait or behavior over time (Cicchetti, 1994). As is commonly understood, reliability serves as a necessary but insufficient condition of validity. That is, for DBR-SC data to demonstrate acceptable validity, it must first exhibit acceptable reliability (e.g., >.80–.90; Cortina, 1993; Nunnally, 1978). One could therefore argue for "stepping back" and expanding the DBR-SC literature via the examination of additional psychometric properties, including reliability. Previous research has supported the temporal reliability of DBR-SIS data in alternative student populations. To date, three DBR-SIS studies employing generalizability theory have been conducted with students in the general education setting. Each investigation has yielded valuable information that informs recommendations regarding the manner in which DBR-SIS should be applied in schools. For instance, findings have spoken to the number of DBR-SIS ratings required to support defensible decisions regarding student service delivery. In particular, research has suggested the amount of DBR-SIS data required to approximate levels of temporal reliability required to support low stakes decisions (i.e., .70), such as screening and progress monitoring. Findings have varied slightly, with estimates ranging between 7 and 10 (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007), 5 and 10 (Chafouleas et al., 2010), and 15 and 20 data points (Briesch, Chafouleas, & Riley-Tillman, 2010). More recently, Chafouleas et al. (2013) employed an alternative approach to reliability evaluation, using intraclass correlations to estimate score reliabilities in accordance with recommendations from Shrout and Fleiss (1979). Reliability estimates exceeded those of prior research, indicating collection of 5 to 10 data points resulted in reliabilities of .90 or greater.

Purpose of the Study

The overall purpose of the current investigation was to build on prior DBR-SIS-specific research in examining the reliability of DBR-SC data within a sample of students with social deficits receiving social competence interventions. The dataset used within this investigation was examined within the aforementioned Riley-Tillman et al. (2015) investigation, with DBR-SC data collected with regard to students with social difficulties across three time points (pre, mid, and post intervention) and two treatment conditions (i.e., SCI-A and BAU). Four research questions were of interest. First, what is the reliability of DBR-SC data within each phase and treatment group? In consideration of prior findings, it was hypothesized DBR-SC data reliability would meet or exceed the threshold for low (.80) or high stakes decisions (.90; Cortina, 1993; Nunnally, 1978) across all phases and groups. Second, does DBR-SC data reliability vary across treatment groups? It was anticipated that relative to the BAU control, SCI would promote behavior that was more appropriate and consistent. Presuming DBR-SC would be sensitive to these differences across groups, it was hypothesized that DBR-SC data collected within the SCI group would be more reliable. Third, does DBR-SC data reliability vary across phases? It was hypothesized reliability would improve over time, reflecting greater consistency in student behavior in response to intervention. Fourth, how many DBR-SC data points are required to reach the low and high stakes reliability thresholds? Given previous findings, it was hypothesized that 5 to 10 data points would be required to reach the low stakes threshold. Answers to this final research question correspond to important

implications for practice, informing recommendations regarding the nature and extent of DBR-SC data collection in applied settings.

Method

Participants and Setting

The current data were collected as part of a larger investigation evaluating the efficacy of the SCI-A. Specifically, the data included in this investigation were those collected in the first year of the broader project across all participating students and teachers. SCI-A is a targeted group social skills intervention comprised of curricula designed for students ages 11 through 14 with deficits in social functioning. SCI-A curricula were developed through an integration of cognitive-behavior techniques, applied behavior analysis, empirically driven methodology, and phenotypic profiling. Previous investigations have demonstrated that SCI-A is associated with improvements in social skills and executive functioning in youth with average to high IQ and social skills deficits (e.g., high functioning autism spectrum disorder). Furthermore, teacher, parent, and student reports lend support for the acceptability, utility, and feasibility of SCI-A school programming. The SCI-A curricula is administered by a teacher to groups of 4 to 6 students and is composed of five scaffolded units: facial expressions, sharing ideas, taking turns in conversation, recognizing emotions and feelings in self and others, and problem solving. Each unit is composed of approximately five to six 1-hr lessons for a total of 31 lessons (Stichter et al., 2010).

The current implementation of SCI-A is part of an IES-funded randomized control trial taking place in middle schools in a Midwestern state. Schools are randomized to either the SCI group or BAU control group. Students enrolled in SCI-A participated in intervention sessions at a scheduled time every other day with a group of 4 to 6 students. Most BAU interventions were delivered to participating students in social studies or science classes. Classroom functioning was measured by DBR-SC and systematic direct observation.

Participants met the following inclusion criteria: (a) student aged 11 to 14, (b) diagnosis of ASD or Special Education eligibility criteria of autism or school-identified social need,[1] and (c) cognitive functioning (i.e., full-scale IQ) within 2.0 standard deviations of the mean. A sample of 33 students at six schools constituted the SCI-A group and 30 students at six schools constituted the BAU group for a total of 63 participants. Two students were dropped from analyses because of misreported IQ scores and one additional student was dropped because of a lack of data on outcome measures. The resulting sample includes 60 students (29 SCI-A and 31 BAU). Parent consent and student assent were obtained before the start of the study.

Across all student participants, 55 students were male and 5 were female. The majority of participants met criteria for special education services, specifically 43.33% in the Autism category, 25% in the Emotional Disturbance category, and 20% in the Other Health Impairment category. Two students met eligibility for Specific Learning Disability, and one student met eligibility for Speech/Language Impairment. Four students did not have a current individualized education plan (IEP), and one student had a Section 504 Plan without an IEP. Full-scale IQ composites were assessed using the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II; Wechsler, 2011) or the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2004). Mean score across groups was 95.75 (SD = 14.92). The majority of the same were Caucasian/White (71.66%), with three participants identifying as bi- or multiracial, two participants identified as African American, one participant identified as Native American, and one participant identified as Hispanic (15% of the sample declined to report race/ethnicity data).

Teachers (n = 23) across all 12 participating middles schools were also recruited to complete assessments on participating students in a general education setting. These teachers were not involved in the delivery of either intervention. (Note: interventions were implemented by school designated educators [typical special education teachers, speech/language pathologists, or counselors], who received specific training

[1] Such social needs were identified as part of normal educational practice. The procedures and criteria through which social needs were identified and specified varied by school in accordance with local practice.

from the researchers regarding intervention implementation.) Neither were participating teachers aware of the condition in which each student was enrolled. Rather, participating teachers were solely responsible for collecting DBR data for 1 to 3 students in their classroom at a time.

Measures – Direct Behavior Rating-Social Competence (DBR-SC)

The DBR-SC is a hybrid assessment tool, which combines elements of systematic direct observation and behavior rating scales (Chafouleas, 2011). Core target behaviors (i.e., AE, DB, and RB) as well as social competence behaviors (i.e., AT and AP) are measured on an 11-point unipolar graphic rating scale after specific observations. The scale extends from 0% to 100%, with increments of 10% and specified qualitative anchors at the beginning, middle, and end of the scale (0 = *0%, never*; 5 = *50%, sometimes*; 10 = *100%, always*). Qualtrics, a Web based survey system, was utilized to distribute surveys and collect data.

Procedures

DBR-SC teacher training. Participating teachers consented to complete daily DBR-SC observations on their target students at the end of a typical school lesson. Teachers were oriented to the DBR-SC during the consent process, including instruction on Qualtrics, the overall use of DBR-SC, and specific training on the five key DBR-SC targets. Teachers were informed that ratings of the five targets were set on an 11-point scale from 0% to 100% with intervals of 10% with 0% representing *none of the time*, 50% representing *sometimes*, and 100% representing *all of the time*. Furthermore, teacher participants were instructed that ratings are independent of each other, such that the five ratings did not have to equal 100%.

Data collection. DBR-SC data was collected using Qualtrics survey system. Links to the survey were emailed to teacher participants at 7:00 a.m. the morning of the targeted observation (typically every day school was in session for five months). Reminder e-mails were sent at 2:00 p.m. to teachers who did not complete their surveys. Surveys expired at 9:00 a.m. the following morning to account for teacher capacity to accurately recall behavior the student exhibited during the observation period. Data were coded in Qualtrics then imported into SPSS for data cleaning and analysis. Response rates were monitored for patterns of inconsistent or inaccurate data. In such cases, teachers were prompted via e-mail to complete accurate surveys for every targeted observation. Across the 23 teachers, response rates remained high (mean response rate = 86.05%; range = 56.18%–100%). A total of 3202 observations (mean *n* per student = 51.66; range = 19–86) were made over approximately five months and the three project phases, defined as pre (Time 1), mid (Time 2), and post (Time 3) intervention. Although the specific timing of each phase varied by student in accordance with his or her start date within the study, Time 1 generally corresponded to January, Time 2 to February and March, and Time 3 to April and May. On average, within each time phase, teachers completed 27.48 ratings in the SCI group and 30.38 ratings in the BAU group for each participating student within each of the five DBR targets.

Data Analysis Plan

To assess DBR-SC performance, intraclass correlations (ICC) coefficients were first calculated to evaluate the consistency of DBR-SC data points across time within students. ICC and other statistics were calculated separately for different time points and different groups (SCI-A and BAU). This resulted in a 2 (treatment groups) × 3 (times of assessment) mixed design with repeated measures on the time of assessment. ICCs were calculated via a two-level unconditional multilevel structural equation model, with DBR-SC observations at level 1 and students at level 2 of the model.[2] For each model, variances and covariances of each DBR-SC subscale were estimated both between and within observations. ICC was computed as the ratio of between group variance to total variance (between group variance + within group variance). Next, ICCs were used to generate reliability estimates in accordance with

[2] In this design there is an additional level of nesting, students nested within schools/teachers. Analyses indicated that a very small proportion of the variance in DBR scores was attributable to differences at the teacher level (ICCs < .01). All analyses were also run correcting for this level of nesting. Results when correcting for nesting within teacher did not differ from the results presented.

recommendations from Shrout and Fleiss (1979). Specifically, reliability coefficients were computed as follows:

$$r_{xx} = \frac{k * ICC}{(1 + (k - 1) * ICC)}$$

Where k is the number of observations and ICC is equal to the ICC described above. For each analysis, k was set to 19, which represented the lowest number of DBR-SC ratings completed for a student within a given time point. Setting k to 19 for each analysis represented an attempt to provide a more conservative estimate of DBR-SC data reliability, as determined via ICC. The minimum number of observations needed to achieve a reliability of 0.80 was computed as follows:

$$\min_k = \frac{0.80 * (1 - ICC)}{ICC * (1 - 0.80)}$$

ICC and other quantities were estimated across groups using multiple group modeling and across time by including measurements from all three time points in a model and allowing all measurements to correlate. Using contrast codes (Schoemann, Short, & Little, 2014; Thompson & Green, 2013), ICCs were compared across groups, across time, and across the interaction of group and time. See Table 1 for a depiction of contrast codes. The relevant ICC (for time and group membership) was multiplied by the each code, and codes for each effect were summed. Tests of main effects and interaction effects were performed by testing whether the relevant contrast codes were different from zero. For example, a test of the main effect of group examined whether the contrast code for group was different from 0, but a contrast code for the interaction between time and group tested whether both contrast codes for the interaction were different from zero. All contrast codes were tested using a likelihood ratio test. All analyses were conducted using Mplus v. 7.12 (Muthén & Muthén, 2014).

Results

See Table 2 for a summary of descriptive statistics associated with each DBR-SC target, broken out by time point and intervention group. ICCs and reliability coefficients for both groups across time are reported in Table 3. Reliability coefficients predominantly fell above the high stakes decision making threshold of .90 (Nunnally, 1978) across all DBR-SC targets, time points, and intervention groups. This was with the exception of the (a) RB target during Time 1 across both SCI-A and BAU, (b) DB target during Time 1 and 2 for the BAU group, and the (c) AT target during Time 1 within the BAU group. For each of these exceptions, reliabilities were found to meet or approach the low stakes decision making threshold of .80 (Cortina, 1993).

Results indicated that to reach a reliability of .80, it was necessary to collect an average (across time and DBR targets) of 7.37 DBR-SC data points (SD = 5.88) within the BAU group and 3.69 (SD = 2.43) data points within the SCI-A group. Within the BAU group, and averaged across DBR-SC targets, the number of necessary data points for Times 1, 2, and 3 were equal to 10.41, 7.66, and 4.03, respectively. Within the SCI-A group, the number of necessary data points was equal to 6.41 for Time 1, 2.90 for Time 2, and 1.76 for Time 3. Within the BAU group, and av-

Table 1
Contrast Codes Used in Tests of Main Effects and Interaction for Differences in Intraclass Correlations (ICCs)

Code	SCI 1	SCI 2	SCI 3	BAU 1	BAU 2	BAU 3
Main effect – Condition	1	1	1	−1	−1	−1
Main effect – Time 1	2	−1	−1	2	−1	−1
Main effect – Time 2	0	1	−1	0	1	−1
Interaction 1	2	−1	−1	−2	1	1
Interaction 2	0	1	−1	0	−1	1

Note. BAU = business-as-usual intervention; SCI = social competence intervention.

Table 2
Descriptive Statistics

DBR-SC target	Group	Time 1		Time 2		Time 3	
		Mean	Range	Mean	Range	Mean	Range
AE	SCI	79.98	100.00	80.79	100.00	80.33	100.00
	BAU	73.66	100.00	78.89	100.00	75.60	100.00
AT	SCI	86.79	100.00	85.83	100.00	82.88	100.00
	BAU	84.97	100.00	87.00	90.00	84.44	100.00
AP	SCI	83.28	100.00	85.28	100.00	83.55	100.00
	BAU	79.64	100.00	82.28	100.00	81.61	100.00
DB	SCI	87.95	100.00	81.07	100.00	86.72	100.00
	BAU	78.85	100.00	64.62	100.00	59.63	80.00
RB	SCI	90.39	100.00	90.25	100.00	90.29	100.00
	BAU	88.59	100.00	90.58	100.00	89.99	80.00

Note. DBR-SC = Direct Behavior Rating–Social Competence; AE = academic engagement; RB = respectful behavior; DB = disruptive behavior; AT = appropriate interactions with teachers; AP = appropriate interactions with peers; BAU = business-as-usual intervention; SCI = social competence intervention.

eraged across time points, the number of necessary data points was equal to 5.94 for AE, 8.66 for RB, 15.34 for DB, 4.82 for AT, and 2.07 for AP. Within the SCI-A group, the number of necessary data points was equal to 2.99, 4.56, 5.13, 2.66, and 3.11 for AE, RB, DB, AT, and AP, respectively.

Tests of main effects of group and time and the interaction of group and time are reported in Table 4. There were no main effects of group membership (i.e., BAU vs. SCI-A) on ICC. In contrast, for all outcomes except RB there was a main effect of time, such that ICC (and thus reliability) increased over time. There were also statistically significant Time by Group interactions for both AE and RB. For both outcomes, ICC increased over time at a greater rate for the SCI-A group relative to the BAU group. This suggests that student academically engaged and respectful behavior became more consistent over time in the SCI-A group compared with the BAU group.

Table 3
Intraclass Correlations (ICCs), Reliability Coefficients, and the Minimum Number of Observations Required for DBR-SC Across Groups

Measure	Time	BAU ICC	BAU reliability	BAU minimum observations	SCI ICC	SCI reliability	SCI minimum observations
AE	1	.373	.919	6.732	.409	.929	5.779
	2	.376	.920	6.633	.709	.979	1.639
	3	.473	.945	4.462	.721	.980	1.550
RB	1	.227	.848	13.597	.310	.895	8.906
	2	.343	.909	7.653	.524	.954	3.628
	3	.459	.942	4.716	.778	.985	1.140
DB	1	.177	.804	18.566	.380	.921	6.516
	2	.167	.793	19.885	.464	.943	4.615
	3	.346	.909	7.575	.484	.947	4.269
AT	1	.285	.883	10.059	.411	.930	5.741
	2	.627	.970	2.381	.689	.977	1.803
	3	.664	.974	2.023	.904	.994	.426
AP	1	.564	.961	3.094	.439	.937	5.121
	2	.696	.978	1.744	.589	.965	2.795
	3	.744	.982	1.378	.740	.982	1.405

Note. All calculations assume an average of 19 observations per student within each phase of data collection. AE = academic engagement; RB = respectful behavior; DB = disruptive behavior; AT = appropriate interactions with teachers; AP = appropriate interactions with peers; BAU = business-as-usual intervention; SCI = social competence intervention.

Table 4
Test of Main Effects and Interaction for Differences in Intraclass Correlations (ICCs)

Measure	Effect	Test
AE	Main effect Time	$\chi^2(2)=20.173$, $p < .0001$
	Main effect Condition	$\chi^2(1)= 1.391$, $p = .2380$
	Time × Condition interaction	$\chi^2(2)= 12.455$, $p = .0002$
RS	Main effect Time	$\chi^2(2)=5.875$, $p = .0530$
	Main effect Condition	$\chi^2(1)= 1.020$, $p = .3122$
	Time × Condition interaction	$\chi^2(2)=6.346$, $p = .0419$
DB	Main effect Time	$\chi^2(2)=10.530$, $p = .0066$
	Main effect Condition	$\chi^2(1)= 2.289$, $p = .1303$
	Time × Condition interaction	$\chi^2(2)=4.325$, $p = .1150$
AT	Main effect Time	$\chi^2(2)=49.751$, $p < .0001$
	Main effect Condition	$\chi^2(1)= 2.022$, $p = .1550$
	Time × Condition interaction	$\chi^2(2)= 5.354$, $p = .0688$
AP	Main effect Time	$\chi^2(2)=85.19$, $p < .0001$
	Main effect Condition	$\chi^2(1)= .484$, $p = .4864$
	Time × Condition interaction	$\chi^2(2)=2.610$, $p = .271$

Discussion

The overarching purpose of the current investigation was to evaluate the reliability of data derived from DBR-SC targets; that is, the consistency of DBR data points collected across time within individual student participants. Results indicated data derived from each DBR-SC target were associated with adequate reliability, with coefficients meeting commonly accepted thresholds for low and high stakes decisions. This was despite the conservative nature of the current approach, wherein the consistency of DBR data was evaluated across the smallest number of ratings collected for an individual student participant. Findings were generally consistent in their support of DBR-SC performance, as all targets met reliability thresholds within each time point across both intervention groups. Results further suggested minimally acceptable reliability levels were achievable with only a small number of ratings. When averaging these estimates of the number of necessary ratings across DBR-SC targets, findings suggested it was possible to achieve acceptable reliability with approximately 4 to 10 data points within the BAU group and 2 to 6 data points within the SCI-A group. These estimates are similar if not superior to those derived from previous DBR-related studies, which supported the need to collect 5 to 20 data points to achieve adequate reliability (Briesch et al., 2010; Chafouleas et al., 2007, 2010). Taken together, these results hold strong implications for practice, as they appear to support the efficiency of the DBR-SC form. That is, it appears possible for educators to attain a relatively reliable estimate of student behavior after only 1 to 2 weeks of daily data collection. This is considered highly encouraging when contrasted with the time and resource intensive nature of alternative progress monitoring methods used in evaluating the effectiveness of social competence interventions (e.g., systematic direct observation), the utility of which has been questioned repeatedly within the literature (Gresham, 2010; Riley-Tillman et al., 2015).

A cursory examination of descriptive statistics might suggest that reliability levels varied in accordance with the time point and intervention group under consideration. As an example, findings suggested that on average, it was possible to achieve reliability of .80 within the SCI-A group with nearly half of the data required within the BAU group (i.e., 7.37 vs. 3.69). Such a finding was in accordance with expectations, as the SCI-A intervention was hypothesized to be more effective than the BAU approach, thus leading to more consistent behavior among students receiving SCI. Yet, the main effect for group membership was not significant, suggesting that although the difference in reliability in the groups was apparent, it was not to a statistically significant extent.

As an additional example, descriptive statistics indicated the reliability of DBR-SC data improved over time, with the average number of required data points falling across each excessive time point. This finding was supported by the statistically significant main effect for time, which indicated that with the exception of the RB target, reliability increased across data collection phases. Such a finding might be the result of several phenomena, with two being particularly likely. First, increased reliability over time might be the result of teachers becoming arbitrarily more consistent in their ratings of student behavior, reflecting potential measurement bias in the form of halo effects.

Second, as hypothesized, increases in reliability might result from actual increases in the consistency of observed student behavior in response to intervention. Unfortunately, in the absence of sufficient time series criterion data with which DBR-SC data might be compared (e.g., systematic direct observation data), it is impossible to verify the cause of observed increases in reliability across time. Yet, partial support for the latter explanation is found in a comparison of reliability findings across treatment groups. Specifically, a review of descriptive statistics suggested the finding of improved reliability over time phenomenon was particularly characteristic of the SCI-A group, within which the number of necessary data points in Times 2 and 3 was equal to half that of its immediately preceding time point (i.e., 6.41 vs. 2.90 vs. 1.76). This was despite a lack of teacher knowledge regarding the intervention condition in which each student was enrolled.

Although inconsistent, Time by Group interaction effects further supported the viability of this second explanation for the increases in reliability over time. Specifically, relative differences in the changes in the reliability of data within the RB and AE targets suggests SCI-A may have been particularly effective in addressing student engagement and respect. Such a finding is somewhat unsurprising given SCI-A's incorporation of curricular elements intended to promote student social functioning and capacity to participate appropriately within the classroom and in relation to adults and peers. That the interaction effect was not noted for the remaining three DBR-SC targets might indicate that (a) SCI-A did not affect DB, AT, or AP, (b) each of these three targets were subject to floor or ceiling effects, thus allowing for limited change following the introduction of intervention, or (c) the effect of SCI on behavior was in terms of mean student response and not the variability of behavior within students across time.

Regardless of the actual reason for the increase in reliability, if any, the current results support for the reliability of DBR-SC data, concurring with hypotheses regarding the magnitude of reliability and its variability across intervention groups and time. Furthermore, findings suggest such levels of reliability might be reached with a rather limited number of time points. This ultimately speaks to the efficiency of the DBR-SC as a progress monitoring instrument and the potential feasibility of its use in evaluating student response to social competence interventions.

Multiple limitations to this investigation ought to be noted. First, the current sample size was considered somewhat restricted. Thus, tests of statistical significance may have possessed limited power to reject null hypotheses. As such, it is recommended that the reader interpret null findings with caution. Second, this investigation did not allow for the evaluation of DBR-SC data reliability relative to an additional behavior assessment standard. It would have been preferable to consider how the reliability of DBR-SC data compared with that of a gold-standard measure, such as systematic direct observation. Such an approach would permit an evaluation of the extent to which consistency in DBR-SC ratings was characteristic of true variance and not measurement error characterized by inappropriately invariant data. The current researchers originally sought to incorporate such an assessment standard via consideration of the reliability of systematic direct observation data collected as part of this investigation. Unfortunately, the quantity of observational data was somewhat limited, with only three observations completed for each student within each of the three phases of data collection. Because the evaluation of such a small amount of data would likely underestimate the reliability of direct observation data, it was ultimately decided that observational data would not be evaluated. It is recommended that future researchers collect additional observational data and thus permit comparisons of reliability across both DBR-SC and systematic direct observation.

Third, given the structure of the current data and data collection procedures, it was only possible to evaluate how the number of ratings within a single rater influenced the reliability of DBR-SC data. It was not possible to evaluate the influence of additional variations in DBR-SC data collection procedures, such as the introduction of additional raters and observation occasions within each day. Previous DBR-related examinations allowed for such evaluations, as researchers incorporated data structures that permitted the use of more advanced statistical procedures in evaluating DBR-SC rating reliability. Specifically, as noted within the introduction, multiple prior DBR-related investigations employed generalizability theory (e.g., Briesch et al., 2010; Chafouleas et al., 2010), which allowed for the examination of various sources of variance in DBR ratings, including students, raters, and rating occasions. It is therefore suggested that future DBR-SC investigations incorporate data collection procedures that allow for the use of generalizability theory, and thus a more nuanced understanding of DBR-SC rating reliability and the conditions under which it might be expected.

An additional area for future research includes the continued evaluation of DBR-SC validity. The current findings suggest that the inconsistent validity documented in the first DBR-SC investigation (Riley-Tillman et al., 2015) was likely not a result of insufficient reliability. Alternative explanations for these findings remain, including a comparison with suboptimal criterion measures and associated constructs. It is therefore recommended that future studies consider alternative criterion measures, including those affording greater convergence with the behavioral constructs assessed via DBR-SC targets. The importance of such research should not be understated, as although the current findings yield important evidence of DBR-SC score reliability, evidence of DBR-SC score validity has yet to be established. Such evidence is considered a necessary prerequisite for applied use of DBR-SC targets.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and direct behavior rating. *School Psychology Review, 39,* 408–421.

Center for Disease Control. (2014). Prevalence of autism spectrum disorders among children aged 8 years: Autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR*. Surveillance Summaries, *63,* 1–22.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education & Treatment of Children, 34,* 575–591. http://dx.doi.org/10.1353/etc.2011.0034

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C., & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of direct behavior rating single item scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology, 48,* 219–246. http://dx.doi.org/10.1016/j.jsp.2010.02.001

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

Chafouleas, S. M., Kilgus, S. P., Jaffery, R., Riley-Tillman, T. C., Welsh, M., & Christ, T. J. (2013). Direct behavior rating as a school-based behavior screener for elementary and middle grades. *Journal of School Psychology, 51,* 367–385. http://dx.doi.org/10.1016/j.jsp.2013.04.002

Chafouleas, S. M., Sanetti, L. M. H., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change using Direct Behavior Rating Single-Item Scales. *Exceptional Children, 78,* 491–505.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34,* 201–213. http://dx.doi.org/10.1177/1534508409340390

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. http://dx.doi.org/10.1037/1040-3590.6.4.284

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Elliott, S. N., & Gresham, F. M. (2008a). *Social Skills Improvement System Performance Screening Guide*. Minneapolis, MN: Pearson.

Elliott, S. N., & Gresham, F. M. (2008b). *Social Skills Improvement System Intervention Guide manual*. Minneapolis, MN: Pearson Assessments.

Frey, J. R., Elliott, S. N., & Gresham, F. M. (2011). Preschoolers' social skills: Advances in assessment for intervention using social behavior ratings. *School Mental Health, 3,* 179–190. http://dx.doi.org/10.1007/s12310-011-9060-y

Gresham, F. M. (1986). Conceptual and definitional issues in the assessment of children's social skills: Implications for classification and training. *Journal of Clinical Child Psychology, 15,* 3–15. http://dx.doi.org/10.1207/s15374424jccp1501_1

Gresham, F. M. (2010). Data-based decision making for students' social behavior. *Journal of Evidence-Based Practices for Schools, 11,* 149–168.

Gresham, F. M., & Elliott, S. N. (2008). *Social skills improvement system: Rating scales*. Bloomington, MN: Pearson.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System-Rating Scales. *Psychological Assessment, 22,* 157–166. http://dx.doi.org/10.1037/a0018124

Gresham, F. M., Elliott, S. N., & Kettler, R. J. (2010). Base rates of social skills acquisition/performance deficits, strengths, and problem behaviors: An analysis of the Social Skills Improvement System—Rating Scales. *Psychological Assessment, 22,* 809–815. http://dx.doi.org/10.1037/a0020255

Gresham, F. M., Sugai, G., & Horner, R. H. (2001). Interpreting outcomes of social skills training for students with high-incidence disabilities. *Exceptional Children, 67,* 331–344.

Kilgus, S. P., Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J., & Welsh, M. E. (2014). Direct behavior rating as a school-based behavior universal screener: Replication across sites. *Journal of School Psychology, 52,* 63–82. http://dx.doi.org/10.1016/j.jsp.2013.11.002

Knoff, H. M. (2001). *The Stop & Think Social Skills Program (Preschool–Grade 1, Grades 2/3, Grades 4/5, Middle School 6–8)*. Longmont, CO: Sopris West.

McFall, R. (1982). A review and reformulation of the concept of social skills. *Behavioral Assessment, 4,* 1–33.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Muthén, L. K., & Muthén, B. O. (2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J., Briesch, A. M., & LeBel, T. J. (2009). The impact of wording and behavioral specificity on the accuracy of Direct Behavior Ratings (DBRs). *School Psychology Quarterly, 24,* 1–12. http://dx.doi.org/10.1037/a0015248

Riley-Tillman, T. C., Stichter, J. P., Kilgus, S. P., Schoemann, A., & Owens, S. (2015). *Examining the concurrent criterion-related validity of Direct Behavior Rating measures of social competence*. Manuscript submitted.

Schoemann, A. M., Short, S. D., & Little, T. D. (2014). *Examining between, within, and mixed factorial designs with structural equation modeling*. Poster presented at the Annual Meeting of the American Psychological Association, Washington, DC.

Schultz, T. R., Schmidt, C. T., & Stichter, J. P. (2011). A review of parent education programs for parents of children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 26,* 96–104. http://dx.doi.org/10.1177/1088357610397346

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Stichter, J. P., Herzog, M. J., Visovsky, K., Schmidt, C., Randolph, J., Schultz, T., & Gage, N. (2010). Social competence intervention for youth with Asperger Syndrome and high-functioning autism: An initial investigation. *Journal of Autism and Developmental Disorders, 40,* 1067–1079. http://dx.doi.org/10.1007/s10803-010-0959-1

Stichter, J. P., O'Connor, K. V., Herzog, M. J., Lierheimer, K., & McGhee, S. D. (2012). Social competence intervention for elementary students with Aspergers syndrome and high functioning autism. *Journal of Autism and Developmental Disorders, 42,* 354–366. http://dx.doi.org/10.1007/s10803-011-1249-2

Thompson, M. S., & Green, S. B. (2013). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 119–169). Charlotte, NC: IAP.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, February). *Early Childhood Education Interventions for Children with Disabilities intervention report: Social skills training*. Retrieved from http://whatworks.ed.gov

Wechsler, D. (2004). *Wechsler Intelligence Scale For Children* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence* (2nd ed.). San Antonio, TX: Pearson.

Received March 26, 2015
Revision received August 23, 2015
Accepted August 26, 2015 ■