School Psychology Review,
2016, Volume 45, No. 1, pp. 39–52

Dependability of Two Scaling Approaches to Direct Behavior Rating Multi-Item Scales Assessing Disruptive Classroom Behavior

Robert J. Volpe and Amy M. Briesch
Northeastern University

Abstract. This study examines the dependability of two scaling approaches for using a five-item Direct Behavior Rating multi-item scale to assess student disruptive behavior. A series of generalizability theory studies were used to compare a traditional frequency-based scaling approach with an approach wherein the informant compares a target student's behavior with that of classroom peers. A total of seven novice raters (i.e., graduate students) used both types of scales to rate 10-min video clips of the classroom behavior of nine middle school students across three occasions. Generalizability of composite scores derived from each type of scale was examined across raters and occasions. Subsequent decision studies were conducted to determine the number of measurement occasions that would be required to obtain an acceptable level of dependability. Results of these studies indicated that the type of scale accounted for a substantial proportion of variance (29%) and that the traditional frequency approach required far fewer assessment occasions to reach the criterion for absolute and relative decisions (4 and 8 occasions, respectively) compared with the comparative scaling approach (>30 occasions). Implications for future research and current practice are discussed.

Schools are increasingly using tiered models of prevention and a problem-solving framework wherein every student is exposed to primary prevention and assessments are used to match the intensity of subsequent intervention to the level of student risk (Gresham, 2014). The success of such models depends, in large part, on the collection of progress-monitoring data to assess student response to intervention and to determine if the level of support provided is adequate or if additional or alternative supports are necessary (National Center on Response to Intervention, 2010). Although the availability of these types of measures for reading and mathematics has facilitated the adoption of problem-solving models in these academic subject areas (e.g., Jimerson, Burns, & VanDerHeyden, 2016), progress in the behavioral domain has been comparatively slow given the lack of appropriate measurement tools (Chafouleas, Volpe, Gresham, & Cook, 2010).

To have utility within a problem-solving model, it has been argued that a progress-monitoring system for social behavior should include the assessment of both specific performance objectives and long-term general objec-

Please address correspondence regarding this article to Robert J. Volpe, Northeastern University, Department of Applied Psychology, 413 International Village, Boston, MA 02115; e-mail: r.volpe@neu.edu

tives or behavioral general outcome measures (GOMs; Kratochwill & Bergan, 1990; Volpe & Gadow, 2010). That is, it is important to assess not only short-term changes in specific behaviors that are targeted for intervention (e.g., calling out, off-task) but also the longer-term impact of the intervention on broader domains of functioning (e.g., disruptive behavior, social behavior). We use the term *behavioral GOMs* to refer to feasible and technically adequate measures that are designed for repeated assessment of broad constructs (e.g., disruptive behavior, academic engagement/motivation, oppositional behavior, social skills).

If behavioral measures are to demonstrate utility for decision making in progress-monitoring applications, they should not only demonstrate defensible psychometric characteristics (e.g., reliability, validity, treatment sensitivity) but also be (a) feasible (cost-effective) and efficient for repeated administration (i.e., quick to complete), (b) flexible to reflect unique concerns, and (c) reflective of meaningful levels of change (Christ, Riley-Tillman, & Chafouleas, 2009; Kazdin, 2011). Although those measures that traditionally have been used to assess treatment response (e.g., systematic direct observation, traditional rating scales) meet some of the requirements for use in problem-solving models, their utility within a progress-monitoring context is limited given concerns related to feasibility (Briesch & Volpe, 2007; Chafouleas, 2011; Hintze & Matthews, 2004). However, the transition from resource-intensive assessment methods designed primarily for classification to more feasible methods has been slow. There has been increased attention directed to the development of feasible and psychometrically sound progress-monitoring measures for behavioral problems, but consensus has not yet been reached concerning the appropriate targets of assessment, the methods that should be used to measure them, or whether any of the candidate methods might serve as GOMs of behavioral functioning akin to those found in the academic realm (Chafouleas, Volpe et al., 2010).

DIRECT BEHAVIOR RATING

One method that has demonstrated much promise within the context of behavioral progress monitoring is Direct Behavior Rating (DBR). DBR can be conceptualized as a category of behavioral assessment that includes several different specific formats or methods; however, at its core, DBR is "an evaluative rating that is generated at the time and place that behavior occurs by those persons who are naturally present in the context of interest" (Christ et al., 2009, p. 205). The timing of DBRs represents a key difference from traditional rating scales. Whereas traditional rating scales typically require the informant (e.g., teacher) to provide a summative rating of behaviors that have occurred over weeks or months, DBRs are designed to rate behaviors observed over much shorter intervals ranging from minutes (e.g., Briesch, Kilgus, Chafouleas, Riley-Tillman, & Christ, 2012; Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012) to hours (Chafouleas, Kilgus, Jaffery, Riley-Tillman, Welsh, & Christ, 2013; Kilgus, Chafouleas, Riley-Tillman, Christ, & Welsh, 2014).

There are multiple options for conducting ratings within the broad category of DBR. The most well-studied method of DBR to date is the single-item scale (DBR-SIS). Using this method, informants are provided with a brief description of the broad behavioral construct being assessed (including examples of behaviors that are considered indicators of the construct) and asked to rate how often the target student demonstrates the behavior. Most typically, the perceived frequency of a behavior has been assessed using a continuous line consisting of several anchors indicating points from 0% to 100% (e.g., Chafouleas, Briesch, et al, 2010); however, Likert-type scales also have yielded favorable results (e.g., Volpe & Briesch, 2012). The research group led by Chafouleas and Riley-Tillman has conducted a programmatic line of research (>30 published studies since 2002) examining this method (for a review, see Chafouleas, 2011).

Another available option is the use of a DBR multi-item scale (DBR-MIS), in which

multiple indicators of a broad behavioral construct are rated simultaneously and summarized in aggregate. A number of different approaches have been investigated within the broad category of DBR-MISs (e.g., Fabiano, Vujnovic, Naylor, Pariseau, & Robins, 2009; LeBel, Chafouleas, Britner, & Simonsen, 2013), but recent work has focused on the psychometric adequacy of multi-item scales that rated several specific indicators (e.g., calls out, is noisy, clowns around) of a broader behavioral construct (e.g., disruptive behavior; Volpe & Briesch, 2012, 2015).

DBR Multi-Item Scales

Volpe and Briesch (2012) recently investigated the psychometric adequacy of two five-item DBR-MISs (academic engagement/motivation and disruptive behavior) designed specifically for progress-monitoring purposes. Raters used a 6-point scale (0 = *never*, 1 = *rarely*, 2 = *sometimes*, 3 = *often*, 4 = *very often*, 5 = *always*) to assess the behavior of eight middle school students across three occasions. Generalizability theory (GT; Cronbach, Gleser, Rajaratnam, & Nanda, 1972) was then used to determine the number of ratings needed to obtain a dependable estimate of each global behavior. GT represents an extension of classical test theory, in which the goal is to determine how accurately one can generalize from a specific sample of behavior to all possible samples of interest. Whereas traditional reliability analyses only allow the user to examine one source of error variance at a time (e.g., raters in assessing interrater reliability), GT can be used to examine multiple sources of rating variance simultaneously. In this way, it is possible to determine which facets (e.g., raters, items, time) are contributing the greatest proportion of rating variance and should therefore be targeted in making improvements in the measurement procedure.

Findings from the Volpe and Briesch (2012) study indicated that few assessments (i.e., four) were necessary to reach a dependability coefficient of .80 or greater when assessing academic engagement within the context of progress monitoring. However, the disruptive behavior DBR-MIS required 3 times as many assessments (i.e., 12). Similar to findings in other studies (e.g., Chafouleas, Briesch et al., 2010), the largest sources of measurement error were those involving time, including changes in overall student behavior across days and changes in the rank order of students across days. Some research has suggested that time-related variance components may be even larger for students with significant behavioral concerns as compared with typically developing children (Briesch, Volpe, & Ferguson, 2014).

Reducing Measurement Error in DBR Assessment

The findings noted earlier concerning time-related variance are not surprising, given that the types of behaviors assessed (e.g., academic engagement, disruptive behavior) are highly influenced by environmental conditions such as the degree of classroom structure or what activity preceded the target observation period. In the analysis of single-case designs, variability of behavior both within and across phases is of particular interest from a behavior-analysis perspective (Cooper, Heron, & Heward, 2007). For example, variability within conditions may be indicative of the level of control one has over the dependent variable, and effective interventions may lead to changes in the level, trend, and variability of the dependent variable. The unfortunate result of behavioral variability within persons, however, is that the number of assessments required to obtain a sufficient level of reliability increases in proportion to the degree to which student behavior fluctuates over time. Indeed, from a test-theory perspective, variability in scores across assessments (under the same conditions) is considered error.

One potential solution to addressing high levels of time-related variance could lie in the scaling of assessment. Likert-type scales have been used across both studies of DBR-SIS and studies of DBR-MIS, with respondents asked to judge how frequently a behavior occurred using either a descriptive scale

(e.g., *never*, *sometimes*, *often*; Volpe & Briesch, 2012) or numeric scale (e.g., 0–10; Chafouleas, Briesch et al., 2010). Although this type of scale is a seemingly intuitive way of measuring teacher perceptions of behavior, ratings of frequency do not consider behavioral norms that may be important. For example, the collection of peer comparison data with direct observation is a way to determine whether the target student's behavior significantly deviates from what is typical in a given situation and therefore warrants intervention (e.g., Whitcomb & Merrell, 2012). Furthermore, when informants are asked to provide their impressions of student behavior using commercially available rating scales, judgments of a target behavior are often influenced by contextual factors. For example, because there is no absolute interpretation of what *often* means (a common anchor on behavior rating scales), rating an item as occurring *often* is likely influenced by the overall rates of the target behavior within the classroom (Reid & Maag, 1994).

It may likewise be helpful to obtain a normative comparison with DBR by asking the informant to rate a target student's behavior in comparison with other students in the classroom. Although there are a variety of comparative scaling approaches used in response-centered measurement (Crocker & Algina, 2008), to our knowledge, no previous work has investigated a scaling approach that explicitly asks raters to compare the behavior of one subject with that of others. Studies of DBR conducted in in vivo settings have highlighted substantial changes in overall student behavior across days (i.e., 16%–20% of variance; Chafouleas, Briesch et al., 2010), suggesting that some fluctuations in behavior are shared across students in the classroom rather than specific to the individual. It may therefore be possible to reduce the degree of rating variance attributable to situational variables by using a normative scale and accounting for this shared variability. The result is a clear departure from traditional rating scale assessment where one seeks to quantify an informant's perception of the frequency of a target student's behavior. Instead, the goal of a normative scale is to obtain an informant's perception of the deviance of a target student's behavior compared with a sample of classroom peers.

STUDY PURPOSE

The aim of this study is to extend the line of research investigating DBR-MIS by investigating the dependability of data generated via two alternative Likert-type scaling approaches involving a 7-point scale. One scale was designed to assess rater perceptions of the frequency of behaviors (e.g., *never* to *almost always*), whereas the comparison scale was designed to assess teacher perceptions of the frequency of behaviors demonstrated by a target student compared with peers (e.g., *much less* to *much more*). We hypothesized that obtaining a dependable estimate of disruptive behavior using the frequency-scaling approach would necessitate at least 2 weeks of ratings, given the large influence of time-related variance found by Volpe and Briesch (2012) using a similar scale. Furthermore, we hypothesized that error associated with differences in student behavior across time, in particular the Person $\times$ Occasion interaction, would be lower for scores generated via the latter approach (comparative scaling) and that this would lead to improvements in dependability and efficiency (i.e., fewer assessment occasions would be required to reach our criterion for dependability).

METHOD

Given that progress monitoring is most likely to be carried out with students at risk for, or currently demonstrating, behavioral problems, we attempted to ensure that the sample represented this target population. It was particularly important to represent students at risk for or demonstrating behavioral difficulties because previous research has shown that using a general sample of students may substantially overestimate the psychometric adequacy of resultant data (Briesch Volpe, & Ferguson, 2014).

Participants and Setting

Participants therefore consisted of nine seventh-grade students (five boys, four girls) who had been nominated by their classroom teacher for participation in a classroom intervention study (Briesch, Hemphill, & Daniels, 2013). Each student was enrolled in one of three general-education mathematics classes, which contained approximately 20 students and were taught by the same teacher. The students were judged by their teacher to demonstrate inadequate response to universal classroom behavior management practices, namely maximizing classroom structure, explicitly teaching behavioral expectations, and providing students with feedback regarding their demonstration of desired and undesired behavior.

All nine participants attended an urban, public, charter middle school in the Northeast comprised entirely of students of color, with more than 70% who were eligible for a free or reduced-price lunch. Ratings were conducted using videotaped footage of classroom instruction that was obtained in accordance with university human subjects institutional review board procedures within the larger intervention study.

Measures

Two scaling approaches (frequency and comparative) were used to assess student behavior using the items comprising a DBR-MIS measuring disruptive behavior developed by Volpe and Briesch (2012). Each disruptive behavior DBR-MIS was comprised of five items (calls out, is noisy, clowns around, talks to classmates when inappropriate, and is out of seat or area). The frequency scale was similar to the original scale used by Volpe and Briesch but included 7 points (0 = *never*, 1 = *rarely*, 2 = *occasionally*, 3 = *sometimes*, 4 = *often*, 5 = *very often*, 6 = *almost always*) instead of 6 points, given findings that at least seven scale gradients are optimal to detect changes in behavior over time (Christ et al., 2009). Instructions for the frequency scale were "Below is a list of behaviors that students may demonstrate in the classroom. Please read each item and rate how the child behaved during the observation interval." The comparative scale required raters to make normative comparisons in completing ratings by comparing the behavior of the target student with the other students in the classroom (0 = *much less*, 1 = *less*, 2 = *somewhat less*, 3 = *about the same*, 4 = *somewhat more*, 5 = *more*, 6 = *much more*). Instructions for the comparative scale were "Below is a list of behaviors that students may demonstrate in the classroom. Please read each item and rate how often the student exhibited the behavior during the observation period compared to other children in the classroom."

Procedures

Video clips gathered during the baseline phase of the aforementioned study were edited to obtain one 10-min continuous segment for each of 3 separate days. It was determined that 10-min segments would afford a sufficient sample of behavior while minimizing rater fatigue (given that each rater would view the same clip on multiple occasions). Clips were edited so that the general structure of the classroom routine was identical across segments. That is, all three video segments began when the class transitioned from independent seatwork to teacher-directed large-group instruction and ended once 10 min had elapsed.

Seven female graduate students served as raters in this study. All were in their first year of graduate study in school psychology and were enrolled in an introductory course in assessment that included brief coverage of rating scale assessment and systematic direct observation. Because these individuals had limited training in assessment procedures, they should be considered novice raters, and as such, they should be considered similar to classroom teachers who are commonly asked to collect DBR data. Raters observed the aforementioned video segments in a quiet room containing a bank of 10 work carrels. All raters attended a 1-hr training session, which involved watching video segments of student classroom behavior and completing ratings using the scales investigated in this study. Raters

also received instruction as to the order in which to rate students using the two scales. No criterion was established for rating student behavior during the training because the primary purpose of the session was to train the raters on study procedures. That is, we intended to study the technical characteristics under minimal training conditions given that training demands could represent a barrier to the adoption of a measurement system. Raters were instructed to conduct their observations and ratings when there were no distractions in the room and not to discuss cases with other raters.

Each rater was provided with detailed instructions including a matrix indicating the order by which students should be observed and the rating scale to be used for each rating. Each rater was instructed to observe one student at a time so that the rater conducted a total of 54 observations (9 students $\times$ 3 occasions $\times$ 2 methods). The forms were designed to counterbalance the order of students to be rated as well as the order of the rating scales. They were also designed to ensure that rating of the same student across two scales would be separated by many other observations. At the end of each 10-min video segment, raters were instructed to rate the student using one of the two scales.

Design and Analyses

Although traditional reliability coefficients only provide information for relative decision making based on rank order, GT can be used to estimate reliability-like coefficients for the purposes of both relative (i.e., generalizability coefficient, ρ^2) and absolute (i.e., dependability coefficient, Φ) decisions. Thus, GT was used to examine differences in dependability between the two scaling approaches. For a detailed discussion of GT, see Briesch, Swaminathan, Welsh, and Chafouleas (2014).

Variance component analyses were performed in SPSS, version 21.0, using analysis-of-variance Type III sum-of-squares estimation, and subsequent generalizability and dependability studies were conducted using Microsoft Excel. First, a model was examined to determine the proportion of variance associated with the two scaling approaches. In this model, person served as the object of measurement, and the three facets of scale (frequency and relative frequency), rater, and occasion were modeled in a fully crossed design (all students were observed on all occasions using both methods by all raters). Next, we conducted generalizability and dependability studies independently for each scale. In these models, person served as the object of measurement, and rater and occasion were modeled as random facets in fully crossed designs.

The goal for each generalizability study was to identify which facets or interactions contributed the largest percentage of rating variance, which could be useful in designing optimally efficient assessment procedures. Results of generalizability studies for each scale were used in a series of dependability studies to calculate reliability-like coefficients for the purposes of relative and absolute decision making. Dependability coefficients are the most relevant index in progress-monitoring applications because the evaluator is most interested in change within the individual (absolute decision making) as compared to the relative position of the individual in comparison with others (relative decision making). In the dependability studies reported later, we investigated how dependability would improve as a function of increasing the number of assessment occasions. Complete data were available for all seven raters.

RESULTS

We calculated proportions of variance in ratings of disruptive behavior in the initial-model generalizability study. The model involved nine students (i.e., person) being rated by seven raters using two scales across three occasions. The largest percentage of variance was attributable to the object of measurement (person, 31%), followed by scale (29%) and the interaction between person and occasion (10%). These findings indicate that intraindividual differences in disruptive behavior were recorded across both scales, that there were

Table 1. Mean Disruptive Behavior Ratings by Rater and Occasion

	Frequency Scale			Comparative Scale		
	Occasion 1	Occasion 2	Occasion 3	Occasion 1	Occasion 2	Occasion 3
Rater 1	8.78 (3.6)	13.89 (6.9)	11.00 (6.6)	13.56 (7.6)	18.67 (10.5)	12.44 (8.9)
Rater 2	9.78 (4.8)	15.00 (7.5)	10.33 (6.9)	17.67 (6.9)	19.44 (6.3)	16.11 (8.7)
Rater 3	10.11 (4.5)	12.11 (5.1)	9.78 (4.5)	19.89 (2.8)	20.11 (4.2)	17.78 (4.1)
Rater 4	9.22 (3.2)	11.67 (4.8)	10.00 (5.7)	19.89 (3.0)	19.22 (5.5)	15.89 (3.7)
Rater 5	12.00 (5.5)	14.67 (6.8)	11.78 (6.4)	20.78 (5.7)	21.67 (10.4)	18.44 (7.7)
Rater 6	9.33 (3.2)	13.67 (6.1)	10.11 (4.0)	18.67 (2.5)	18.67 (4.5)	15.00 (4.5)
Rater 7	13.11 (7.6)	17.33 (7.5)	12.44 (6.7)	18.78 (4.9)	20.56 (8.8)	14.78 (8.1)
Mean	10.33 (4.9)	14.05 (6.38)	10.78 (5.70)	18.46 (5.38)	19.76 (7.28)	15.78 (6.79)

Note. Numbers in parentheses are standard deviations.

notable differences in the level of scores across scale types (means and standard deviations for each method are summarized in Table 1), and that to some extent, students changed rank across occasions. Interactions including scale type accounted for relatively small proportions of variance (between 0% and 3%), which suggested that although the level of scores was somewhat different across the two scales, raters used the scales consistently and students tended to maintain their rank order across scales. Although the results of the full model provide important information, the purpose of this study was to examine important differences between the two scales included in the starting model. Therefore, reduced models (fully crossed Person × Rater × Occasion design) were examined for each scale type.

Reduced-Model Generalizability Studies

Separate generalizability studies were conducted for each scale type to examine the proportion of variance attributable to the object of measurement (students), the facets of occasion (day) and rater, and interactions between these facets (see Table 2). The object of measurement accounted for the largest proportion of variance for both the frequency and comparative scales (47% and 49%, respectively), which indicated that both scales were

Table 2. Variance Components for Reduced Models for Frequency and Comparative Scales

	Frequency Scale		Comparative Scale	
	Variance	% Variance	Variance	% Variance
Person	18.03	47.00	24.13	49.10
Occasion	3.32	8.60	3.28	6.70
Rater	1.70	4.40	1.82	3.70
Person × Occasion	6.32	16.50	6.07	12.30
Person × Rater	0.21	0.60	5.60	11.40
Occasion × Rater	0	0	0.31	0.60
Residual	8.80	22.90	7.97	16.20
Total	38.37	100	49.18	100

roughly equivalent in terms of their ability to discriminate interindividual differences in student disruptive behavior. The next largest source of variance for both scales was the three-way interaction (Person × Occasion × Rater) plus residual. The proportion of unexplained variance was somewhat higher for the frequency scale (23%) than for the comparative scale (16%). The interaction between person and occasion accounted for somewhat more variance for the frequency scale (17%) than for the comparative scale (12%), which indicated that changes in rank across students over days were slightly less pronounced when using the comparative scale. The proportion of variance attributed to the facet of rater was small across scales (4% for both scales), indicating only minor differences in how raters judged disruptive behavior overall, but there were notable differences in the consistency with which raters judged the disruptive behavior of different students (<1% for frequency, 11% for comparative). Finally, the proportion of variance attributable to the interaction between occasion and rater was small (<1% for both scales), indicating that raters were consistent in their overall judgments of disruptive behavior over time.

Reduced-Model Dependability Studies

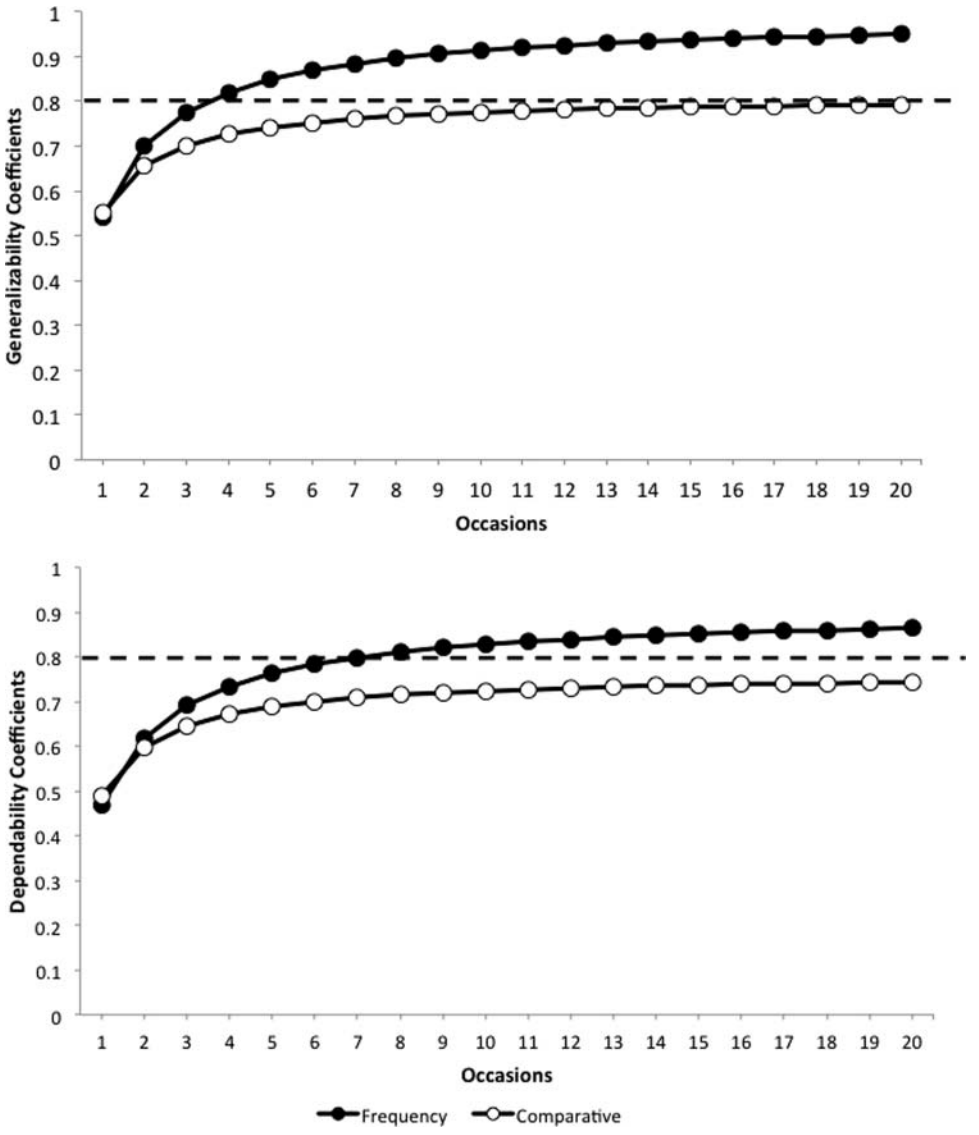
Using the variance components from the reduced models described earlier, we conducted a series of dependability studies to inform data collection for progress-monitoring purposes. First, for the starting models wherein seven raters completed ratings on three occasions, the frequency ($\rho^2 = .88$, $\Phi = .82$) and comparative ($\rho^2 = .88$, $\Phi = .84$) scales showed favorable levels of dependability for both relative and absolute decisions. Next, on the basis of typical school-based progress-monitoring practice, we examined how many assessment occasions would be required for scores from a single rater to achieve a reliability-like coefficient of .80 or greater. This criterion of .80 was selected because it has been recommended as a minimal standard for progress-monitoring purposes (Salvia, Ysseldyke, & Bolt, 2010). The results of the dependability analyses are summarized in Figure 1, in which generalizability coefficients are shown in the top panel and dependability coefficients are shown in the bottom panel.

Results for both relative and absolute decision making differed markedly between scales. First, regarding relative decision making, the comparative scale required more than 8 times as many assessment occasions as the frequency scale to reach the criterion of .80 (i.e., 33 versus 4). The reliability criterion for absolute decision making was reached after 8 assessment occasions for the frequency scale, but the criterion would not be reached even after 100 assessment occasions for the comparative scale. Dependability estimates across scales and indices rose quickly over the first five or six assessment occasions, reaching the more liberal criterion of .70 after three (relative) to seven (absolute) occasions.

DISCUSSION

The success of a multitiered model of school service delivery relies on the availability of tools to monitor student response to intervention that are both psychometrically defensible and feasible for repeated use. Within the context of behavioral assessment in particular, much attention has been paid in recent years to the development and evaluation of single- and multi-item DBR scales as a way to balance defensibility and feasibility. One recurring finding within the literature has been the dampening effect of time-related variance (i.e., changes in student behavior over time) on obtained generalizability and dependability coefficients. That is, the more that student behavior is found to fluctuate over time, the more rating occasions are needed to obtain a sufficient level of dependability. It was proposed that one potential way to reduce some of the noise in these data would be to rate the target student's behavior in comparison with other students in the classroom, which would potentially control for rating variance attributable to situational variables outside of the student (e.g., structure of classroom lesson). The present study therefore sought to determine whether the use of a comparative method of

Figure 1. Summary of Dependability Studies



scaling when completing ratings of disruptive behavior on a five-item DBR-MIS would result in improvements in the dependability of data as compared with a traditional frequency-based scaling approach such as *sometimes* and *often*.

Within the current study, descriptive statistics indicated that raters assigned higher mean scores of disruptive behavior when using the comparative scale (range = 13.56–21.67) as opposed to the frequency scale (range = 8.78–17.33). However, discrepancies in the level of ratings across scales are not unexpected given the differences across qualitative descriptors. For example, although a score of 3 reflects the midpoint of both scales, this score indicates that the behavior *sometimes* occurred on the frequency scale and that the target student exhibited the behavior *about the same* as other students in the classroom on the comparative scale. The same disruptive behavior may therefore occur infrequently in

an absolute sense but more frequently as compared with peers. Results of the full-model generalizability study further supported the fact that differences in rating across scales were substantial, with 29% of the total rating variance attributable to differences across the two scales.

Results of the reduced-model generalizability studies were overall fairly consistent across scales. The largest proportion of rating variance across both scales (47% for frequency, 49% for comparative) was attributable to differences in disruptive behavior across students, which was a desirable finding. In fact, both percentages were higher than those identified in previous research examining ratings of disruptive behavior conducted by research assistants (i.e., 29% variance in DBR-SIS ratings, Chafouleas, Briesch et al., 2010; 37% variance in DBR-MIS ratings, Volpe & Briesch, 2012). Furthermore, only 16% (comparative) to 23% (frequency) of the rating variance was left unexplained by the facets modeled. This finding was comparable with results from Volpe and Briesch (2012), wherein residual error accounted for 26% of the variance in disruptive behavior ratings.

Despite these similarities, one notable exception across scaling approaches involved the interaction between person and rater. Whereas the Person × Rater interaction accounted for a negligible amount of rating variance when using the frequency scale, this same interaction accounted for roughly 11% of the variance when using the comparative scale. This finding suggests that there was greater variability in raters' perceptions of particular students when using the comparative scale as opposed to the frequency scale. Person × Rater interaction effects have been documented in other studies using DBR in recent years. For example, Volpe and Briesch (2012) found that 10% of the variance in DBR-MIS ratings of disruptive behavior was attributable to this interaction when using a similar rater sample such as graduate research assistants. Likewise, in exploring the use of DBR-SIS with classroom teachers, Briesch, Chafouleas, and Riley-Tillman (2010) found that 20% of variance in ratings of academic engagement was attributable to a Person × Rater interaction. Both teachers in the Briesch et al. (2010) study were fairly consistent in their assessments of most students, but discrepancies were noted in ratings for the two students demonstrating the lowest levels of academic engagement. When the Person × Rater plots in the current study were examined, however, there did not appear to be greater inconsistencies among raters at specific levels of disruptive behavior.

An additional finding of interest involved the Person × Occasion interaction. It was hypothesized that the comparative scale would reduce the percentage of variance attributable to changes in student behavior over time (i.e., Person × Occasion interaction), but the size of this reduction was much smaller than expected. Whereas the Person × Occasion interaction accounted for 17% of the variance in ratings using the frequency scale, it accounted for 12% of the variance for ratings using the comparative scale.

One potential explanation for the substantial Person × Rater and Person × Occasion interactions with the comparative scale may be that the raters considered the behavior of all visible peers in the classroom. Because more than one student in each classroom had been identified by the teacher as being in need of intervention, some of the comparison students inevitably had elevated levels of disruptive behavior because the comparison group may have included other target students. It may be that selecting only typically developing peers for comparison purposes might have yielded more favorable results for the comparative scale. Several systematic direct-observation systems involve selecting peers in the classroom for this purpose. Typically, three or four randomly selected peers are observed to generate an estimate of average classroom behavior for that observation session (for a review of methods, see Volpe, DiPerna, Hintze, & Shapiro, 2005). Although the comparative scale did not perform as expected in terms of generating dependable estimates of student disruptive behavior, the results of the current study warrant further attention to this approach. Future studies of this scaling approach

should consider the use of a feasible training protocol for raters (Chafouleas, Riley-Tillman, Jaffery, Miller, & Harrison, 2015; Chafouleas et al., 2012; Harrison, Riley-Tillman, & Chafouleas, 2014), including specific attention to the selection of comparison peers.

Despite a number of similarities across variance component analyses, large discrepancies were identified regarding the number of rating occasions needed to achieve a .80 level of dependability when using a single novice rater. A total of 8 rating occasions were needed for absolute decision making using the frequency scale, which is comparable with previous recommendations concerning DBRs of disruptive behavior (i.e., 12 occasions; Volpe & Briesch, 2012). In contrast, an adequate level of dependability could not be reached using the comparative scale because of the large percentage of variance (16%) attributable to rater-related effects.

Results of this study highlight one of the key advantages of using GT over a classical test theory approach. That is, it is possible to isolate the specific sources of rating variance (i.e., generalizability study) and to use this information to suggest improvements in the measurement procedures (i.e., dependability studies). The largest source of alterable (i.e., nonperson, nonresidual) variance differed across scales. For the frequency scale, the size of the Person × Occasion interaction suggested that ratings would need to be collected over a greater number of days to improve dependability. Although this adds to informant load (Volpe, Briesch, & Gadow, 2011), the addition of rating occasions represents a realistic modification to data-collection procedures. However, both the Person × Occasion and Person × Rater interactions with the comparative scale were sizable. Thus, improvements in dependability would require increasing both the number of occasions and the number of raters. We modeled the use of a second rater in a supplementary dependability study and found dependability and generalizability coefficients that reached .80 after only six assessment occasions. Unfortunately, in many classroom settings, it may be logistically challenging to collect DBR data from more than one adult. If this cannot be achieved, one option may be to focus additional efforts on the training of raters to ensure that the scale is used consistently across target students.

Limitations and Future Directions

Findings of this study should be evaluated in the context of several limitations. First, although the five-item disruptive behavior scale was designed to obtain ratings of student classroom behaviors from teacher informants, first-year graduate students were used as raters in this study. Although disruptive student behavior is highly observable (cf. Volpe, McConaughy, & Hintze, 2009) and salient to teachers during typical instruction, the use of videotape over in vivo observations and the use of graduate student observers who could focus on student behavior without competing demands warrant consideration. Moreover, teachers are accustomed to rating student behavior and have relationships with students that may affect the way they rate students. As with any study, the investigator using GT must try to balance considerations for internal validity and external validity. A fully crossed design such as the one used in this study, wherein all raters rate each student across all occasions, allows modeling of each facet (raters, occasions) and interactions between these facets and interactions between these facets and the object of measurement (students). However, the investigator must decide which of two alternatives will yield more authentic (and hence informative) data. One option is to have more than one teacher rate many students during the same observation sessions, and the other is to capture student behavior on video and use as many raters as is deemed adequate. The second option was chosen for the current study given that this is a preliminary investigation of a novel scaling approach and because we were interested in investigating variability among a large group of raters.

Second, the length of the rating sessions used in this study is shorter than typically would be used in progress-monitoring applications. The 10-min duration was deemed adequate to address the primary research ques-

tions pertaining to the dependability of the two scaling approaches given that disruptive behavior typically occurs at a moderate rate in students referred for behavioral concerns. However, future studies of DBR-MIS should study teacher ratings over longer observation intervals (a full class period, a full school day). The length of the rating interval should be based on several considerations including the base rate of the behavior of interest and the purpose of assessment. The base rate of the behavior of interest is an important consideration in the assessment of social behavior (Gresham, Elliott, & Kettler, 2010; Meehl & Rosen, 1955). Behaviors with lower base rates (e.g., physical aggression) often are not suitable for rating intervals of shorter durations because the lack of variability between and within persons over short intervals impedes decision making. In addition, base rates have an influence on the accuracy of raters (Harrison et al., 2014).

Third, although traditional considerations related to statistical power do not apply given that GT analyses do not involve statistically significant null hypothesis testing, the total number of data points (i.e., 9 students $\times$ 7 raters $\times$ 3 occasions $=$ 189 data points) may seem small. Multiple examples of the application of GT to a small sample (i.e., roughly 10) can be found in the literature (Chafouleas, Briesch et al., 2010; Fawson, Reutzel, Smith, Ludlow, & Sudweeks, 2006; Hintze & Matthews, 2004; Marzano, 2002); however, one consequence of including too few data points within a complex design is that the resulting variance components may be negative or unstable (see Briesch, Swaminathan et al., 2014, for a discussion of these issues). Sample size was not deemed a significant issue within the current study, given that no issues with estimation were encountered during analyses and the obtained results were found to be fairly consistent with previous investigations (e.g., Volpe & Briesch, 2012). For example, for the frequency-scaling approach, both person variance and the number of occasions required to reach the criterion for dependability were fairly stable across the current study (47% and 8, respectively) and the Volpe and Briesch (2012) study (37% and 12, respectively). Nevertheless, findings of the current study are best viewed as preliminary. Future studies may strive to include a greater number of data points to assess the stability of the obtained variance components.

Finally, this study compared two scaling approaches applied to one five-item DBR-MIS designed to assess disruptive behavior. The extent to which findings of this study generalize to other measures of disruptive behavior and DBR assessment of other constructs are topics for future study.

Implications for School-Based Practice

The results of this study have several implications for progress monitoring using the disruptive behavior DBR-MIS in school settings. Results for the comparative scale, in which student behavior was judged in comparison with that of classroom peers, suggested that a satisfactory level of dependability could not be achieved with a single rater. Moving forward, attention will need to be paid to the development and implementation of training procedures to determine whether dependability can be improved. Caution is therefore prescribed for the use of this scaling approach at the present time. In contrast, the results for the traditional frequency scale (i.e., how often did the behavior occur?) both replicated and extended the work of Volpe and Briesch (2012) in support of the disruptive behavior DBR-MIS (using either a 6-point or 7-point scale) and added to the literature supporting the use of the frequency-scaling approach.

Given that progress monitoring involves examining changes in a target student's behavior over time (i.e., intraindividual decision making), obtained dependability coefficients are of greatest relevance to school-based practitioners. Research to date examining the frequency approach of scaling DBR-MIS suggests that between 8 (current study) and 12 (Volpe & Briesch, 2012) DBR-MIS ratings are needed to obtain a dependable estimate of disruptive behavior, and far fewer assessments (between 2 and 4) are needed to obtain a dependable estimate of academic engagement/

motivation (Volpe & Briesch, 2012). Although dependable estimates of these behaviors can be obtained efficiently, further studies are needed to examine other important technical characteristics of these scales including criterion-related validity and treatment sensitivity. To date, research on DBR-MIS for disruptive behavior and academic engagement/motivation looks promising, but further evidence of the aforementioned technical characteristics is needed before these scales can be recommended for school-based progress monitoring of social behavior.

REFERENCES

Briesch, A., & Volpe, R. J. (2007). Selecting progress monitoring tools for evaluating social behavior. *School Psychology Forum, 1*, 59–74.

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, 34*, 408–421.

Briesch, A. M., Hemphill, E. M., & Daniels, B. (2013). Check your SLANT: Use of self-management as a class-wide intervention. *School Psychology Forum, 7*, 29–39.

Briesch, A. M., Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J., (2012). The influence of alternative scale formats on the generalizability of data obtained from Direct Behavior Rating Single Item Scales (DBR-SIS). *Assessment for Effective Intervention, 38*, 127–133. doi:10.1177/1534508412441966

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*, 13–35. doi:10.1016/j.jsp.2013.11.008

Briesch, A. M., Volpe, R. J., & Ferguson, T. D. (2014). The influence of student characteristics on the dependability of observation data. *School Psychology Quarterly, 29*, 171–181.

Chafouleas, S. M. (2011). Direct behavior rating: A review of the issues and research in its development. *Education and Treatment of Children, 34*, 575–591.

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.

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. doi:10.1016/j.jsp.2013.04.002

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. doi:10.1016/j.jsp.2011.11.007

Chafouleas, S. M., Riley-Tillman, T. C., Jaffery, R., Miller, F. G., & Harrison, S. E. (2015). Preliminary investigation of the impact of a web-based module on direct behavior rating accuracy. *School Mental Health, 15*, 92–104. doi:10.1007/s12310-014-9130-z

Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010). Behavioral assessment within problem-solving models: Current status and future directions. *School Psychology Review, 39*, 343–349.

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. doi:10.1177/1534508409340390

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

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory.* Mason, OH: Cenage Learning.

Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements.* New York, NY: Wiley.

Fabiano, G. A., Vujnovic, R., Naylor, J., Pariseau, M., & Robins, M. (2009). An investigation of the technical adequacy of a Daily Behavior Report Card (DBRC) for monitoring progress of students with attention-deficit hyperactivity disorder in special education placements. *Assessment for Effective Intervention, 34*, 231–241.

Fawson, P. C., Reutzel, D. R., Smith, J. A., Ludlow, B. C., & Sudweeks, R. (2006). Examining the reliability of running records: Attaining generalizable results. *Journal of Educational Research, 100*, 113–126.

Gresham, F. M. (2014). Best practices in diagnosis of mental health and academic difficulties in a multitier problem-solving approach. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 147–158). Bethesda, MD, National Association of School Psychologists.

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.

Harrison, S. E., Riley-Tillman, T. C., & Chafouleas, S. M. (2014). Direct behavior rating: Considerations for rater accuracy. *Canadian Journal of School Psychology, 29*, 3–20.

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.

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.). (2016). *Handbook of response to intervention: The science and practice of assessment and intervention* (2nd ed.). New York, NY: Springer Science.

Kazdin, A. E. (2011). *Single case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., 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. doi:10.1016/j.jsp.2013.11.002

Kratochwill, T. R., & Bergan, J. R. (1990). *Behavioral consultation in applied settings: An individual guide.* New York, NY: Plenum.

LeBel, T. J., Chafouleas, S. M., Britner, P. A., & Simonsen, B. (2013). Use of a daily report card in an intervention package involving home-school communication to reduce disruptive behavior in preschoolers. *Journal of Positive Behavior Interventions, 15*, 103–112. doi:10.1177/1098300712440451

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*, 249–267.

Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.

National Center on Response to Intervention. (2010). *Essential components of RTI—A closer look at response to intervention.* Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

Reid, R., & Maag, J. W. (1994). How many fidgets in a pretty much: A critique of behavior rating scales for identifying students with ADHD. *Journal of School Psychology, 32*, 339–354.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Boston, MA: Houghton Mifflin.

Volpe, R. J., Briesch, A., & Gadow, K. D. (2011). The efficiency of behavior rating scales to assess disruptive classroom behavior: Applying generalizability theory to streamline assessment. *Journal of School Psychology, 49*, 131–155.

Volpe, R. J., & Briesch, A. M. (2012). Generalizability and dependability of single item and multiple item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review, 41*, 246–261.

Volpe, R. J., & Briesch, A. M. (2015). Multi-item direct behavior ratings: Dependability of two levels of assessment specificity. *School Psychology Quarterly, 30*, 431–442. doi:10.1037/spq0000115

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review, 34*, 454–474.

Volpe, R. J., & Gadow, K. D. (2010). Creating abbreviated rating scales to monitor classroom inattention-overactivity, aggression, and peer conflict: Reliability, validity, and treatment sensitivity. *School Psychology Review, 39*, 350–363.

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the Direct Observation Form. *School Psychology Review, 38*, 382–401.

Whitcomb, S., & Merrell, K. W. (2012). *Behavioral, social, and emotional assessment of children and adolescents* (4th ed.). New York, NY: Lawrence Erlbaum.

Date Received: October 28, 2014
Date Accepted: June 2, 2015
Associate Editor: Stephen P. Kilgus
Article accepted by previous Editor ■

Robert J. Volpe is an associate professor in the Department of Applied Psychology at Northeastern University and Co-Director of the Center for Research in School-based Prevention. His research focuses on designing academic and behavioral interventions for students with disruptive behavior disorders, as well as feasible systems for assessing student behavior in problem-solving models. He is President-Elect of the Society for the Study of School Psychology.

Amy M. Briesch is an associate professor in the Department of Applied Psychology at Northeastern University and Co-Director of the Center for Research in School-based Prevention. Her research interests involve the role of student involvement in intervention design and implementation, as well as the development of feasible and psychometrically sound measures for the assessment of student behavior in multitiered systems of support.