School Psychology Review,
2012, Volume 41, No. 3, pp. 246–261

Generalizability and Dependability of Single-Item and Multiple-Item Direct Behavior Rating Scales for Engagement and Disruptive Behavior

Robert J. Volpe and Amy M. Briesch
Northeastern University

Abstract. Direct behavior rating (DBR) has been described as a hybrid of systematic direct observation and behavior rating scales. Although single-item (DBR-SIS) and multi-item (DBR-MIS) methods have been advocated, the overwhelming majority of research attention has focused on DBR-SIS. This study employed generalizability theory to compare the dependability of the two DBR methods for two behaviors (Academic Engagement/Motivation and Disruptive Behavior). Two graduate students used both methods to rate 10-min video clips of the classroom behavior of 8 middle school students on three occasions. Generalizability of ratings was examined across raters and occasions, and decision studies were conducted to determine the minimal number of ratings necessary to obtain an acceptable level of dependability. Results favor the DBR-MIS method over the DBR-SIS method for making timely decisions across decision-making contexts. Results are discussed in terms of their implications for research and practice.

A refer-test-place model of academic and social-emotional assessment has predominated in school-based practice for decades, wherein emphasis has been placed on quantifying child deficits for the purpose of classification (Reschly, 2008). However, focus has recently shifted toward tiered models of prevention wherein every student is exposed to primary prevention efforts and those who are not responsive to these efforts receive higher intensity supports aligned with their needs (Tilly, 2008). The success of a preventative model depends heavily on the availability of appropriate measurement tools for assessing student response to intervention.

Direct behavior rating (DBR) is one method for measuring student response to intervention that has received substantial attention in recent years. There are two central features of DBR assessments: (a) the behavior is operationally defined and (b) a brief and low-inference rating of that behavior is conducted over a specified period (Christ, Riley-Tillman, & Chafouleas, 2009). As such, DBR has been considered a hybrid between systematic direct observation (informants complete the form in close proximity to the actual behavioral occurrence) and behavior rating scales (impressions from the observation are rated on a scale; Chafouleas, Riley-Tillman, & Christ, 2009). Although the DBR literature has

Correspondence concerning this article should be addressed to Robert Volpe, Northeastern University, 404 International Village, Boston, MA, 02115; e-mail: r.volpe@neu.edu

focused largely on academic engagement, disruptive behavior, and respectful behavior (Chafouleas, 2011), one of the often-emphasized strengths of DBR is its inherent flexibility. It has been noted that "DBR is not defined by a single scale, form, or number of rating items; rather, it is likely that lines of research will (and should) investigate multiple versions and applications of DBR as a method of assessment" (Chafouleas, Riley-Tillman, & Christ, 2009, p. 196). Although flexibility can certainly be viewed as an advantage of this method, it also means that much work must be carried out to validate different assessment approaches. To date, however, the psychometric research involving DBR has focused largely on the use of single-item scales (DBR-SIS), for which data summarization and interpretation take place at the individual item level. One preliminary study found that raters were generally more accurate when rating a single, globally worded item (e.g., academic engagement) than a single and more discrete behavior (e.g., raising hand; Riley-Tillman, Chafouleas, Christ, Briesch, & LeBel, 2009), suggesting that DBR-SIS measuring global behaviors may be the most efficient way to assess student behavior in an ongoing fashion (Christ et al., 2009). However, this research was limited in that the comparison was made at the level of an individual item. Unfortunately, it is not known how a single global item would compare to a composite score derived from multiple indicators of the construct of interest.

An important reason to investigate a DBR-MIS approach is that using multiple items to assess a construct may accelerate decision making. Researchers have previously suggested that between 7 and 10 ratings of a single item are needed to obtain a reliable estimate of behavior (Chafouleas, Riley-Tillman, & Sugai, 2007; Chafouleas, Christ, & Riley-Tillman, 2009; Chafouleas, Kilgus, & Hernandez, 2009). Although it has been argued that the brevity of a DBR assessment makes this data collection schedule reasonable, the need for 7–10 ratings indicate that classroom teachers must wait approximately 2 weeks before informed decisions can be made about student performance. Alternatively, recent research has suggested that fewer occasions may be required if utilizing a MIS. Volpe, Briesch, and Gadow (2011) found that 8–10 rating occasions were needed to achieve adequate levels of reliability using a SIS derived from the IOWA Conners Teacher Rating Scale (Loney & Milich, 1982). However, the number of necessary rating occasions quickly diminished as additional items were added, such that only three rating occasions were needed for a 4-item scale. Although preliminary, these results suggest that adequate levels of reliability may be achieved within a school week given that a sufficient number of items are employed.

An additional advantage of the MIS approach is that it is possible to customize each scale based on an individual's unique pattern of problem behavior (Volpe, Gadow, Blom-Hoffman, & Feinberg, 2009; Volpe & Gadow, 2010). Recent research has focused on developing change-sensitive brief behavior rating scales (Gresham et al., 2010; Meier, McDougal, & Bardos, 2008), but the scales were created to monitor general behavioral performance (i.e., general outcome measurement) rather than specific behavioral constructs (e.g., attention problems, disruptive behavior). Therefore, the same form is completed for all students regardless of the individual presenting problem, much like traditional rating scales. One noted limitation of such an approach is that the standard set of items may not adequately reflect those behaviors deemed most problematic for an individual student (Volpe et al., 2009). A DBR-MIS approach makes it is possible to select items that are most directly relevant to the presenting problem and student. As a result, it is likely not only that decision making surrounding intervention effectiveness will be enhanced, but also that the assessment procedures will be seen as more socially valid by potential consumers (Volpe et al., 2009, Volpe & Gadow, 2010).

Despite the potential advantages of a DBR-MIS approach, research to date has been limited. Several studies investigating school-based medication titration have reported the

use of brief rating scales administered one or more times per week (Volpe, Heick, & Guerasko-Moore, 2005), and Pelham (1993) has advocated for the daily administration of such scales. However, little data exist on the psychometric properties of measures used this way (Volpe et al., 2009; Volpe & Gadow, 2010). Moreover, with few exceptions, these studies have investigated scales composed of items that were not developed specifically for frequent administration or for monitoring the effects of intervention.

The purpose of the current study was to employ generalizability theory (Cronbach, Gleser, Rajaratnam, & Nanda, 1972) to comparatively examine the dependability of DBR-MIS and DBR-SIS designed to assess student behavior during classroom instruction. Rather than comparing a global DBR-SIS (e.g., academic engagement) to a DBR-SIS measuring one discrete behavior (e.g., raising hand; e.g., Riley-Tillman et al., 2009), we sought to compare DBR-MIS and DBR-SIS with comparable item content, in order to identify the most efficient and reliable method. Specifically, we developed individual DBR-MIS and DBR-SIS to measure both Academic Engagement/Motivation and Disruptive Behavior. These constructs were recently investigated by Chafouleas and colleagues (2010) in an examination of the dependability of DBR-SIS for 7 middle-school students across the facets of rater and occasions. Given the similarities in sample, design, and measures (with regard to DBR-SIS), the current study represents a partial replication and extension of that study. Generalizability and dependability studies were conducted to determine the number of rating occasions necessary within each method to achieve an acceptable level of dependability. This is an essential consideration as it relates both to how often informants must complete the rating task and how quickly data can be used to inform decision making. Given initial evidence that increasing the number of items may increase the accuracy of decision making (Volpe, Briesch, & Gadow, 2011), it was hypothesized that higher levels of dependability would be achieved for the DBR-MIS than for the comparable DBR-SIS.

Method

Participants and Setting

Participants included 8 (4 male, 4 female) seventh-grade students attending a public charter middle school, located in an urban setting in the Northeast. The school was comprised entirely of students of color, more than 70% of whom were eligible for free or reduced-price lunch. All participants were enrolled in the same general education math class, which contained roughly 20 students. Each day, the class followed the same structural format, beginning with independent seatwork followed by teacher-directed large group instruction and independent practice. Video footage of students in this classroom was collected as part of a previous study examining the effectiveness of a classwide intervention. For the purposes of the current investigation, 3 days of baseline video footage were selected and then edited to obtain one 10-min segment for each of 3 days. It was determined that a 10-min segment would provide raters with a sufficient sample of behavior while ensuring that the length of the rating task did not become overly burdensome.

Although the specific content of the lesson varied between days, the structure was identical across segments. Specifically, all three video segments began when the class transitioned from independent seatwork to teacher-directed large-group instruction and ended once 10 min had passed. The 8 students for whom observations were collected were selected because they were (a) clearly visible within the camera frame and (b) present across all 3 days of video footage.

DBR Measures

DBR measures were developed to assess academic engagement/motivation and disruptive behavior. Based on a review of the literature and a review of extant rating scales and observation measures assessing academic engagement (on-task) and off-task and disruptive classroom behavior, an initial pool of items (9 items for Academic Engagement/Motivation and 10 items for Disruptive Behavior) was

developed by the first author for the creation of multiple-item scales. Items were adapted from existing measures by rewording them (rating scale items) or creating one or more items from the operational definitions of observation measures. In selecting items, preference was given to items considered by the author to be fairly broad indicators of the construct of interest, to be readily observable, and to represent socially valid targets for intervention.

Each of the items was evaluated by a panel of nine doctoral students (most of which had experience with children with disruptive classroom behavior) and two faculty who rated each item on three dimensions using a 5-point Likert scale (1 = *Strongly Disagree*; 5 = *Strongly Agree*). The three dimensions rated were (a) criterion relatedness (how well each item was thought to measure either academic engagement/motivation or disruptive behavior), (b) observability (how observable the behaviors would be in a typical classroom), and (c) treatment validity (i.e., the degree to which the items represented malleable and socially valid targets for intervention). Ratings and recommendations for wording were considered when creating multiple-item scales for each of the two constructs of interest, although differences between ratings for deleted and retained items were small because the three dimensions guided the development of the initial pool of items. Average ratings for the retained items were between 4.31 (observability) and 4.58 (criterion relatedness) for Engagement/Motivation items and between 4.56 (criterion relatedness) and 4.72 (observability) for Disruptive Behavior items. Ratings for deleted items were only slightly lower for both the Academic Engagement/Motivation (between 3.91 and 4.34) and Disruptive Behavior (between 4.45 and 4.54) items.

DBR-MIS. Each DBR-MIS (Academic Engagement/Motivation and Disruptive Behavior) consisted of 5 items. On the Academic Engagement scale, the 5 items were "finishes work on time," "actively participates in class," "raises hand when appropriate," "works hard," and "stays on task." The 5 items on the Disruptive Behavior scale were "calls out," "noisy," "clowns around," "talks to classmates when inappropriate," and "out of seat/area." The scales were developed so that (a) each scale gradient was associated with a unique descriptor and (b) respondents would be able to meaningfully discriminate between points on the scale. Recent work related to DBR found that at least six scale gradients are needed to generate defensible data (Chafouleas, Christ, & Riley-Tillman, 2009; Christ et al., 2009) and that changing the number of DBR scale gradients (e.g., 5 vs. 10) results in minimal differences with regard to obtained scores (Briesch, Kilgus, Chafouleas, Riley-Tillman, & Christ, 2012). Given that use of fewer scale gradients may improve applied feasibility without sacrificing reliability, a 6-point scale was utilized ranging from N (*did not occur* = 0) to A (*occurred always* = 5).

DBR-SIS. Each DBR-SIS consisted of a single-item, which was rated using the 6-point scale described above. However, a brief definition of the overall construct was provided for the SIS. In addition, individual items from the corresponding DBR-MIS were provided as examples to support the general definition (see Appendix A).

Procedures

Two male school psychology doctoral students served as raters in the current study (hereafter referred to as raters). Raters observed video segments in a graduate student office that contained a bank of 10 student carrels, and they were instructed to observe one student at a time, to carry out the procedures when there were no distractions in the room, and to avoid discussing ratings with one another. Although these raters received no explicit training in conducting the DBR ratings, both had previous experience with completing DBR within their coursework and applied practice. At the end of each 10-min video segment, raters were instructed to rate the student on both constructs (Academic Engagement/Motivation, Disruptive Behavior) using one of the two DBR methods (DBR-MIS, DBR-SIS). The order of students, video seg-

Table 1
Full Model Variance Components and (Percent of Variance) Results across Scales (N = 8)

	Academic Engagement/ Motivation		Disruptive	
Facet	DBR-MIS	DBR-SIS	DBR-MIS	DBR-SIS
Person	532.33 (67)	27.31 (46)	49.81 (37)	15.25 (29)
Occasion	1.29 (0)	2.00 (3)	0.79 (1)	1.29 (2)
Rater	40.33 (5)	1.69 (3)	0.52 (0)	0.08 (0)
Person x Occasion	143.04 (18)	15.00 (25)	41.88 (31)	25.38 (49)
Person x Rater	35.67 (4)	3.48 (6)	5.31 (4)	5.25 (10)
Occ. x Rater	1.29 (0)	1.50 (3)	0.04 (0)	0.04 (0)
Person x Occasion x Rater + Residual	41.71 (5)	8.83 (15)	34.63 (26)	4.63 (9)

Note. DBR-MIS = direct behavior rating with multi item scales, DRB-SIS = direct behavior rating with single-item scales.

ments, DBR methods, and constructs were provided to raters in a table to ensure that all were counterbalanced to control for order effects and spillover across methods. Moreover, the order of observations was arranged to minimize the influence of one rating method on the other—that is, observations of the same student in the same segment using the two methods were separated by many observations. Each rater conducted a total of 48 observations (8 students × 3 occasions × 2 methods) over a period of 2 weeks.

Design and Analyses

Generalizability theory was used to conduct both generalizability (G) and decision (D) studies. Within a G study, the goal is to determine the percentage of rating variance attributable to relevant facets (e.g., raters, occasions) and interactions (Brennan, 2001). By understanding which facets or interactions contribute the largest percentage of rating variance, researchers may more effectively design future measurement procedures. Results of a G study are then used in a series of D studies to calculate reliability-like coefficients for the purposes of relative and absolute decision making. Within the context of a D study, researchers can also estimate how dependability would improve if particular aspects of the measurement situation were altered (e.g., increasing number of raters, increasing number of rating occasions).

Variance components were first derived in SPSS 18.0 using an analysis of variance with Type III sum of squares. In a series of subsequent decision studies, we then compared the dependability of each measurement approach examining the facets of rater and time (i.e., measurement occasions). For each method we conducted a series of decision (D) studies wherein we calculated both the generalizability coefficient (ρ, the relevant index for making relative decisions) and the dependability coefficient (ϕ, the relevant index for absolute decisions).

Results

G Studies

For both Academic Engagement/Motivation and Disruptive Behavior, separate G studies were conducted within each DBR method (i.e., DBR-MIS, DBR-SIS) to examine the proportion of variance independently attributable to the facets of person (i.e., student), rater, and occasion (i.e., day), as well as the interactions between these facets. Results for the starting model wherein 8 students were rated by two raters on three occasions are

presented in Table 1. For both methods of assessing Academic Engagement/Motivation, the greatest proportion of variance was attributable to the facet of person (67% DBR-MIS, 46% DBR-SIS). This finding suggests that although intraindividual differences in student engagement were identified across methods, the DBR-MIS was more sensitive to these differences (i.e., roughly 20% more variance attributable to persons for the DBR-MIS as compared to DBR-SIS).

The next largest source of variance was the person by occasion interaction (DBR-MIS = 18%, DBR-SIS = 25%) indicating changes in rank across persons over repeated assessments. Combined, the facet of rater (5% DBR-MIS, 3% DBR-SIS) and the interaction between person and rater (4% MIS, 6% SIS) accounted for a small, but notable, proportion of variance. As such, there were small differences in the ways that the two raters judged Academic Engagement/Motivation overall, as well as the way in which they rated particular students.

For both the DBR-MIS and DBR-SIS methods, variance attributable to occasion (0% and 3% respectively) and the interaction between occasion and rater (0% and 3% respectively) were relatively small. These findings indicate that the overall level of Academic Engagement/Motivation for the rated students as a group was fairly consistent, as was the overall consistency of raters across time. Finally, the three-way interaction between person, occasion, and rater, plus residual, accounted for approximately 5% of the variance for DBR-MIS and 15% for DBR-SIS.

Results were somewhat different for ratings on the Disruptive scales, with a smaller proportion of variance attributable to differences among students (37% DBR-MIS, 29% DBR-SIS) and a larger proportion explained by the interaction between persons and occasions (31% DBR-MIS, 49% DBR-SIS). Ratings indicated that the disruptive behavior of individual students fluctuated more across time than was noted for Academic Engagement/Motivation. Furthermore, these fluctuations were more pronounced with the DBR-SIS than the DBR-MIS. The only other notable interaction was that between persons and raters, which was somewhat higher for the DBR-SIS than the DBR-MIS (10% compared to 4% respectively). This reflects the degree to which raters were consistent with regard to their stringency across occasions. For all remaining facets and interactions (i.e., occasion, rater, occasion by rater), variance estimates were negligible. These findings indicate that the overall level of ratings on the Disruptive scale was fairly consistent across occasions and raters. Of interest, the three-way interaction plus residual was substantially larger for the Disruptive DBR-MIS as compared to the DBR-SIS (26% compared to 9%).

D Studies

Utilizing the results of the aforementioned G studies, we conducted a series of D studies to determine the number of assessment occasions necessary to achieve an acceptable level of reliability for each scale and method. For the starting model, wherein raters completed forms over three occasions, reliability-like coefficients for Academic Engagement/Motivation were found to be higher for the DBR-MIS (i.e., $\rho^2 = .85$, $\Phi = .82$) as compared to the DBR-SIS ($\rho^2 = .73$, $\Phi = .70$). Reliability-like coefficients for Disruptive were notably lower overall, although coefficients were again found to be higher for the DBR-MIS (i.e., $\rho^2 = .64$, $\Phi = .63$) as compared to the DBR-SIS ($\rho^2 = .50$, $\Phi = .49$).

Subsequent decision studies were conducted to estimate the number of times a single rater would need to provide ratings to reach a benchmark of .80. We chose .80 as a criterion because it is a commonly accepted criterion for both screening and progress monitoring decisions (Salvia, Ysseldyke, & Bolt, 2010). Results of the D studies are summarized in Figure 1 with the segmented horizontal line representing the .80 criterion.

For the Academic Engagement/Motivation scale, sizable differences were noted between the DBR-MIS and DBR-SIS methods, as well as between generalizability and dependability coefficients. For the DBR-MIS,

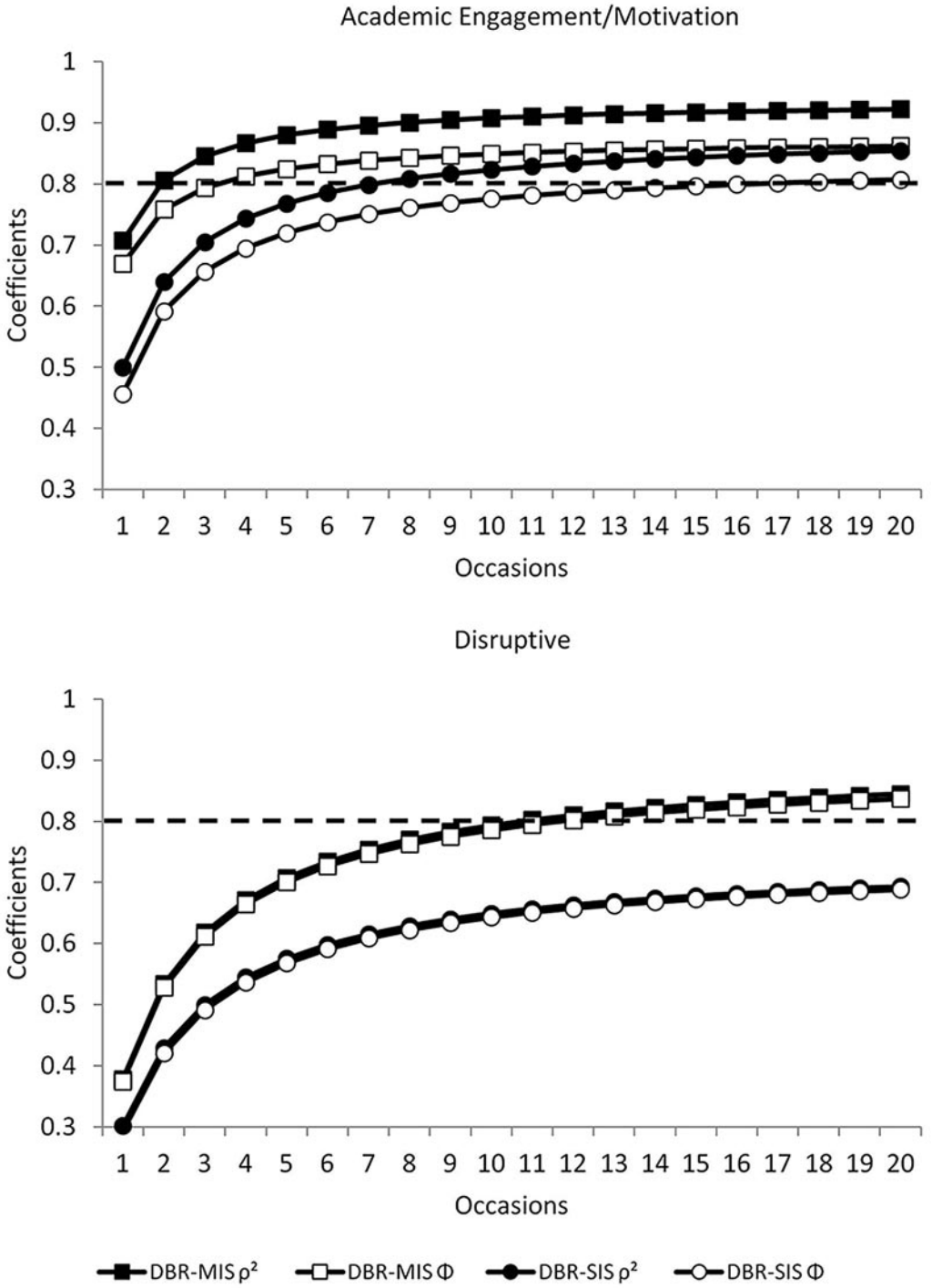

Figure 1. Summary of D studies.

the criterion was reached after only 2 occasions for relative decision-making (ρ^2) and after 2 occasions for absolute decision-making (Φ). In comparison, 8 and 17 occasions were

necessary to reach the criterion for relative and absolute decisions respectively using the DBR-SIS.

Results for relative and absolute decision-making purposes for the Disruptive scales were more consistent, but coefficients were substantially lower than those found for Academic Engagement/Motivation. For the DBR-MIS for Disruptive, the criterion was reached after 11 and 12 occasions respectively for relative and absolute decisions. In comparison, the reliability-like coefficients for the DBR-SIS had not reached .70 after 20 occasions ($\rho^2 = .69$; $\Phi = .69$), and negligible improvement was noted even when projecting out to 100 occasions ($\rho^2 = .73$; $\Phi = .73$).

Discussion

The purpose of the current study was to directly compare the dependability of ratings of common classroom behaviors using both a single-item and multi-item DBR approach. The principal aim of this study was to determine how many times one would be required to complete ratings before the obtained score would be sufficiently repeatable. The method requiring the fewest occasions to reach the criterion would be deemed the most efficient because it would take less time to achieve the desired level of dependability of measurement. Overall, results suggested that more efficient decision making was facilitated when using the DBR-MIS than the DBR-SIS. Important differences, however, were noted across the two behaviors evaluated, which are discussed next.

Across both behaviors, the overwhelming majority of variance was attributable to either overall differences in behavior across students or changes in the behavior of individual students across occasions. Although the aim of traditional assessment has been to maximize the percentage of variance attributable to the person facet, in behavioral assessment it is expected that the behavior of individuals will vary across time and settings (Cone, 1977). As such, it is desirable that more than two-thirds of the rating variance across scales and behaviors was attributable to either the facet of persons or the interaction between persons and occasions. It is also notable that greater variability in student behavior was observed across time when rating Disruptive (31% DBR-MIS, 49% DBR-SIS) than when rating Academic Engagement/Motivation (18% DBR-MIS, 25% DBR-SIS). This finding is consistent with previous research comparing DBR-SIS ratings of Academic Engagement and Disruptive Behavior conducted by research assistants within a middle school classroom (Chafouleas et al., 2010); however, the size of the discrepancy between the two behaviors was found to be much smaller in the 2010 study (14% Disruptive Behavior vs. 12% Academic Engagement/Motivation).

Facets included in the model (i.e., occasions, raters) explained all but a relatively small proportion of the variance in ratings. The smallest proportion of unexplained variance (5%) was noted for the DBR-MIS when rating Academic Engagement/Motivation, whereas the largest proportion (26%) was noted for the DBR-MIS when rating Disruptive Behavior. When research assistants in Chafouleas and colleagues' 2010 study conducted in vivo DBR-SIS ratings of these two constructs, roughly one-third of the variance was not explained by the model including rater, rating period, and day. Stronger findings related to DBR-SIS in the current study may be in part attributable to the fact that research assistants rated the behavior of 1 student at a time rather than conducting multiple ratings simultaneously (e.g., 7 students in the study by Chafouleas and colleagues, 2010). As such, raters were able to more closely attend to the behavior of a particular student rather than dividing their attention across several students. This suggests that future attention should likely be paid to better understanding how the dependability of measurement relevant to DBR changes as a function of the number of students being rated simultaneously.

Although their contribution to the overall model was relatively small, rater-related effects were noted for both scales and behaviors. First, a larger proportion of variance was attributable to overall differences across raters for both the MIS (5%) and SIS (3%) when

rating Academic Engagement/Motivation than Disruptive Behavior (0% for both scales), which seems to suggest that raters were more consistent overall when rating disruptive behavior than academic engagement. Previous findings have been mixed with regard to the influence of rater. Although negligible differences (i.e., 0%) were identified in the rating behavior of two research assistants evaluating academic engagement (Chafouleas et al., 2010), a larger proportion of variance (i.e., 8%) has been attributable to rater effects when examining in vivo teacher ratings of the same behavior (Briesch, Chafouleas, & Riley-Tillman, 2010; Chafouleas et al., 2010). However, the opposite was found when assessing disruptive behavior, in that the magnitude of the rater effect was slightly larger (3%) in the case of the research assistants than in the case of the teachers (1%). Although these combined results suggest that rater effects may be somewhat idiosyncratic, further research should be conducted to better understand whether particular features of a behavior (e.g., saliency) are related to greater inter-rater agreement.

Fairly consistent results were identified across scales and behaviors, however, related to the interaction between persons and raters. This interaction reflects small discrepancies (4–10% of variance explained) in the way that the two research assistants rated the behavior of specific individuals. This rater bias effect was present to a slightly greater degree for the DBR-SIS (10% Disruptive Behavior, 6% Academic Engagement/Motivation) than the DBR-MIS (4% for both scales). Although Chafouleas and colleagues (2010) found that this interaction did not explain any of the variance in ratings, evidence of notable rater bias has been identified in other work related to teacher-completed DBR (Briesch et al., 2010). Raters in the current study were not provided with explicit training in use of these scales; however, recent research has suggested that some training that incorporates modeling, practice, and feedback may improve the accuracy of DBR ratings (Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012).

Results of decision studies suggested that a dependable estimate of student behavior could be obtained much more quickly when utilizing the DBR-MIS than the DBR-SIS across decision-making contexts. If DBR-MIS scales were to be used to make intraindividual decisions, such as within the context of progress monitoring individual student performance, 4 occasions were needed to dependably measure Academic Engagement/Motivation, whereas 12 occasions were needed when measuring Disruptive Behavior. This observed discrepancy between scales can be explained in large part by differences in observed variability over time. That is, the disruptive behavior of particular students was observed to fluctuate to a greater degree over time (31% of variance attributable to person by occasion interaction) than was true for academic engagement (18% of variance attributable to person by occasion interaction). In contrast, when using the DBR-SIS, a dependable estimate of academic engagement was obtained after 17 rating occasions, the .80 criterion was not met for disruptive behavior even after 100 occasions. Similarly, the weaker dependability estimates obtained for the DBR-SIS scales can be explained to some degree by the large percentages of variance attributable to changes in academic engagement (25%) and disruptive behavior (49%) over time. Further research is therefore warranted to understand whether the DBR-SIS scales are more sensitive to actual behavior change over time or whether the rank order of students changes more across occasions given the global nature of the evaluation. A similar pattern of results was identified for the purposes of rank-order decision making (e.g., screening), as depicted in Figure 1.

Obtained coefficients for the DBR-SIS Academic Engagement/Motivation scale were fairly comparable to those previously identified when utilizing external raters (i.e., research assistants). In the study by Chafouleas and colleagues (2010), 7 (relative) to 20 (absolute) occasions were recommended, whereas 8 (relative) to 17 (absolute) were needed in the current investigation. It is worth noting, however, that when teachers have completed DBR-SIS ratings in vivo, dependability estimates have been found to be much weaker. In fact, across two studies (Briesch et

al., 2010; Chafouleas et al., 2010), absolute coefficients remained at a level of approximately .60 subsequent to 20 days of data collection when rating Academic Engagement. Results relevant to Disruptive Behavior have been slightly more promising, with dependability coefficients subsequent to 20 days of data collection ranging from .68 (current study) to .80 (Chafouleas et al., 2010). Furthermore, although Volpe and colleagues (2011) investigated a different set of constructs, results were similar in that a minimum of 20 occasions were needed when using a SIS. Taken together, these findings appear to suggest that it may not be possible to dependably estimate student behavior utilizing single-item scales within a time frame conducive to progress monitoring. Given the more reasonable estimates obtained within the context of relative (i.e., interindividual) decision making across studies, single-item scales may be better suited for rank-order purposes. Overall, the findings of this study support our hypothesis that higher levels of dependability would be achieved for the DBR-MIS than for comparable DBR-SIS, and illustrate how reliance on DBR-SIS could lead to delays in decision making.

Limitations and Future Directions

Although efforts were made to address potential threats to internal and external validity, limitations of the current study should be noted. First, observations were conducted by graduate research assistants who did not have the competing demands on their attention typical in actual classroom settings. Although there exists preliminary research to suggest that the overall pattern of DBR ratings may be similar between teachers and external raters (e.g., research assistants; Chafouleas et al., 2010), it is notable that a larger percentage of rating variance remained unexplained when examining teacher ratings (48% Academic Engagement, 52% Disruptive Behavior) than those conducted by research assistants (32% Academic Engagement, 34% Disruptive Behavior). In addition, although the raters in this study had no familiarity with the students they were asked to rate, teachers would typically have a preexisting relationship with their ratees. As such, we cannot discount the fact that rater biases such as halo or stringency effects would likely be more influential in the context of a preexisting relationship. Although the procedural conditions in the current study (i.e., use of research assistants, videotaped instruction) were deemed necessary to conduct multiple ratings of the same sample of behavior, additional research is certainly warranted to explore the extent to which the current results replicate when actual teachers are responsible for conducting ratings.

Second, the number of measurements within a facet, and consequently the number of overall data points, was somewhat small (8 students $\times$ 2 raters $\times$ 3 occasions $=$ 48 data points). It has been noted that variance component estimates typically become more stable as the number of data points increases (Smith, 1981); however, guidelines regarding the minimum number of data points needed to detect an effect do not exist as they do for other statistical analyses. Within the context of the current study, one goal was to ensure that the number of data points was sufficient to run the model. We balanced this goal against the concern that we did not want to overburden our raters as this might lead to an artificial elevation in the error variance related to time. Given minimal differences identified between raters, emphasis in future studies may instead be placed on collecting DBR data for a larger number of students across a greater number of occasions.

Lastly, more research is needed to investigate other psychometric indicators for the scales investigated in this study, including criterion-related validity with stringently evaluated measures of the constructs of interest (systematic direct observation, commercially available rating scales) and treatment sensitivity. Moreover, the pool of DBR scales should be expanded to other constructs that typically are targets for school-based intervention (social skills, inattention, overactivity, interpersonal aggression, and academic behaviors other than engagement) or relate to intervention side effects

Table 2.
Comparison across Direct Behavior Rating (DBR) Studies

	Briesch et al. (2010)	Chafouleas et al. (2010)	Current Study
Population	12 kindergarten students	7 middle school students	8 middle school students
Design	Persons x Raters x Occasions/Day	Persons x Raters x Occasions/Day	Persons x Raters x Occasion/Day
DBR scale	Single-item scale (SIS) with continuous line	SIS with continuous line	SIS and multi-item scale (MIS) with 6-point scale
Rater type	Classroom teachers	Classroom teachers and research assistants (RAs)	RAs
Length of observation	15 minutes	15 minutes	10 minutes
Number of observations needed to facilitate intra-individual decision making (.80)	Teachers: Academic Engagement SIS > 100	Teachers: Academic Engagement SIS: > 20 Disruptive Behavior SIS: > 20	Teachers: N/A
	RAs: N/A	RAs: Academic Engagement SIS: 20 Disruptive Behavior SIS: 15-20	Academic Engagement SIS: 17 Academic Engagement MIS: 4 Disruptive Behavior SIS: > 20 Disruptive Behavior MIS: 12

N/A = Not applicable.

(e.g., social withdrawal, medication side effects; cf. Volpe & Gadow, 2010).

Potential Implications for Practice

Results of the current study add to the extant literature supporting the use of DBR within the realm of behavioral assessment. Although there exists psychometric evidence to support the use of single-item DBR to measure student behavior (Briesch et al., 2010; Chafouleas et al., 2010), it is notable that between 7 and 10 ratings have typically been recommended to obtain a dependable estimate of student behavior (Chafouleas, Christ, & Riley-Tillman, 2009). An examination of Table 2 would suggest that using single-item scales to measure the constructs of engagement and disruptive behavior would require 4 weeks or more of daily data collection before dependable decisions can be made regarding student performance. As evidenced in the current study, however, use of a MIS substantially decreases the number of rating occasions needed to obtain an adequate level of dependability.

One limitation of the use of DBR-SIS is that it represents only one level of measurement of interest to decision makers. However, both global general objectives and specific performance objectives should be monitored when evaluating student response to school-based interventions (Kratochwill & Bergan, 1990; Volpe, Briesch, & Chafouleas, 2010). General social or academic functioning may be described through the measurement of global objectives (e.g., Pelham, Fabiano, & Massetti, 2005), but global objectives are less sensitive to the effects of treatment (Fabiano et al., 2007). However, performance objectives are closely aligned with treatment targets (e.g., raises hand) and are highly sensitive short-term indicators of response to intervention (Volpe & Gadow, 2010). If assessment is focused on broad domains alone, then important changes in specific child behaviors may be missed, and if measurement is restricted to one or two specific target behaviors, then the potential broader effects of intervention may not be fully detected.

Using a multi-item scale (DBR-MIS) seems to be a potentially psychometrically sound approach that combines specificity with global measurement. Raters within a DBR-MIS approach are asked to assess several specific indicators (e.g., talking out, out of seat) of a more general behavioral construct (e.g., disruptive behavior; Christ et al., 2009). As a result, it is possible to obtain information regarding student performance at two levels. In addition to monitoring specific behaviors using the individual item data, all items within a scale could be summed to provide a global measure of that construct. In addition, several constructs could be combined to produce a composite score, representing the student's overall level of functioning, and aggregate data could be used to monitor overall performance over time. Although these applications might meet several important assessment needs, a substantial amount of research is required to evaluate the psychometric characteristics of such methods. Caution should therefore be exercised in adopting the scales developed in this study until further evaluation is carried out.

References

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

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.

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. doi:*10.1177/1534508412441966

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., Christ, T. J., & Riley-Tillman, T. C. (2009). Generalizability of scaling gradients on Direct Behavior Ratings. *Educational & Psychological Measurement, 69,* 157–173.

Chafouleas, S. M., Kilgus, S. P., & Hernandez, P. (2009). Using Direct Behavior Rating (DBR) to screen for school social risk. *Assessment for Effective Intervention, 34,* 214–223.

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

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

Chafouleas, S. M., Riley-Tillman, T. C., & Sugai, G. (2007). *School-based behavioral assessment: Informing intervention and instruction.* New York: Guilford Press.

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.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8,* 411–426.

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

Fabiano, G. A., Pelham, W. E., Gnagy, E. M., Burrows-MacLean, L., Coles, E. K., Chacko, A., et al. (2007). The single and combined effects of multiple intensities of behavior modification and multiple intensities of methylphenidate in a classroom setting. *School Psychology Review, 36,* 195–216.

Gresham, F. M., Cook, C. R., Collins, T., Dart, E., Rasetshwane, K., Truelson, E., & Grant, S. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating System Teacher Form. *School Psychology Review, 39,* 364–379.

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

Loney, J., & Milich, R. (1982). Hyperactivity, inattention, and aggression in clinical practice. In M. Wolraich & D. K. Routh (Eds.), *Advances in developmental and behavioral pediatrics* (Vol. 3, pp. 113–147). Greenwich, CT: JAI Press.

Meier, S. T., McDougal, J. L., & Bardos, A. (2008). Development of a change-sensitive outcome measure for children receiving counseling. *Canadian Journal of School Psychology, 23,* 148–160.

Pelham, W. E. (1993). Pharmacotherapy for children with attention-deficit hyperactivity disorder. *School Psychology Review, 21,* 199–223.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of attention-deficit hyperactivity disorder in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34,* 449–476.

Reschly, D. J. (2008). School psychology RTI paradigm shift and beyond. In A. Thomas & J. Grimes (Eds.) *Best practices in school psychology* (5th ed., pp. 3–15). Bethesda, MD: National Association of School Psychologists.

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

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

Smith, P. L. (1981). Gaining statistical accuracy in generalizability theory: The use of multiple designs. *Journal of Educational Measurement, 18*, 147–154.

Tilly, W. D. (2008). The evolution of school psychology to science-based practice. In A. Thomas & J. Grimes (Eds.) *Best practices in school psychology* (5th ed., pp. 17–36). Bethesda, MD: National Association of School Psychologists.

Volpe, R. J., Briesch, A. M., & Chafouleas, S. M. (2010). Linking screening for emotional and behavioral problems to problem-solving efforts: An adaptive model of behavioral assessment. *Assessment for Effective Intervention, 35,* 240–244.

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., & 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., Gadow, K. D., Blom-Hoffman, J., & Feinberg, A. B. (2009). Factor analytic and individualized approaches to constructing brief measures of ADHD behaviors. *Journal of Emotional and Behavioral Disorders, 17,* 118–128.

Volpe, R. J., Heick, P., & Gureasko-Moore, D. (2005). An agile model for monitoring the effects of stimulant medication in schools. *Psychology in the Schools, 42,* 509–523.

Appendix A

Disruptive- MIS

Directions: 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 period.

Circle **N** if the behavior did not occur
Circle **R** if the behavior rarely occurred or for a slight or ambiguous occurrence
Circle **S** if the behavior occurred sometimes
Circle **O** if the behavior occurred often
Circle **V** if the behavior occurred very often
Circle **A** if the behavior occurred always during the observation period

1. Calls out

N **R** **S** **O** **V** **A**

2. Noisy

N **R** **S** **O** **V** **A**

3. Clowns around

N **R** **S** **O** **V** **A**

4. Talks to classmates when inappropriate

N **R** **S** **O** **V** **A**

5. Out of seat/area

N **R** **S** **O** **V** **A**

Disruptive- SIS

Disruptive is intended to measure the degree to which student behaviors are distracting to others or otherwise interfere with student learning. Examples include calls out, noisy, clowns around, talks to classmates when inappropriate, and out of seat/area.

Directions: Circle the letter that best reflects the student's behavior during the observation period.
Circle **N** if the behavior did not occur
Circle **R** if the behavior rarely occurred or for a slight or ambiguous occurrence
Circle **S** if the behavior occurred sometimes
Circle **O** if the behavior occurred often
Circle **V** if the behavior occurred very often
Circle **A** if the behavior occurred always during the observation period

Disruptive

N **R** **S** **O** **V** **A**

Appendix B

Academic Engagement/Motivation - MIS

Directions: 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 period.

Circle **N** if the behavior did not occur
Circle **R** if the behavior rarely occurred or for a slight or ambiguous occurrence
Circle **S** if the behavior occurred sometimes
Circle **O** if the behavior occurred often
Circle **V** if the behavior occurred very often
Circle **A** if the behavior occurred always during the observation period

1. Finishes work on time

N **R** **S** **O** **V** **A**

2. Actively participates in class

N **R** **S** **O** **V** **A**

3. Raises hand when appropriate

N **R** **S** **O** **V** **A**

4. Works hard

N **R** **S** **O** **V** **A**

5. Stays on task

N **R** **S** **O** **V** **A**

Academic Engagement/Motivation - SIS

Academic engagement/Motivation is intended to measure the degree to which students are actively and enthusiastically engaged in academic tasks. Examples include, finishes work on time, actively participates in class, raises hand when appropriate, works hard, and stays on task.
Directions: Circle the letter that best reflects the student's behavior during the observation period.
Circle **N** if the behavior did not occur
Circle **R** if the behavior rarely occurred or for a slight or ambiguous occurrence
Circle **S** if the behavior occurred sometimes
Circle **O** if the behavior occurred often
Circle **V** if the behavior occurred very often
Circle **A** if the behavior occurred always during the observation period

Academic Engagement/Motivation

N **R** **S** **O** **V** **A**

Date Received: February 29, 2012
Date Accepted: June 5, 2012
Action Editor: Amanda VanDerHeyden ■

Robert J. Volpe is an associate professor in the Bouvé College of Health Sciences at Northeastern University. His primary research interests concern academic problems experienced by children with ADHD, and the development and evaluation of feasible behavioral assessment and intervention procedures.

Amy M. Briesch is an assistant professor in the Bouvé College of Health Sciences at Northeastern University. Her research interests include the identification and examination of feasible and psychometrically sound measures for the formative assessment of student social behavior; the use of self-management as an intervention strategy for reducing problem behaviors in the classroom; and the role of student involvement in intervention design and implementation.

Accepted Manuscripts for Forthcoming Issues

Dobbs-Oats, J., & Robinson, C. *Preschoolers' Mathematics Skills and Behavior: Analysis of a National Sample*

Dart, E., Cook, C., Collins, T., Gresham, F., & Chenier, J. *Test-driving Interventions to Increase Treatment Integrity and Student Outcomes*

DuPaul, G., Eckert, T., & Vilardo, B. *The Effects of School-Based Interventions for Attention-Deficit/Hyperactivity Disorder: A Meta-Analysis 1996–2010*

Hansen, B., & Wills, H. *Effects of Aligning Self-Management Interventions with Functional Behavioral Assessment*

Gutierrez, G., & Vanderwood, M. *A Growth Curve Analysis of Literacy Performance among Second-Grade, Spanish-Speaking, English-Language Learners*