Hammill Institute on Disabilities

Journal of Emotional and Behavioral Disorders
2021, Vol. 29(2) 71–82

DOI: 10.1177/1063426619882345
jebd.sagepub.com
SAGE

Content Validation of Direct Behavior Rating Multi-Item Scales for Assessing Problem Behaviors

Brian Daniels, PhD[1], Amy M. Briesch, PhD[2], Robert J. Volpe, PhD[2], and Julie Sarno Owens, PhD[3]

Abstract

Direct Behavior Rating (DBR) is an efficient method for monitoring changes in student behavior in response to intervention. Emerging research on DBR Multi-Item Scales (DBR-MIS) indicates that DBR-MIS has promising characteristics as a progress-monitoring assessment. Specifically, the multiple items within DBR-MIS allow stakeholders to measure specific behaviors at the item level, as well as global constructs at the scale level. In addition, studies have shown that fewer rating occasions are necessary to reach acceptable levels of dependability when using DBR-MIS as opposed to single-item scales (DBR-SIS). The purpose of the study was to develop and validate DBR-MIS problem behavior scales (Disruptive, Oppositional, Interpersonal Conflict, and Conduct Problems) that may be used to evaluate students' response to social-emotional or behavioral intervention. Within the first phase of development, item content was generated and subjected to evaluation by panels of researchers and school-based consumers. Exploratory factor analysis (EFA) was then used in the second phase to identify items that represented the strongest indicators of each construct. Teachers (N = 307) in Grades K–3, from 35 school districts across 13 states in the Northeastern, Midwestern, Southern, and Southwestern United States each completed ratings for one randomly selected student (N = 307). Results of the EFA using a starting pool of nine to 11 items for each DBR-MIS initially indicated one-factor solutions for the Disruptive and Oppositional scales and a two-factor solution for the Interpersonal Conflict scale. Consequently, a new Conduct Problems scale was created from items loading on the second factor. Implications for progress monitoring and future research are discussed.

Keywords

behavioral, assessment, externalizing, behavior(s), rating scales, behavior(s), psychometrics

An essential component of multitiered systems of support (MTSS) is the frequent and repeated assessment of student performance over time to evaluate the effectiveness of interventions and inform subsequent decisions (Sugai & Horner, 2009). Although numerous assessment tools are available for measuring growth in academic skills, few technically adequate assessment tools exist for measuring changes in student behavior (National Center on Intensive Intervention, 2018). Commonly used behavior assessment methods, such as systematic direct observation (SDO) and behavior rating scales have several limitations that limit their appropriateness and usability as progress-monitoring tools in applied (i.e., nonresearch) settings. For example, SDO requires external personnel (e.g., a school psychologist) and substantial time to conduct observations, particularly when multiple students' behavior is monitored (Briesch & Volpe, 2007). In addition, SDO data reflect student behavior during a short time frame within the school day when the observer is present, and thus generalizability of resultant data across other situations or occasions may be limited (Briesch, Chafouleas, & Riley-Tillman, 2010; Hintze & Matthews, 2004). On the contrary, behavior rating scales, which were originally developed for diagnostic purposes, may not be sensitive to short-term changes in student behavior. Furthermore, the time necessary to repeatedly complete rating scales may be prohibitive, especially as the number of students being monitored increases (Volpe & Gadow, 2010; Volpe, Gadow, Blom-Hoffman, & Feinberg, 2009).

Direct Behavior Rating (DBR) has emerged as a feasible and acceptable assessment method that is sensitive to changes in student behavior following intervention and has some advantages over traditional rating scales and SDO (Chafouleas, 2011; Chafouleas, Riley-Tillman, & Christ, 2009). DBR

[1]University of Massachusetts Boston, MA, USA
[2]Northeastern University, Boston, MA, USA
[3]Ohio University, Athens, OH, USA

Corresponding Author:
Brian Daniels, Department of Counseling and School Psychology, University of Massachusetts Boston, Boston, MA 02125, USA.
Email: Brian.Daniels@umb.edu

focuses on directly rating observable, operationally defined behaviors in close proximity to the time at which they are exhibited in the setting(s) of interest (Christ, Riley-Tillman, & Chafouleas, 2009). DBR has several strengths that make it useful for school-based progress monitoring. First, DBR is completed by a teacher or someone else who spends a substantial amount of time with the target student in the context of interest. Thus, DBR need not burden additional personnel, as is the case with SDO. Second, the fact that the rater has the opportunity to observe the target student over an extended period of time (e.g., entire school day) means that DBR can be used to assess low-frequency behaviors that may not be captured through SDO. Finally, less inference is likely involved when using DBR to assess student behavior when compared with traditional behavior rating scales because ratings are conducted in close temporal proximity to their occurrence (i.e., immediately after the time frame of interest), whereas traditional behavior rating scales ask informants to summarize behavior exhibited over several months.

Within the broader category of DBR, there are two general methods for which psychometric evidence has been established. DBR Single-Item Scales (DBR-SIS), which measure a global construct using a single item (i.e., *Academically Engaged, Disruptive*, and *Respectful*), have been researched most extensively. Numerous studies have demonstrated the reliability/dependability (Chafouleas et al., 2010; Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007), validity (Riley-Tillman, Chafouleas, Sassu, Chanese, & Glazer, 2008), and treatment sensitivity (Chafouleas, Sanetti, Kilgus, & Maggin, 2012; Fabiano, Pyle, Kelty, & Parham, 2017) of DBR-SIS when measuring these three constructs. DBR Multi-Item Scales (DBR-MIS), on the contrary, include several items that assess specific behaviors (e.g., calls out or out of seat) and can be summed to obtain a composite score measuring a broad construct (e.g., Disruptive Behavior; Volpe & Briesch, 2012). The primary advantages of DBR-MIS are that they can be interpreted at either the item- or composite-level (Volpe & Briesch, 2015) and acceptable dependability for progress monitoring may be achieved in relatively few rating occasions (Volpe & Briesch, 2012, 2016). However, a limitation of DBR in general is that psychometric characteristics only have been established for only a small number of constructs.

Just as the bulk of psychometric evidence for DBR-SIS exists for the constructs of Academic Engagement and Disruptive Behavior (Briesch, Chafouleas, & Riley-Tillman, 2016), prior studies of DBR-MIS have largely focused on these two constructs as well (Volpe & Briesch, 2012, 2015, 2016). Whereas one study found that as few as four ratings of Academic Engagement may be needed to achieve adequate levels of dependability (i.e., .80) for progress-monitoring purposes (Volpe & Briesch, 2012), dependability results for DBR-MIS designed to measure Disruptive Behavior have been somewhat more variable (i.e., 5–12 ratings; Volpe & Briesch, 2012, 2015, 2016) However, two studies have found promising evidence for a brief DBR-MIS assessing Inattention/Overactivity, Aggression, and Peer Conflict (Daniels, Volpe, Briesch, & Gadow, 2017; Volpe & Gadow, 2010).

Although these studies provide initial support for the use of DBR-MIS to assess additional behavioral targets, broadening DBR assessment to include other constructs that may interfere with student academic and social functioning is warranted. In particular, externalizing problems are logical targets for the development of DBR-MIS because aggressive, oppositional, and disruptive behaviors in early childhood are associated with significant negative social and academic outcomes including limited peer relationships (Wehmeier, Schacht, & Barkley, 2010), increased risk for grade retention (Barbaresi, Katusic, Colligan, Weaver, & Jacobsen, 2007), placement in self-contained special education classrooms (Bierman et al., 2013), and academic underachievement (Frick et al., 1991; Nelson, Benner, & Gonzalez, 2003). In addition, they are problems commonly reported by teachers. In one study, a least 90% of teachers reported disruptive and defiant behaviors as major concerns during the prior school year, and 78% of the same teachers indicated aggressive behavior as a major concern (Reinke, Stormont, Herman, Puri, & Goel, 2011). Furthermore, there is evidence that proactive interventions (e.g., Schoolwide Positive Behavior Support) reduce externalizing behaviors (Bradshaw, Mitchell, & Leaf, 2010; Horner et al., 2009), which in turn is associated with improved academic skills (DiPerna, Volpe, & Elliott, 2002; Myers, Simonsen, & Sugai, 2011). These reductions in externalizing problems indirectly affect academic performance by increasing the amount of time students spend actively engaged in instruction (Horner et al., 2009). However, feasible and technically adequate assessment tools, such as DBR-MIS, are needed to monitor student progress with regard to specific externalizing behavior constructs over time and facilitate data-based evaluation of interventions.

Finally, although externalizing behaviors are often discussed in terms of a unitary overarching construct, they encompass several related but distinct classes of behavior (e.g., aggression and conduct problems). Results of prior factor analytic studies indicate that hyperactive/impulsive, oppositional, and conduct problems are related but distinct constructs (Burns, Walsh, Owen, & Snell, 1997; Hartman et al., 2001). More specifically, Hartman et al. (2001) conducted confirmatory factor analysis using behavior rating scale data from the *Teacher Report Form* (Achenbach, 1991), *Ontario Child Health Study Scales–Revised* (OCHS-R; (Boyle et al., 1993; Macleod, McNamee, Boyle, Offord, & Friedrich, 1999), and *Child Symptom Inventory–4* (Gadow & Sprafkin, 1997), for 11 samples of clinically referred and nonreferred Canadian, Dutch, and American youth (ages 6–17 years). Results indicated that four distinct

dimensions related to externalizing problems: *Problems with Attention* (e.g., distractibility, inattention, disorganization, or failure to complete work), *Hyperactivity/Impulsivity* (e.g., interrupting or disrupting others, impulsivity, or excessive motor activity or movement), *Conduct Disorder* (e.g., physical aggression, fighting, bullying, stealing, property destruction, or violating rules/laws), *and Oppositional Defiant Disorder* (e.g., defiance, arguing with adults, or temper tantrums).

Given that several related but distinct dimensions exist within the overarching externalizing behavior construct, progress-monitoring assessment tools should be sensitive to changes in behavior within each distinct class of behavior. As such, extant Disruptive DBR (i.e., Chafouleas et al., 2010; Volpe & Briesch, 2012) would likely not capture the extent to which other related but distinct behaviors (e.g., opposition or aggression) indirectly affect student learning through teacher–student and peer relationships. Consequently, DBR assessment of an expanded range of externalizing constructs is needed because existing DBR may not be adequately sensitive to changes across the full range of behaviors relevant to student success.

Purpose of Study

Although initial evidence supports the dependability and treatment sensitivity of DBR-MIS focused on externalizing problem behaviors (e.g., Disruptive Behavior, Inattention/Overactivity, or Aggression), there are two limitations to this work to date. First, the construct validity of the Disruptive Behavior DBR-MIS, particularly the extent to which each item measures the underlying construct, has not been explored. Although Volpe and Briesch (2012) followed a process of content validation (i.e., asking a panel to evaluate the criterion relatedness, observability, and treatment validity of potential items) to develop the Disruptive Behavior DBR-MIS, no formal evaluation of construct validity was conducted for the resultant scale. Second, although the studies by Volpe et al. (2009), Volpe and Gadow (2010), Volpe, Briesch, and Gadow (2011), and Daniels et al. (2017) employed abbreviated versions of existing rating scales with established construct validity (i.e., IOWA Conners—Loney & Milich, 1982; Peer Conflict Scale—Gadow, 1986), both scales were designed to assess students with disruptive behavior disorders. As such, some of these items may be less socially valid (e.g., fidgeting) or amenable to change (e.g., excitable) within a typical classroom environment.

The current article aims to describe the process of item development used to create three DBR-MIS focused on externalizing problem behaviors, which can be used to efficiently evaluate student response to social-emotional or behavioral intervention. Within the first phase of development, item content was generated and subjected to content evaluation by panels of researchers and school-based staff. Exploratory factor analysis (EFA) was then used in the second phase to identify those items related to disruptive behavior, oppositional behavior, and interpersonal conflict that represented the strongest indicators of each particular construct of interest.

Method

Disruptive, Oppositional, and Interpersonal Conflict scales were chosen as the focus of this initial content validation study due to their relationship to academic performance and their importance in school-based intervention. All three constructs have the potential to interfere with learning and are viewed to be the highest priority targets for intervention among teachers (Bramlett, Murphy, Johnson, Wallingsford, & Hall, 2002; Briesch, Ferguson, Volpe, & Briesch, 2013).

Item Development

Item content for the Disruptive, Oppositional, and Interpersonal Conflict scales was initially developed through three stages. First, an initial pool of potential items was generated based on a national teacher survey of common referral concerns (Briesch et al., 2013), a review of existing measures (e.g., rating scales and observation codes) assessing the constructs of interest and a review of frequently used targets for Daily Report Card interventions (Owens et al., 2012).

Second, a Consumer Advisory Panel (CAP) consisting of four K–3 classroom teachers, one special education teacher of a substantially separate program for students with emotional and behavior disorders, one elementary school principal of an urban public elementary school, two elementary school psychologists with expertise in social-emotional and behavioral assessment, and four parents of children in kindergarten through third grade refined the pool of potential DBR-MIS items by evaluating them with regard to the following three criteria rated on a 5-point scale ranging from 0 = *strongly disagree* to 4 = *strongly agree*: (a) observability (*this item represents a behavior a teacher could see in a classroom, if it happened*), (b) suitability (*this item represents a behavior that would be a suitable target for intervention*), and (c) social validity (*this item represents a behavior that if changed would be helpful to the student and/or the classroom environment*). One classroom teacher was male, and the remaining 11 members of the CAP were female.

Third, a Scientific Advisory Panel (SAP) comprised of five researchers with expertise in scale development, the constructs of interest, and statistical methods reviewed and rated the items on the following four dimensions, each rating on the same 5-point scale (0 = *strongly disagree* to 4 = *strongly agree*): (a) representativeness (*this item is a strong indicator of the construct to which it was assigned*), (b) observability

(*this item represents a behavior that would be readily observable by classroom teachers*), (c) malleability (*this item represents a malleable behavior that could be targeted for treatment*), and (d) social validity (*this item represents a socially valid target for treatment*). Every SAP member held a doctoral degree in a field of psychology, which includes the specific fields of educational and school psychology. In addition, each SAP member had published more than 60 articles in peer-reviewed psychology and education journals at the time this article was published, and four of the five members have coauthored commercially published behavior rating scales.

Item retention decisions were subsequently made based on the feedback from the CAP and SAP and ratings were averaged across all members of each respective panel. Items with mean ratings less than or equal to 3.00 on a 0 to 4.00 scale on any of the three criteria rated by the CAP or the four criteria rated by the SAP were considered for elimination from a scale. This resulted in the deletion of four items designed to assess Interpersonal Conflict, four items designed to assess Oppositional Behavior, and two items designed to assess Disruptive Behavior. Items remaining in each scale after this initial phase were subsequently used in the EFA described in the following section.

EFA

Participants. Teacher recruitment followed a two-stage process. First, elementary school psychologists and principals identified through school district websites were contacted via email to serve as local coordinators for data collection. Second, local coordinators who expressed interest were asked to secure permission from administrators and subsequently recruit K–3 teachers in their school(s). The demographics for the total of 307 general and special education teachers in kindergarten through third grade who completed ratings are provided in Table 1. Each teacher completed ratings for one randomly selected student in his or her class (details regarding random student selection are provided under the "Procedures" section). The mean age of the student sample was 7.3 years (SD = 1.3), and the majority of students were male (60.9%), White (67.1%), and not receiving special education services (63.8%).

Measures. Following the multistage item development process described above, initial pools of items for each construct were included for rating by teachers. Specifically, the Disruptive scale included nine items, such as *disturbs others, calls out*, and *interrupts teacher*; the Oppositional scale included 11 items, such as *uncooperative, argues with teacher*, and *breaks rules*; and the Interpersonal Conflict scale contained 11 items, such as *insults peers, argues with other students*, and *difficulty resolving conflicts*. Teachers were instructed to rate the target student's behavior over the previous 5 school days using a 7-point scale ranging from *not a problem* to *serious problem* for all three scales.

Table 1. Teacher Demographics (N = 307).

Demographics	Frequency	Percentage of sample
Gender		
Male	13	4.2
Female	294	95.8
Years of experience		
0–5	73	23.8
6–10	61	19.9
11–20	107	34.9
21–30	48	15.6
>30	18	5.9
U.S. region (number of school districts)		
Northeast (25)	224	73.0
Midwest (5)	43	14.0
West/Southwest (3)	30	9.8
South (2)	10	3.3

Although Likert-type scales have been used in previous DBR-MIS research (e.g., Volpe & Briesch, 2012, 2015), respondents have typically been asked to rate the frequency with which a behavior was observed (i.e., did not occur or occurred always). One of the difficulties in using a frequency scaling method to assess externalizing problem behaviors, however, is that it does not account for the intensity of the behavior, which is an important dimension for disruptive, defiant, or aggressive behaviors. For example, a student might only physically aggress upon a peer on one occasion; however, this would still be considered to be a highly problematic behavior. By asking the teacher to rate the degree to which he or she believed that the behavior was a problem, it was believed that perceptions of both frequency and intensity would be incorporated into one rating.

Procedures. Local coordinators (e.g., school psychologist and principals) forwarded an email containing a link to an online Qualtrics rating form to teachers who agreed to participate. The first page of the online rating form included instructions for identifying the target student for rating, using an embedded random number generator. Specifically, the teacher was instructed to rate the student on his or her alphabetical class roster, who corresponded with a randomly generated number between 1 and 20 (if fewer than 20 students were in the class, the teacher continued counting from the top of the roster). This approach ensured that teachers did not select only students with significant behavioral concerns, which could potentially restrict score variance. Each teacher conducted ratings for only one student to yield independent observations suitable for EFA. In addition to item ratings, each teacher provided demographic information about the student and himself or herself in the online rating form. All information was automatically stored in Qualtrics and exported as an SPSS data file suitable for analysis.

Data analysis. Although EFA is traditionally used in scale development to identify latent constructs that explain item-level variance, EFA served a different purpose in the present study. Individual latent constructs were identified a priori and items were mapped on to these constructs following the aforementioned iterative multistage item-development process. Prior to EFA, interitem correlations within each scale were evaluated to identify the potential for multicollinearity. When interitem correlation coefficients exceeded .90, one item within the pair was selected for exclusion based on CAP and SAP feedback.

EFA was then performed for each scale to identify items that were most representative of the latent construct. Given that items were developed to measure a single construct, each DBR-MIS scale would ideally consist of items that loaded highly on a single factor. Principal Axis Factoring (PAF) was used to extract factors because it is more robust to violations of multivariate normality than other methods of extraction (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Parallel analysis (PA; Horn, 1965) was used to identify the optimal number of factors within each scale by computing eigenvalues from multiple sample correlation matrices generated from permutations of raw data. Factors with values exceeding the 95th percentile of eigenvalues were considered viable. Given that factors were expected to be correlated, direct quartimin rotation was applied in instances in which more than one factor was identified per scale (Costello & Osborne, 2005; Fabrigar & Wegener, 2012). Individual items were selected for continued inclusion in scales after EFA based on factor loadings and base rates.

Results

Item Development

CAP and SAP ratings on the aforementioned 0 to 4.00 scale were used to reduce the number of items in each of the three scales prior to EFA (see Table 2). Two items, *shows off* and *loud*, were eliminated from the Disruptive scale due to low CAP and SAP ratings, and the Disruptive scale consisted of nine items at the end of the initial item-development phase. Four items were removed from the Oppositional scale: *sassy, annoyed, pushes*, and *annoys*. After removal of the aforementioned items, the Oppositional scale consisted of 11 items. Finally, four items were removed from the Interpersonal Conflict scale: *provokes, excludes others, bullies*, and *hurts*. The Interpersonal Conflict scale consisted of 11 items after removal of the aforementioned items.

Interitem Correlations

Interitem Pearson correlation coefficients within each scale were reviewed to evaluate suitability for EFA (see Tables 3 to 5). Most coefficients were sufficiently high to conduct EFA (rs > .50), but not so high as to indicate the potential for multicollinearity (rs < .90), with the exception of three item pairs in the Oppositional scale: *talks back* and *disrespectful* (r = .90), *talks back* and *argues* (r = .91), and *directions* and *noncompliant* (r = .90). As a result, *talks back* and *directions* were removed prior to EFA because they were deemed to be redundant with other items in the scale.

EFA

No missing values were present in the data set as a result of the Qualtrics procedures, which required teachers to rate every item of every scale to submit their responses. Distributions of teacher ratings were skewed and kurtotic for some items, and a few multivariate outliers were identified for each scale. Data transformation was performed on all variables to reduce skewness, kurtosis, and the influence of multivariate outliers. The method of transformation used for each scale was based on the severity of the skewness and kurtosis in the data for each scale. That is, data were transformed using a progressively more intensive method recommended by Tabachnick and Fidell (2007), beginning with square root transformation and ending with inverse transformation, until skewness and kurtosis were reduced to acceptable values. Square root transformation was performed on all variables in the Disruptive scale because data were only moderately skewed and kurtotic, whereas inverse transformation was performed on variables in the Oppositional and Interpersonal Conflict scales because data were substantially skewed and kurtotic. Bartlett's Test of Sphericity was significant (p < .001) for all scales, which indicated that the data were suitable for factor analysis, and the following Kaiser–Meyer–Olkin Measure of Sampling Adequacy values were obtained: Disruptive = .94, Oppositional = .94, and Interpersonal Conflict = .95.

Disruptive

Results of parallel analysis (PA) indicated a one-factor solution for the Disruptive scale, with the single factor consisting of nine items accounting for 67.12% of the variance in ratings. All base rates fell between 17% and 26%, indicating that behaviors were exhibited by 17% to 26% of the sample (see Table 6). Finally, the overall mean factor loading was .82 (range = .75–.86), suggesting that all items were strong indicators of the underlying factor. Within the Disruptive scale, seven items were selected for retention, and two items (*active* and *disturbs*) were removed. *Active* was removed due to a low base rate (17.3%) and low factor loading (.80) relative to other items in the scale, as well as feedback from the SAP, which indicated a low rating on malleability (M = 2.80; range = 2.00–3.00). *Disturbs* was removed based on a low CAP rating on suitability (M = 2.82; range = 2.00–4.00). Coefficient alpha for the revised seven-item Disruptive scale was .93.

Table 2. Consumer and Scientific Advisory Panel Ratings.

	Consumer advisory panel mean ratings			Scientific advisory panel mean ratings			
Item description	Observability	Malleability	Socially valid	Construct	Observability	Malleability	Socially valid
Disruptive							
Active	3.3	2.8	3.0	3.0	3.8	2.8	3.2
Clowns	2.7	2.2	2.4	3.8	3.6	3.6	3.8
Noisy	3.1	2.7	3.0	3.8	3.8	3.8	3.8
Invades	3.3	3.2	3.2	4.0	3.4	3.4	3.4
Out of seat	3.9	3.1	3.3	3.4	4.0	4.0	3.6
Interrupts	3.4	2.8	3.0	4.0	4.0	4.0	4.0
Talks	3.5	3.3	3.5	3.8	3.4	4.0	3.8
Disturbs	3.1	2.8	3.1	4.0	3.4	4.0	4.0
Calls out	3.8	3.2	3.4	4.0	4.0	4.0	4.0
Shows off[a]	2.3	1.7	1.9	3.0	3.4	3.0	3.0
Loud[a]	3.0	2.6	2.9	3.4	4.0	3.6	3.6
Oppositional							
Blames	3.0	2.5	2.9	3.4	3.4	3.4	3.6
Frustrated	3.2	3.0	3.0	3.0	2.8	2.4	3.0
Temper	3.7	3.4	3.4	4.0	4.0	3.8	4.0
Disrespectful	3.5	3.3	3.4	4.0	4.0	3.8	3.8
Defiant	3.5	3.3	3.4	4.0	3.6	3.4	3.8
Uncooperative	3.3	2.9	3.2	4.0	3.8	3.6	3.8
Argues	3.5	3.0	3.0	3.8	4.0	3.8	3.4
Talks back	3.3	3.0	3.3	4.0	4.0	3.8	3.6
Rules	3.5	3.2	3.3	4.0	3.6	4.0	4.0
Noncompliant	3.3	2.9	3.1	3.8	3.8	4.0	4.0
Refuses							
Sassy[b]	1.6	1.3	1.5	3.2	3.4	3.0	3.0
Annoys[b]	3.0	2.2	2.4	3.4	3.8	3.4	3.2
Pushes[b]	2.7	2.0	2.2	3.0	3.4	3.4	3.2
Annoyed[b]	3.1	2.8	3.0	3.4	2.6	2.4	3.4
Interpersonal conflict							
Intimidates	2.8	3.3	3.3	4.0	3.2	2.8	3.6
Threatens	3.1	3.6	3.7	4.0	3.0	3.8	3.8
Annoys	2.8	2.7	2.9	3.6	2.8	3.0	3.4
Insults	2.9	2.9	3.3	3.8	3.2	3.6	3.6
Fights	3.1	3.3	3.6	3.8	3.6	4.0	4.0
Grabs	3.2	2.8	3.2	3.2	3.6	3.4	3.4
Argues	3.1	2.8	3.1	3.4	3.4	3.4	3.6
Throws	3.8	3.6	3.8	3.4	3.8	4.0	3.8
Teases	2.7	3.2	3.4	3.6	3.2	3.8	3.6
Hits	3.7	3.8	3.9	4.0	3.6	4.0	4.0
Bullies[c]	2.6	3.6	3.7	4.0	3.2	3.8	4.0
Excludes[c]	2.9	3.0	3.0	2.6	1.6	2.2	2.8
Provokes[c]	2.8	2.9	2.8	3.8	2.6	2.8	3.4
Hurts[c]	3.2	3.5	3.8	3.8	2.4	3.6	3.8
Conflicts[d]	—	—	—	3.2	2.6	3.4	3.4

[a]Items removed from the Disruptive scale based on Consumer Advisory Panel (CAP) and Scientific Advisory Panel (SAP) feedback.
[b]Items removed from the Oppositional scale based on CAP and SAP feedback.
[c]Items removed from the Interpersonal Conflict scale based on CAP and SAP feedback.
[d]*Conflicts* was not one of the original items reviewed by members of CAP. It was subsequently added based on feedback from the CAP and rated only by members of the SAP.

Table 3. Disruptive Interitem Correlations.

Item	1	2	3	4	5	6	7	8	9
1. Active	__								
2. Clowns	.67	__							
3. Noisy	.73	.73	__						
4. Invades	.67	.60	.66	__					
5. Out of seat	.75	.64	.76	.69	__				
6. Interrupts	.68	.64	.75	.64	.74	__			
7. Talks	.55	.67	.71	.61	.63	.70	__		
8. Disturbs	.68	.67	.78	.71	.77	.76	.75	__	
9. Calls out	.71	.68	.76	.66	.74	.85	.66	.72	__

Table 4. Oppositional Interitem Correlations.

Item	1	2	3	4	5	6	7	8	9	10	11
1. Blames	__										
2. Frustrated	.60	__									
3. Temper	.63	.78	__								
4. Disrespectful	.59	.64	.71	__							
5. Defiant	.60	.64	.72	.85	__						
6. Uncooperative	.59	.71	.71	.82	.86	__					
7. Argues	.62	.68	.75	.88	.84	.83	__				
8. Talks	.57	.61	.69	.90	.84	.82	.91	__			
9. Rules	.68	.63	.71	.77	.80	.78	.76	.73	__		
10. Noncompliant	.54	.70	.71	.82	.87	.87	.81	.80	.77	__	
11. Refuses	.53	.66	.71	.81	.88	.86	.80	.80	.76	.90	__

Table 5. Interpersonal Conflict Interitem Correlations.

Item	1	2	3	4	5	6	7	8	9	10	11
1. Intimidates	__										
2. Threatens	.70	__									
3. Annoys	.60	.58	__								
4. Insults	.71	.72	.69	__							
5. Fights	.70	.73	.62	.68	__						
6. Grabs	.55	.54	.61	.56	.61	__					
7. Argues	.63	.67	.70	.76	.77	.62	__				
8. Throws	.48	.52	.44	.49	.43	.47	.40	__			
9. Teases	.68	.63	.65	.79	.62	.50	.63	.37	__		
10. Hits	.61	.72	.60	.60	.68	.61	.59	.50	.61	__	
11. Conflicts	.59	.60	.67	.65	.68	.49	.75	.39	.55	.54	__

Oppositional

Results of PA indicated a one-factor solution for the Oppositional scale, and the single factor consisting of 11 items accounted for 69.27% of the variance in teacher ratings. Base rates in the sample fell between 15% and 26% for items in the Oppositional scale. Although all factor loadings were found to be strong ($M = .83$; range = .68–.90), two items (i.e., *blames* and *frustrated*) were removed due to relatively lower factor loadings than the other seven items (see Table 6). Feedback from the CAP and SAP supported the aforementioned item-retention and exclusion decisions. Specifically, *blames* was rated low on suitability ($M = 2.50$; range = 0–4.00) and social validity ($M = 2.90$; range = 0–4.00) by the CAP. *Frustrated* was rated low on observability ($M = 2.80$; range = 1.00–4.00) and malleability ($M = 2.4$; range = 1.00–4.00) by the SAP. Coefficient alpha for the revised seven-item Oppositional scale was .96 and all of the remaining interitem correlations were found to be strong (range = .71–88).

Table 6. Disruptive and Oppositional Scales: Item Means, Standard Deviations, Base Rates, and Factor Loadings.

Item	*M*	*SD*	Base rate %	Factor	Factor loading
Active	1.51	1.91	17.3	D	.80
Clowns[a]	1.91	2.01	17.9	D	.80
Noisy[a]	1.42	2.02	18.6	D	.86
Invades[a]	1.24	1.79	16.9	D	.77
Out of seat[a]	1.47	1.98	23.1	D	.83
Interrupts[a]	1.45	1.86	21.8	D	.84
Talks[a]	1.31	1.89	25.4	D	.75
Disturbs	1.20	1.83	26.1	D	.85
Calls out[a]	1.58	1.91	21.8	D	.86
Blames	1.51	1.91	18.6	O	.69
Frustrated	1.91	2.01	25.7	O	.68
Temper[a]	1.42	2.02	21.8	O	.82
Disrespectful[a]	1.24	1.79	15.0	O	.86
Defiant[a]	1.47	1.98	20.5	O	.90
Uncooperative[a]	1.45	1.86	19.9	O	.89
Argues[a]	1.31	1.89	18.2	O	.89
Rules[a]	1.58	1.91	23.1	O	.85
Noncompliant[a]	1.50	2.03	22.8	O	.88

Note. Parallel analysis indicated a one-factor solution for both the Disruptive and Oppositional scales. Square root transformation was performed on all Disruptive items; inverse transformation was performed on all Oppositional items.
[a]Item retained in the scale.

Interpersonal Conflict

Results of PA indicated a two-factor solution for the Interpersonal Conflict scale. The first factor consisted of five items accounting for 59.45% of the variance in ratings and the second factor, consisting of six items, accounted for an additional 4.07% of the variance. Item descriptives and factor loadings are reported in Table 7. After reviewing the content of each factor, we labeled the first factor *Conduct Problems* (which consisted largely of physically aggressive behaviors) and retained the label, *Interpersonal Conflict*, for the second factor. EFA was subsequently conducted on the new Conduct Problems scale (PA indicated a one-factor solution) consisting of five items loading on a single factor (see Table 7), which accounted for 61.30% of the variance.

Table 7. Interpersonal Conflict and Conduct Problems Item Means, Standard Deviations, Base Rates, and Factor Loadings.

Item	*M*	*SD*	Base rate %	Factor 1 loading	Factor 2 loading	Single IC factor loading	Single CD factor loading
Conduct problems							
Threatens	.67	1.30	6.5	.87			.87
Hits	.82	1.51	9.1	.83			.82
Intimidates	.90	1.55	10.1	.71			.81
Grabs	.94	1.42	7.8	.49	.29		.72
Throws	.50	1.20	5.2	.75			.68
Interpersonal conflict							
Insults	1.06	1.58	11.4	.42	.47	.84	
Argues	1.64	1.88	20.8		.88	.84	
Fights	.93	1.54	12.4	.46	.41	.79	
Annoys	1.74	1.93	21.5		.66	.77	
Teases	.82	1.40	7.8	.44	.36	.76	
Conflicts	2.01	2.01	27.0		.78	.74	

Note. Parallel analysis initially indicated a two-factor solution. Factor 1 Loading and Factor 2 Loading columns report loadings for the two-factor solution. Inverse transformation was performed on all variables. Factor loadings less than .20 were suppressed. The Single CD Factor Loading column reports loadings for a one-factor EFA conducted using only items in the newly formed Conduct Problems scale. The Single IC Factor Loading column reports loadings for a one-factor EFA conducted using only items retained in the Interpersonal Conflict scale.

The mean factor loading was .78 (range = .68–.87) and coefficient alpha for the five-item Conduct Problems scale was .86. EFA conducted on the six-item Interpersonal Conflict scale (PA now indicated a one-factor solution) indicated all items substantially loaded on a single factor (see Table 7), which accounted for 62.46% of the variance. The mean factor loading was .79 (range = .74–.84) and coefficient alpha for the six-item scale was .91.

Discussion

The purpose of the study was to develop DBR-MIS measuring externalizing behaviors, specifically disruptive, oppositional, and interpersonal problems. The study builds upon prior DBR-MIS research by (a) broadening assessment of externalizing problems beyond the most frequently researched constructs (e.g., Disruptive Behavior) to include assessment of interpersonal conflict with peers and adults and physically aggressive behaviors, and (b) evaluating the extent to which specific items comprising this broadened pool of DBR measures represent the constructs they purport to measure.

Initial item content was generated and evaluated by panels of school-based consumers and researchers with expertise in scale development and the constructs of interest. Items that were rated poorly by consumers and/or researchers with regard to construct representativeness, observability, suitability for school-based intervention, malleability, or social validity were removed prior to EFA. Input from consumers and researchers in the first stage of development increases the likelihood that resultant scales are perceived by consumers (e.g., teachers, school psychologists, and administrators) and experts to measure behaviors that are highly observable, functionally relevant to success in school, and responsive to treatment. Similarly, many items that represent more molar behaviors (e.g., *frustrated*) or clinical symptoms as opposed to functional targets were excluded based on empirical data, including relatively low base rates and factor loadings, as well as low ratings from consumers and experts. As such, the resultant scales are likely to align with socially valid targets for school-based intervention and be sensitive to changes in student behavior over time following intervention.

EFA was conducted in the second phase to identify items within each DBR-MIS that represented the strongest indicators of the constructs of interest. Results indicated one-factor solutions for the Disruptive and Oppositional scales and a two-factor solution for the original 11-item Interpersonal Conflict scale. Items with relatively low base rates and/or low factor loadings in comparison with other items in the same scale were removed, which resulted in seven-item Disruptive and Oppositional scales. Given that the 11 items within the original Interpersonal Conflict scale loaded on two factors, the items appeared to measure two distinct constructs. Consequently, five items with the highest loadings on the first factor were used to create a separate Conduct Problems scale, and the six items loading on the second factor were retained in the Interpersonal Conflict scale. The single factor measured by each of the four scales accounted for a substantial amount of variance in teacher ratings and each DBR-MIS demonstrated adequate internal consistency.

The fact that the initial pool of items in the Interpersonal Conflict scale, which were rated by consumers and experts to purportedly measure a single class of behavior, loaded on

two related but distinct constructs, highlights the limitations of using existing single-item scales (e.g., *Academically Engaged, Disruptive*, and *Respectful*) to capture the distinct features of different externalizing behaviors. That is, a scale measuring a single global construct, such as disruptive behavior, is unlikely to accurately measure behaviors associated with related, albeit distinct constructs such as aggression, which may be less overt. The need for multiple DBR to measure the full range of externalizing behaviors is also supported by the results of prior confirmatory factor analyses, which indicate that although hyperactive/impulsive, oppositional, and conduct problems are correlated, they are indeed separate constructs (Burns et al., 1997; Hartman et al., 2001). A few notable differences between results of the present study and prior factor analytic research are worthy of discussion. First, items similar to those in the Disruptive DBR-MIS primarily loaded on a Hyperactivity/Impulsivity factor in the Hartman et al. (2001) study; however, *out-of-seat behavior* had a secondary loading on an Attention Problems factor in a general youth sample (ages 5–13) rated by U.S. teachers. Second, items similar to those in the Oppositional DBR-MIS primarily loaded on an Oppositional Defiant Disorder factor in the Hartman et al. (2001) study; however, *rule-breaking behavior* also loaded on Hyperactive/Impulsive and Conduct Disorder factors in a sample of clinically referred youth (ages 3–19) rated by U.S. teachers. Third, fighting behaviors loaded on the Interpersonal Conflict DBR-MIS separate from Conduct Problems (e.g., *threatens; hits*) in the present study, whereas Hartman et al. (2001) found items measuring fighting had primary loadings on a Conduct Disorder factor (along with destruction and stealing) and secondary loadings on an Oppositional Defiant Disorder factor in samples of clinically referred (ages 3–19) and nonreferred (ages 5–13) youth rated by U.S. teachers. Although there is some overlap in constructs across DBR-MIS developed in the present study and measures used in prior factor analytic research, it is important to note that assessment tools used in the Hartman et al. (2001) study (e.g., Teacher Report Form; Achenbach, 1991; Child Symptom Inventory–4; Gadow & Sprafkin, 1997) were developed for diagnosis and classification, whereas DBR-MIS are intended for measuring changes in behavior in the short term (i.e., progress monitoring). Consequently, results of factor analyses would not be expected to align perfectly across studies.

Although these initial results are promising, it is important to note that EFA is relatively new to DBR research, given that the majority of research to date has focused on single-item scales. As a result, there is a limited frame of reference to which the construct validity of DBR-MIS developed in the present study may be compared. Nevertheless, the results may be interpreted in light of prior studies that have shown that three- to six-item scales constructed using items with the highest factor loadings on a full-length scale are sensitive to changes in students' externalizing behaviors following treatment (Daniels et al., 2017; Volpe & Gadow, 2010; Volpe et al., 2009). In addition, several studies have demonstrated the technical adequacy of Disruptive DBR-SIS (e.g., Chafouleas et al., 2010; Fabiano et al., 2017; Riley-Tillman et al., 2008) and initial research on DBR-MIS (which include somewhat different items than the scales developed in the present study) indicate that dependable measures of disruptive behavior may be obtained in eight to 12 assessment occasions (Volpe & Briesch, 2012, 2016). Given the results of these studies, it is likely that similarly strong dependability and treatment sensitivity will be found for the DBR-MIS developed in the present study; however, the psychometric properties of the newly formed DBR-MIS are yet to be fully investigated.

Limitations

Although results provide initial evidence to support the construct validity of the newly developed DBR-MIS, there are a few limitations to the present study that warrant discussion. First, although teachers included in the sample represented different geographical regions of the United States, students rated by the teachers were mostly White and/or male. Therefore, results may not generalize to populations that include higher percentages of minority or female students. Second, the majority of teachers who completed ratings were female (95.8%). Although this pattern is generally consistent with the current demographics of teachers across the United States (76.6%; U.S. Department of Education, National Center for Education Statistics, 2014), it is possible that the results primarily reflect the perspective of female teachers. Similarly, information regarding teachers' race/ethnicity and the primary population they were responsible for educating (e.g., students in general education classrooms or students in substantially separate special education classrooms) was not collected; therefore, conclusions regarding effects attributable to rater characteristics cannot be assessed. Third, although the overall sample size is sufficient for EFA, samples of constituent groups (e.g., gender and grade level) are too small to evaluate the extent to which DBR-MIS items measure latent constructs (i.e., disruptive behavior, oppositional behavior, interpersonal conflict, and conduct problems) consistently across groups, and thus measurement invariance remains unexplored in the present study (Finch & French, 2008).

Implications for Practice

Disruptive, Oppositional, Interpersonal Conflict, and Conduct Problems scales described in the present study were developed as part of a larger web-based system designed to assess a broad range of student behaviors (e.g., engagement, study skills, and social skills) using DBR-MIS. These brief scales,

comprised of five to seven items each, will primarily be used by K–3 classroom teachers to monitor changes in student behaviors associated with academic success in response to social, emotional, and behavioral interventions. That is, teachers and school-based problem-solving teams may select one or two scales for each student receiving targeted (Tier 2) or intensive (Tier 3) intervention that are aligned with the individual student's intervention goals. Teachers will subsequently use the scales to rate student behavior immediately following prespecified periods of time (e.g., at the end of a class period, school day, or week) to generate data streams, which may be used to inform intervention decisions within MTSS. Because student behaviors are rated in close temporal proximity to the actual occurrence of behaviors and within the context of interest (e.g., the classroom), inference is likely to be lower when compared with teacher ratings using traditional behavior rating scales. Finally, teachers will have the ability to complete ratings and view data on their web-enabled devices (e.g., smart phones, tablets, and laptops), which increases feasibility and facilitates timely decision-making.

Future Directions

The present study represents only one step in a larger scale development process. In the next phase, the consistency of the five- to seven-item DBR-MIS will be evaluated through a series of generalizability (G) and dependability (D) studies. In addition, the treatment sensitivity of the DBR-MIS will be examined by evaluating the extent to which the scales measure changes in individual students' behavior following implementation of evidence-based behavioral intervention (e.g., Daily Report Card, Volpe & Fabiano, 2013). Results of these studies will be considered along with the results of EFA conducted in the present study to determine which items will be included in the final Disruptive, Oppositional, Interpersonal Conflict, and Conduct Problems scales.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Institute of Education Sciences National Center for Special Education Research (R324A150071).

References

Achenbach, T. M. (1991). *Manual for the teacher's report form and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., & Jacobsen, S. J. (2007). Long-term school outcomes for children with attention deficit/hyperactivity disorder: A population-based perspective. *Journal of Developmental and Behavioral Pediatrics*, *28*, 265–273. doi:10.1097/DBP.0b013e31811ff87d

Bierman, K. L., Coie, J., Dodge, K., Greenberg, M., Lochman, J., McMohan, R., & Pinderhughes, E. (2013). School outcomes of aggressive-disruptive children: Prediction from kindergarten risk factors and impact of the Fast Track prevention program. *Aggressive Behavior*, *39*, 114–130.

Boyle, M. H., Offord, D. R., Racine, Y., Fleming, J. E., Szatmari, P., & Sanford, M. (1993). Evaluation of the revised Ontario child health study scales. *Journal of Child Psychology and Psychiatry*, *34*, 189–213.

Bradshaw, C. R. S., Mitchell, M., & Leaf, P. (2010). Examining the effects of schoolwide positive behavior interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*, 133–148. doi:10.1177/1098300709334798

Bramlett, R. K., Murphy, J. J., Johnson, J., Wallingsford, L., & Hall, J. D. (2002). Contemporary practices in school psychology: A national survey of roles and referral problems. *Psychology in the Schools*, *39*, 327–335.

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., Chafouleas, S. M., & Riley-Tillman, T. C. (2016). *Direct behavior rating (DBR): Linking assessment, communication, and intervention*. New York, NY: Guilford Press.

Briesch, A. M., Ferguson, T. D., Volpe, R. J., & Briesch, J. M. (2013). Examining teachers' perceptions of social-emotional and behavioral referral concerns. *Remedial and Special Education*, *34*, 249–256.

Briesch, A. M., & Volpe, R. J. (2007). Important considerations in the selection of progress monitoring measures for classroom behaviors. *School Psychology Forum*, *1*, 59–74.

Burns, G. L., Walsh, J. A., Owen, S. M., & Snell, J. (1997). Internal validity of attention deficit hyperactivity disorder, oppositional defiant disorder, and overt conduct disorder symptoms in young children—Implications from teacher ratings for a dimensional approach to symptom validity. *Journal of Clinical Child Psychology*, *26*, 266–275.

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., & 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. doi:10.1016/j.jsp.2010.02.001

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

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. doi:10.1177/1534508409340391

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

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

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*. Retrieved from http://pareonline.net/pdf/v10n7.pdf

Daniels, B., Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2017). Dependability and treatment sensitivity of multi-item Direct Behavior Rating scales for interpersonal peer conflict. *Assessment for Effective Intervention*, *43*, 48–59.

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2002). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review*, *31*, 298–312.

Fabiano, G. A., Pyle, K., Kelty, M. B., & Parham, B. R. (2017). Progress monitoring using direct behavior rating single item scales in a multiple-baseline design study of the daily report card intervention. *Assessment for Effective Intervention*, *43*, 21–33.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York, NY: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299. doi:10.1037/1082-989X.4.3.272

Finch, W. H., & French, B. F. (2008). Using exploratory factor analysis for locating invariant referents in factor invariance studies. *Journal of Modern Applied Statistical Methods*, *7*, 223–233.

Frick, P. J., Kamphaus, R. W., Lahey, B. B., Loeber, R., Christ, M. A. G., Hart, E. L., & Tannenbaum, L. E. (1991). Academic underachievement and the disruptive behavior disorders. *Journal of Consulting and Clinical Psychology*, *59*, 289–294.

Gadow, K. D. (1986). *Peer Conflict Scale*. Stony Brook: Department of Psychiatry, State University of New York.

Gadow, K. D., & Sprafkin, J. (1997). *Child Symptom Inventory 4: Norms manual*. Stony Brook, NY: Checkmate Plus.

Hartman, C. A., Hox, J., Mellenbergh, G. J., Boyle, M. H., Offord, D. R., Racine, Y., . . . Sergeant, J. A. (2001). DSM-IV internal construct validity: When a taxonomy meets data. *Journal of Child Psychology and Psychiatry*, *42*, 817–836.

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.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. doi:10.1007/BF02289447

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, waitlist controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. *Journal of Positive Behavior Interventions*, *11*, 133–144. doi:10.1177/1098300709332067

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.

Macleod, R. J., McNamee, M. A., Boyle, M. H., Offord, D. R., & Friedrich, M. (1999). Identification of childhood psychiatric disorder by informant: Comparisons of clinic and community samples. *Canadian Journal of Psychiatry*, *44*, 144–150.

Myers, D. M., Simonsen, B., & Sugai, G. (2011). Increasing teachers' use of praise with a response-to-intervention approach. *Education and Treatment of Children*, *34*, 35–59.

National Center on Intensive Intervention. (2018). *Behavioral progress monitoring tools*. Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Intensive Intervention. Retrieved from https://intensiveintervention.org/chart/behavioral-progress-monitoring-tools

Nelson, R. J., Benner, G. J., & Gonzalez, J. (2003). Learner characteristics that influence the treatment effectiveness of early literacy interventions: A meta-analytic review. *Learning Disabilities Research and Practice*, *18*, 255–267.

Owens, J. S., Holdaway, A. S., Zoromski, A. K., Evans, S. W., Himawan, L. K., Girio-Herrera, E., & Murphy, C. E. (2012). Incremental benefits of a daily report card intervention over time for youth with disruptive behavior. *Behavior Therapy*, *43*, 848–861.

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School Psychology Quarterly*, *26*, 1–13.

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A. M., & Glazer, A. D. (2008). Examining the agreement of direct behavior ratings and systematic direct observation for on-task and disruptive behavior. *Journal of Positive Behavior Interventions*, *10*, 136–143. doi:10.1177/1098300707312542

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, *17*, 223–237.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

U.S. Department of Education, National Center for Education Statistics. (2014) *Teacher attrition and mobility: Results from the 2012–13 teacher follow-up survey* (NCES 2014-077). Retrieved from https://nces.ed.gov/programs/digest/d17/tables/dt17_209.10.asp?current=yes

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., & Briesch, A. M. (2016). Dependability of two scaling approaches to Direct Behavior Rating Multi-Item Scales assessing disruptive classroom behavior. *School Psychology Review*, *35*, 39–52. doi:10.17105/SPR45-1.39-52

Volpe, R. J., Briesch, A. M., & 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. doi:10.1016/j.jsp.2010.09.005

Volpe, R. J., & Fabiano, G. A. (2013). *The Guilford practical intervention in the schools series. Daily behavior report cards: An evidence-based system of assessment and intervention*. Neywork: Guilford Press.

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. doi:10.1177/1063426608323370

Wehmeier, P. M., Schacht, A., & Barkley, R. A. (2010). Social and emotional impairment in children and adolescents with ADHD and the impact on quality of life. *Journal of Adolescent Health*, *46*, 209–217. doi:10.1016/j.jadohealth.2009.09.009