School Psychology

2578-4218/19/$12.00

2019, Vol. 34, No. 1, 86–95
http://dx.doi.org/10.1037/spq0000263

Developing a Direct Rating Behavior Scale for Depression in Middle School Students

Stephen P. Kilgus, Michael P. Van Wie, James S. Sinclair, T. Chris Riley-Tillman, and Keith C. Herman
University of Missouri

Research has supported the applied use of Direct Behavior Rating Single-Item Scale (DBR-SIS) targets of "academic engagement" and "disruptive behavior" for a range of purposes, including universal screening and progress monitoring. Though useful in evaluating social behavior and externalizing problems, these targets have limited utility in evaluating emotional behavior and internalizing problems. Thus, the primary purpose of this study was to support the initial development and validation of a novel DBR-SIS target of "unhappy," which was intended to tap into the specific construct of depression. A particular focus of this study was on the novel target's utility within universal screening. A secondary purpose was to further validate the aforementioned existing DBR-SIS targets. Within this study, 87 teachers rated 1,227 students across two measures (i.e., DBR-SIS and the Teacher Observation of Classroom Adaptation—Checklist [TOCA-C]) and time points (i.e., fall and spring). Correlational analyses supported the test–retest reliability of each DBR-SIS target, as well as its convergent and discriminant validity across concurrent and predictive comparisons. Receiver operating characteristic (ROC) curve analyses further supported (a) the overall diagnostic accuracy of each target (as indicated by the area under the curve [AUC] statistic), as well as (b) the selection of cut scores found to accurately differentiate at-risk and not at-risk students (as indicated by conditional probability statistics). A broader review of findings suggested that across the majority of analyses, the existing DBR-SIS targets outperformed the novel "unhappy" target.

Impact and Implications

Research suggests that although many students exhibit within internalizing concerns, schools struggle to identify them in a timely manner. The results of the study indicate how the Direct Behavior Rating Single-Item Scales (DBR-SIS) can be used to identify students exhibiting such concerns.

Keywords: universal screening, direct behavior rating, behavioral assessment

Mental health difficulties have an extraordinary impact on youth, adults, and society at large (Perou et al., 2013). Internalizing behavior problems, including depression and anxiety, represent particularly burdensome psychological conditions faced by many youths (Merikangas et al., 2010). Specifically, research indicates 10 to 20% of adolescents and 18.1 to 36.1% of adults display depressive symptoms (Kessler et al., 2009; Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993). Such comparability in rates between adolescents and adults suggests that depression in adults often begins earlier in life (Birmaher et al., 1996). Additional longitudinal research supports this conclusion, documenting that while externalizing behavior problems begin to decrease as youth enter early adolescence, internalizing behavior problems begin to increase (Cyranowski, Frank, Young, & Shear, 2000; Masten et al., 2005). Such early display of internalizing behavior problems is associated with several negative outcomes, including those (a) proximal in nature, such as low academic achievement and poor peer relationships (Fergusson & Woodward, 2002; Grover, Ginsburg, & Ialongo, 2007), as well as those (b) distal in nature, such as adult psychopathology, substance abuse, and suicidality (Fergusson & Woodward, 2002; Fombonne, Wostear, Cooper, Harrington, & Rutter, 2001; Perroud et al., 2009).

Despite the prevalence of internalizing problems among adolescents, as well as their long-term consequences, many youths with mental health concerns go without the supports they require to be successful (Forness, Freeman, Paparella, Kauffman, & Walker, 2012; Weist, 1999). This is likely in part because of the inade-

This article was published Online First June 18, 2018.

Stephen P. Kilgus, Michael P. Van Wie, James S. Sinclair, T. Chris Riley-Tillman, and Keith C. Herman, Department of Educational, School, and Counseling Psychology, University of Missouri.

T. Chris Riley-Tillman is an author of DBR Connect.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130143 to the University of Missouri (PI: Keith C. Herman). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

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

quacy of student identification methods commonly used in schools, such as teacher referral or office discipline referrals. Research has found these two methods to be associated with limited diagnostic accuracy, particularly for those students struggling with internalizing behavior (Nelson, Benner, Reid, Epstein, & Currin, 2002; Pearcy, Clopton, & Pope, 1993). Such limited detection of students with internalizing challenges is unfortunate, as many students might otherwise benefit from evidence-based interventions designed to prevent the progression of such problems (Forness, 2005).

Universal Screening

One means by which to support the early identification of youths exhibiting internalizing concerns is through universal screening, defined as a systematic process through which a population of individuals is evaluated to detect those possessing some condition of interest (Jenkins, Hudson, & Johnson, 2007). Universal screening is commonly used by professionals who have a legal or ethical responsibility to identify children who they believe would benefit from intervention or support. Within schools, particularly those structuring their service delivery via multitiered systems of support, universal screening can be used to identify students who are unresponsive to universal supports at Tier 1, thereby requiring more targeted or intensive supports at Tiers 2 or 3, respectively.

Many school-based screening tools have been developed over recent years, such as the *Student Internalizing Behavior Screener* (SIBS; Cook et al., 2011) and the *Student Risk Screening Scale-Internalizing and Externalizing* (SRSS-IE; Lane et al., 2012). Yet, despite the prevalence of internalizing problems, the importance of addressing such concerns, and repeated researcher calls for the development of internalizing measures, recent reviews of the literature suggest internalizing assessment methods are lacking both in number and quality (e.g., McIntosh, Ty, & Miller, 2014). This state of affairs within the internalizing area is in stark contrast to that within both academic and social behavior domains, within which numerous assessment methods have been developed (Kilgus, Reinke, & Jimerson, 2015).

Of the available academic and social behavioral screening tools methods, one of the most commonly researched and notable methodological categories corresponds to general outcome measures (GOMs). The purpose of a GOM is to support the collection of psychometrically defensible data related to variables that are not necessarily targeted for intervention, but are rather predictive of broad general functioning within a domain of interest (Fuchs & Deno, 1991). Within the area of reading, such a general indicator is found in oral reading fluency, which is used to predict student general reading proficiency. GOM tools possess multiple defining characteristics that make them particularly well suited for use within multitiered systems of support. First, GOMs are highly efficient, requiring minimal time and effort to be administered. Second, GOMs typically possess utility across multiple purposes of assessment, such as universal screening and progress monitoring (Hintze, Owen, Shapiro, & Daly, 2000). Third, when used for multiple purposes within a school setting, GOMs also enhance the broader efficiency of service delivery procedures. This is given that schools are required to dedicate resources and training to support the adoption of only a single measure across multiple tiers of support (Kilgus, Riley-Tillman, Chafouleas, Christ, & Welsh, 2014).

Within the academic domain, the most commonly used and researched GOM tool is found in curriculum-based measurement, a measure around which a great deal of psychometric evidence has been amassed across the past several decades (Kilgus, Methe, Maggin, & Tomasula, 2014; Reschly, Busch, Betts, Deno, & Long, 2009). Within the social behavior domain, such a tool is found in Direct Behavior Rating Single-Item Scales (DBR-SIS; Chafouleas, Riley-Tillman, & Christ, 2009). Unlike other traditional rating scales (e.g., SIBS, SRSS-IE), which are founded upon multiple items and accordingly less sensitive to short-term changes in behavior, DBR-SISs are founded upon only one item, which is designed to capture small changes of student behavior over small increments of time (e.g., hours or days). Research to date has supported DBR-SIS in both screening and progress monitoring, with findings corresponding to the tool's validity, reliability, diagnostic accuracy, and sensitivity to change (Briesch, Chafouleas, & Riley-Tillman, 2010; Chafouleas, Sanetti, Kilgus, & Maggin, 2012; Miller et al., 2015). Unfortunately, researchers have yet to develop and validate a GOM within the internalizing behavior domain. Yet, consideration of DBR-SIS methodology suggests it might be suited for use within this latter domain pending the development of novel DBR-SIS targets specific to internalizing behavior.

Direct Behavior Rating Single-Item Scales

DBR-SIS represents a hybrid assessment methodology, incorporating elements of two existing methodologies. First, DBR-SIS is like systematic direct observation in that DBR-SIS data are collected within a prespecified time and setting (e.g., large group reading instruction; 2:00–2:45 p.m.) relative to operationally defined behaviors. Second, DBR-SIS is akin to a behavior rating scale in that DBR-SIS data are collected via brief user ratings of student behavior. Specifically, after an informal observation of student behavior within the aforementioned context, DBR-SIS raters use a unipolar graphic rating scale to record their perceptions regarding the extent to which the student engaged in specific behaviors within the aforementioned prespecified time and setting.

The majority of research to date has supported the defensibility of multiple social behavior DBR-SIS targets, including disruptive behavior and academic engagement (Chafouleas, 2011). Each of these targets represents a "single-item scale," in that it serves as a sole indicator of a particular class of behavior. In using DBR-SIS to evaluate student disruptive behavior, a user would rate a single item on a repeated basis (e.g., across 5–10 occasions) to derive reliable information regarding the student's behavior. This stands in contrast to more traditional behavior rating scales, through which a rater would complete multiple items at a single time to acquire reliable information regarding the student's disruptive behavior.

Though the majority of DBR-SIS research to date pertains to social behavioral targets, a single study has also examined the tenability of DBR-SIS targets specific to internalizing problems. Specifically, von der Embse, Scott, and Kilgus (2015) examined the association between DBR-SIS academic anxiety targets and the Test Anxiety Inventory (TAI; Spielberger, 1980). Results supported the (a) concurrent validity of DBR-SIS targets, as well as

(b) their sensitivity to change in evaluating student response to a self-monitoring intervention. Unfortunately, this investigation is the only study to date that has examined internalizing-specific DBR targets. Furthermore, the measure considered by von der Embse et al. (2015) was specific to one particular narrow form of internalizing concerns (i.e., academic anxiety). Finally, the study yielded no information regarding the DBR-SIS' target defensibility in universal screening. Taken together, the limitations associated with the von der Embse et al. (2015) study, as well as the measure considered within that investigation, leaves room for additional research in establishing internalizing-specific DBR-SIS targets. More specifically, there is an apparent opportunity to develop and validate a DBR-SIS target that is specific to depression.

The benefits from extending the DBR-SIS framework to depression may be amplified when applied to adolescents in middle-school settings. Adolescence is an ideal developmental period for screening depression because the presence of depressive symptoms during adolescence increases the likelihood of developing full-syndrome mood disorders in adulthood (Klein, Shankman, Lewinsohn, & Seeley, 2009). In addition, the prevalence of some depressive disorders increases during the transition from adolescence to adulthood (Costello, Copeland, & Angold, 2011). Because adolescents attending middle school are typically under the allowable age for voluntary school withdrawal, screening efforts within middle-school contexts have the ability to reach many individuals within an opportune developmental timeframe. In addition, because many school systems provide mental health services to their students (Slade, 2002), linking screening and early intervention efforts can be easily facilitated within a school system.

Purpose of the Study

Taken together, a review of the literature reveals the availability of multiple targeted depression screeners (e.g., Children's Depression Inventory; Kovacs, 1992), defined as measures intended for use with students at elevated risk for depression (Levitt, Saka, Romanelli, & Hoagwood, 2007). In contrast, there are relatively fewer quality universal depression screeners, defined as measures used with all students regardless of risk status (Levitt et al., 2007). An understanding of DBR-SIS methodology suggests it might be an appropriate means by which to screen for such concerns pending the development of internalizing targets. The purpose of this study was to therefore develop and initially examine a depression-specific target, while considering its defensibility for use as a universal screener within a middle school sample. A secondary purpose was to further evaluate the existing core DBR-SIS targets of academic engagement and disruptive behavior. The current investigation followed an argument-based approach to validation (Kane, 2013), through which we examined evidence regarding the tenability of each target's (a) *interpretation* as an indicator of its corresponding area and (b) *use* as a screening tool for the purpose of differentiating at-risk and not at-risk students. Three research questions specific to the novel target's interpretation are as follows:

1. To what extent do DBR-SIS cores exhibit stability over time (i.e., test–retest reliability)?
2. Do DBR-SIS targets demonstrate concurrent and predictive validity relative to theoretically convergent measures of behavioral and emotional functioning, including scales derived from the *Teacher Observation of Classroom Adaptation, Checklist* (TOCA-C; Koth, Bradshaw, & Leaf, 2009)?
3. Do DBR-SIS targets demonstrate discriminant validity relative to theoretically divergent measures?

Two additional research questions specific to the use of DBR-SIS targets for universal screening purposes were as follows:

4. To what extent do DBR-SIS targets exhibit overall diagnostic accuracy, as measured via the area under the curve (AUC) statistic, relative to TOCA-C scales?
5. Which of the possible DBR-SIS cut scores is most suited for use in differentiating at-risk and not at-risk students (as defined via the TOCA-C)? Such suitability was evaluated via a series of conditional probability statistics, including sensitivity and specificity (among others).

Method

Participants

Middle school student and teacher participants were recruited from urban school districts in the Midwest. Participants were recruited as part of a larger Institute of Education Sciences (IES)-funded randomized controlled trial of a behavior management and coaching system. Eligible teacher participants included sixth to eighth Grade English language arts or math teachers who consented to participate in the aforementioned IES grant. All students within each consenting teacher's classrooms were invited to participate in the study. No other inclusionary or exclusionary criteria were specified for teacher and student participants. Before the investigation, none of the schools had a history of using DBR-SIS as part of their systematic service delivery efforts.

Across the broader school district, all English language arts and math teachers were invited to participate in the study. Approximately 73% of teachers were recruited, with the majority of those who elected to not participate citing concerns related to available time. Within the recruited classrooms, parental consent was received for 80% of students. Of this group of students, 100% assented to participation. To note, as part of assent procedures, students were alerted that their teachers would be rating their behavior as part of the study. Accordingly, such awareness may have influenced their behavior across the investigation.

A final teacher sample of 86 and student sample of 1,227 agreed to participate in the present study. Student participants were 50.0% female and 76.0% Black, 19.9% White, 2.2% Hispanic/Latino(a), and 1.1% Asian, and 0.9% other. The percentage of students in 6th, 7th, and 8th grade was equal to 41.8, 33.3, and 24.9%, respectively. Overall, 64.1% of students qualified for free/reduced-priced lunch. Teacher participants were 79.1% female and 70.9% White, 25.6% African American, 2.3% Asian, and 1.2% other. Teachers' ages ranged from 23 to 63 years ($M = 37.8$, $SD = 8.8$), whereas teaching experience ranged from 1.0 to 23.0 ($M = 10.4$, $SD = 6.3$).

Measures

Direct Behavior Rating Single Item Scales (DBR-SIS). The standard DBR-SIS form consists of three targets, including "academic engagement" (AE), "disruptive behavior" (DB), and "respectful behavior" (RB; Chafouleas et al., 2009). For the purposes of this study, only AE and DB were considered (RB was not considered because of the lack of relevant criterion scale within the TOCA-C). AE is defined as a student's active or passive participation in the classroom activity. Examples include writing, hand raising, answering a question, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials. DB is defined as student actions that interrupt regular or classroom activities. Examples included being out of one's seat, fidgeting, playing with objects, acting aggressively, or talking or yelling about things that are unrelated to classroom instruction.

Within the present study, teachers rated an additional DBR-SIS target corresponding to "unhappy" (UN; Rohrer & Herman, 2014). This particular item was intended to serve as a broad and general indicator of student depression. UN was defined as the expression of sadness, gloom, joylessness, or discontentment through words, body posture, tone of voice, facial expressions, or social cues. Examples included a limited range of facial expressions or animation, downward cast eyes and mouth, infrequent smiling or laughing, crying, inactivity, limited social participation, engagement in few pleasurable activities, recurrent expressions of worry or guilt, frequent physical complaints, pessimism, and negative self-statements.

DBR-SIS ratings corresponded to the percentage of time during an activity that the teacher observed the student to be engaging in each behavior. To clarify, teachers were not required to differentiate among any of the discrete behaviors comprising the broader DBR-SIS target definitions, but consider all behaviors simultaneously in estimating an overall percentage rating. During each phase of the study, teachers completed a single rating of each student across all three DBR-SIS targets. Each rating corresponded to behavior teachers observed of that particular student across the school day. The consideration of such lengthy rating periods is consistent with prior DBR-SIS research, including that related to the use of DBR-SIS for universal screening purposes (e.g., Kilgus, Riley-Tillman, et al., 2014).

Teacher Observation of Classroom Adaptation Checklist (TOCA-C). The TOCA-C (Koth et al., 2009) is a teacher-completed checklist of classroom behavior. Teachers rate the 24 TOCA-C items using a 6-point scale, with response options ranging from *never* to *almost always*. Three TOCA-C subscale scores were utilized in this study, including Concentration Problems, Disruptive Behavior, and Internalizing Problems. Previous research has supported the internal consistency of each subscale, with coefficient alphas ranging between .86 and .96 for all three subscales (Koth et al., 2009). The TOCA-C has also been found to predict various outcomes, including office discipline referrals (Pas, Bradshaw, & Mitchell, 2011) and alternate measurement methods such as the TOCA-R (Werthamer-Larsson, Kellam, & Wheeler, 1991). In the present study, internal consistency statistics ranged between .82 and .96 for TOCA-C subscales. Previous research has supported the present TOCA-C factor structure among students from the same geographic region as the present sample (Wang et al., 2015).

Procedures

Phase 1: Item development. Phase one of the current study centered on the creation of a single DBR-SIS target specific to depression. After the DBR-SIS framework, the final version of the measure would display the single word prompt in addition to a brief description of behaviors typical of the construct. After the selection of a satisfactory single-item prompt by an expert review, a behavioral definition was written to help raters comprehend the meaning of the construct. The single term unhappy was selected as the general prompt and behavioral descriptors of negative affectivity and low positive affect were used in the definition of the term. The inclusion of low positive affect terms was a purposeful decision to closely mirror the gold standard test for depression in the present study. Research has shown that low positive affect can differentiate depression from other internalizing problems (Watson, Clark, & Carey, 1988). Additionally, by emphasizing specific components of depression in comparison with other internalizing problems, the screening utility for the single-item was expected to increase.

Before administering the measure to study participants, a focus group of five teachers unaffiliated with the study was conducted to evaluate teachers' ability to use the measure as intended. Teachers were first provided the unhappy target prompt in written form. Upon reviewing the written prompt, teachers were asked a series of interview questions intended to assess both their understanding of the unhappy target and their capacity to use the target in evaluating student behavior. Overall, responses to interview questions suggested teachers understood the unhappy target and the broader depression construct being measured. The five teachers were each able to describe at least one student from their current roster that seemed to meet the profile of a student with substantial depression. The teachers also indicated that the student would receive a score of 7 or higher on a typical day. The teachers were able to discern the differences between scores appropriate for disruptive students with moderate DBR-SIS ratings and the most depressed students in the classroom. This suggests that the teachers were able to comprehend the construct as it was intended, and to measure depression independently from negative affect and low positive affect associated with externalizing problems.

Phase 2: Data collection. Following Phase 1 item development, large data collection efforts began with the entire sample. Data were collected in two rounds, first during the fall semester (late September to early October) followed by a second round in the spring semester approximately 6 months later (late April to May). Within each round of assessment, all DBR-SIS and TOCA-C ratings were completed within a 1-month time window. When completing the TOCA-C, teachers considered the student's behavior over the last 3 weeks. When completing DBR-SIS, teachers considered the student's behavior across that particular day (that researchers had preselected). All teacher-rated data were collected online using Qualtrics (2017) online survey solutions. Before data collection, a research assistant provided teachers with a brief in-person overview of each measure. When completing the measures via Qualtrics, teachers were provided specific instructions regarding how to complete each measure. For instance, when

completing DBR-SIS ratings, teachers were provided the operational definition of each behavior and reminded that their ratings should correspond to the percentage of time the student was observed engaging in each behavior.

Data Analysis Plan

Research question 1. The test–retest reliability of scores from each DBR-SIS target were evaluated via the calculation of a series of stability coefficients. Such coefficients represented Pearson correlations (*r*) between Time 1 and 2 scores within each target. No interpretive criteria were proposed for the evaluation of stability coefficients, as researchers have yet to suggest what threshold might correspond to "adequate" test–retest reliability. Furthermore, the extent to which data used for screening purposes should be stable over time is somewhat questionable. This is given that screening typically occurs in the context of prevention and intervention efforts, which might alter the rank ordering of students within a sample over time, thus, attenuating coefficients.

Research questions 2 and 3. Pearson correlation coefficients were once again calculated in evaluating the convergent and discriminant validity of each DBR-SIS target. Expected convergent relations were between (a) DBR-SIS AE and TOCA-C Attention/Concentration Problems, (b) DBR-SIS DB and TOCA-C Aggressive/Disruptive Behavior, and (c) DBR-SIS UN and TOCA-C Internalization. Expected discriminant relations were between all other DBR-SIS and TOCA-C pairings. In accordance with these expectations, it was anticipated that convergent coefficients would exceed those of discriminant relations within each DBR-SIS target. Convergent and discriminant validity was evaluated in both a concurrent and predictive fashion. Concurrent relations were evaluated within time point (e.g., Time 1 DBR-SIS and Time 1 TOCA-C), whereas predictive relations were evaluated across time points (e.g., Time 1 DBR-SIS and Time 2 TOCA-C). In accordance with interpretive benchmarks considered within prior DBR-SIS screening research, small, medium, and large coefficients corresponded to >.41, >.58, and >.69, respectively (Kilgus, Riley-Tillman, et al., 2014).

Research question 4. The overall diagnostic accuracy of each DBR-SIS target was evaluated via AUC statistics (and associated 95% confidence intervals, CIs), which were derived using receiver operating characteristic (ROC) curve analyses. AUCs are interpreted as the probability of a randomly selected at-risk individual yielding a more problematic score on a screener than a randomly selected not at-risk individual. AUC values range from 0 to 1, where 0 = perfectly incorrect decision making (where all risk decisions should be reversed), .50 = random decision making, and 1.00 = perfect decision making. In accordance with previous scholarly recommendations, AUC values >.50 were considered low, >.70 moderate, and >.90 high (Streiner & Cairney, 2007). Similar to correlational analyses, three sets of analyses were conducted, allowing for two evaluations of concurrent diagnostic accuracy (Time 1 DBR-SIS compared with Time 1 TOCA-C, as well as Time 2 DBR-SIS compared with Time 2 TOCA-C) and one evaluation of predictive diagnostic accuracy (Time 1 DBR-SIS compared with Time 2 TOCA-C). This approach to conducting analyses three times was repeated in the context of research Question 5 analyses described next.

Research question 5. The performance of individual DBR-SIS scores within screening decisions was evaluated via a series of conditional probability statistics. These included (a) sensitivity (SE), defined as the proportion of truly at-risk students (per the TOCA-C) identified as such via DBR-SIS, (b) specificity (SP), defined as the proportion of truly not at-risk individuals identified as such via DBR-SIS, (c) positive predictive values (PPV), or the proportion of individuals identified as at-risk via DBR-SIS who were actually at risk, and (d) negative predictive values (NPV), or the proportion of individuals identified as not at-risk via DBR-SIS who were actually not at risk.

In determining which cut score within each DBR-SIS scale was best suited for screening, we selected the cut score with the smallest difference between SE (true positive rate) and SP (true negative rate). We then evaluated the selected cut score relative to each of the four conditional probability statistics. In accordance with prior DBR-SIS research, it was expected selected cut scores would yield acceptable SE (>.80) and SP (>.70; Kilgus, Riley-Tillman, et al., 2014). It was further anticipated that NPV statistics would be high and approaching 1.00, whereas PPV statistics would be comparatively lower. This was given the inherent dependency of PPV and NPV statistics on the prevalence (i.e., base rate) of the condition in question (Petscher, Kim, & Foorman, 2011). In the context of low base rate conditions (e.g., social-emotional and behavioral risk), PPV is expected to be lower and NPV is expected to be higher. In the presence of a higher base rate condition, PPV will be higher and NPV will be lower.

Results

Research Question 1

See Table 1 for a summary of descriptive statistics specific to each DBR-SIS and TOCA-C scale. Analyses examined the extent to which Time 1 DBR-SIS scores were correlated with DBR-SIS scores collected at Time 2. Stability coefficients were found to equal .69 for DB, .70 for AE, and .48 for UN. All coefficients were statistically significant at the $p < .001$ level.

Research Question 2

See Table 2 for a summary of correlational findings specific to research Questions 2 (convergent validity) and 3 (discriminant validity). For each DBR-SIS, relative to all TOCA-C scales, the largest correlation was with the theoretically expected convergent TOCA-C scale. Thus, within each time point comparison, (a) DBR-SIS AE was most related to TOCA-C Concentration Problems, (b) DBR-SIS DB to TOCA-C Disruptive Behavior, and (c) DBR-SIS UN to TOCA-C Internalization. For DBR-SIS AE and DB, all convergent concurrent relations were large, whereas convergent predictive relations were medium. For DBR-SIS UN, convergent concurrent relations were medium, whereas convergent predictive relations were small.

Research Question 3

DBR-SIS UN exhibited the best discriminant validity of the DBR-SIS, with expected discriminant correlations all falling below the small threshold (>.41). Discriminant validity was more

Table 1
Descriptive Statistics for Direct Behavior Rating Single-Item Scales (DBR-SIS) and Teacher Observation of Child Adaptation, Checklist (TOCA-C) Scales Across Time Points

Scale	Time point	Scale	*M*	*SD*	Skewness	Kurtosis	Score at the 75th percentile
TOCA-C	1	Concentration problems	20.72	8.84	.02	−1.07	28
		Disruptive behavior	16.59	6.76	1.00	.75	21
		Internalization	8.85	3.63	1.18	2.03	11
	2	Concentration problems	20.20	9.10	.09	−1.12	27
		Disruptive behavior	17.47	7.17	.82	.23	22
		Internalization	9.56	3.75	.70	.45	12
DBR-SIS	1	Academic engagement	7.30	2.39	−1.00	.26	
		Disruptive behavior	2.43	2.76	1.12	.08	
		Unhappy	1.44	1.93	1.95	3.77	
	2	Academic engagement	7.39	2.42	−1.09	.51	
		Disruptive behavior	2.34	2.63	1.22	.48	
		Unhappy	1.57	1.85	2.01	4.75	

variable for DBR-SIS AE and DB. Although their relations with the TOCA-C Internalization scale were consistently below the small threshold, correlations between DBR-SIS DB and TOCA-C Concentration Problems, as well as DBR-SIS AE and TOCA-C Disruptive Behavior, fell in the small or medium range.

Research Question 4

See Table 3 for a summary of diagnostic accuracy findings specific to research Questions 4 (overall diagnostic accuracy). DBR-SIS AE yielded high AUC values within both concurrent analyses and a medium value within the predictive analysis. In contrast, DBR-SIS DB and UN yielded medium AUC values across all concurrent and predictive analyses. To note, 95% CIs also fell within their respective medium or high ranges. This was with the exception of DBR-SIS UN within the predictive analyses, where the lower end of the interval fell within the small range.

Research Question 5

See Table 3 for an overview of cut scores selected as best performing for each DBR-SIS within each ROC curve analysis (selected values are bolded). A review of conditional probability findings indicated DBR-SIS AE yielded cut scores associated with acceptable SE and SP across all three ROC curve analyses. Notably, across both concurrent analyses, the cut score selected as best performing was 6 (7 was selected within the predictive analysis). DBR-SIS DB performance was variable, with concurrent analyses yielding acceptable SE and SP values, but the predictive analysis yielding only an acceptable SP value (note that the cut score of 3 was selected across all analyses). DBR-SIS UN performance was also variable, with the scale yielding consistently acceptable SP but unacceptable SE. It should be noted, however, that SE values were found to approximate the SE acceptability threshold within concurrent analyses (i.e., .74 and .78). Across all three analyses, the cut score of 2 was selected as best performing.

Discussion

Previous research has supported the development and validation of multiple DBR-SIS targets. The majority of these targets have been relevant to student social behavior and externalizing concerns, with less focus given to targets specific to emotional behavior and internalizing concerns. Thus, researchers have recently begun to develop such internalizing targets via research similar to that conducted by early DBR-SIS researchers (e.g., von der Embse

Table 2
Pearson (r) Correlations Between Direct Behavior Rating Single-Item Scales (DBR-SIS) and Teacher Observation of Child Adaptation, Checklist (TOCA-C) Scales

		TOCA-C Scale		
Time point comparisons	DBR-SIS	Concentration problems	Disruptive behavior	Internalization
Time 1 DBR-SIS and Time 1 TOCA-C	AE	−.85**	−.55**	−.28**
	DB	.68**	.72**	.10**
	UN	.36**	.40**	.63**
Time 1 DBR-SIS and Time 2 TOCA-C	AE	−.67**	−.50**	−.25**
	DB	.53**	.62**	.07*
	UN	.27**	.34**	.45**
Time 2 DBR-SIS and Time 2 TOCA-C	AE	−.87**	−.54**	−.32**
	DB	.65**	.75**	.15**
	UN	.41**	.40**	.61**

Note. AE = academic engagement; DB = disruptive behavior; UN = unhappy.
* $p < .01$. ** $p < .001$.

Table 3
Diagnostic Accuracy Statistics Across Each Direct Behavior Rating Single-Item Scale (DBR-SIS) Relative to the Teacher Observation of Child Adaptation, Revised (TOCA-C) Scales

DBR-SIS	TOCA Scale	AUC [95% CI]	Cut	SE	SP	PPV	NPV
		Time 1 DBR-SIS → Time 1 TOCA-C					
AE	Concentration problems	.94 [.93, .96]	5	.78	.93	.75	.94
			6	**.89**	**.85**	**.61**	**.97**
			7	.96	.73	.48	.99
DB	Disruptive behavior	.88 [.86, .90]	2	.91	.66	.42	.97
			3	**.81**	**.78**	**.50**	**.94**
			4	.71	.86	.57	.92
UN	Internalization	.84 [.82, .87]	1	.97	.49	.33	.98
			2	**.74**	**.80**	**.49**	**.92**
			3	.52	.90	.56	.88
		Time 1 DBR-SIS → Time 2 TOCA-C					
AE	Concentration problems	.85 [.82, .88]	5	.55	.92	.68	.86
			6	.69	.85	.59	.89
			7	**.80**	**.73**	**.49**	**.92**
DB	Disruptive behavior	.82 [.79, .85]	1	.95	.41	.34	.96
			2	.80	.68	.44	.91
			3	**.68**	**.79**	**.50**	**.89**
UN	Internalization	.71 [.67, .75]	1	.82	.47	.28	.91
			2	**.54**	**.77**	**.37**	**.87**
			3	.38	.88	.45	.85
		Time 2 DBR-SIS → Time 2 TOCA-C					
AE	Concentration problems	.95 [.94, .96]	5	.69	.96	.84	.90
			6	**.85**	**.91**	**.77**	**.95**
			7	.96	.76	.57	.98
DB	Disruptive behavior	.89 [.87, .91]	2	.94	.68	.48	.97
			3	**.82**	**.81**	**.58**	**.94**
			4	.70	.89	.68	.90
UN	Internalization	.83 [.80, .85]	1	.98	.37	.28	.98
			2	**.78**	**.75**	**.45**	**.93**
			3	.56	.88	.54	.89

Note. AUC = area under the curve; CI = confidence interval; AE = academic engagement; DB = disruptive behavior; UN = unhappy; SE = sensitivity; SP = specificity; PPV = positive predictive value; NPV = negative predictive value. Bolded values correspond to cut scores selected as best performing within each analysis.

et al., 2015). The primary purpose of this study was to build upon this research base by initially validating a depression-specific target of unhappy. A secondary purpose was to expand the broader DBR-SIS research base by also considering the core social behavior DBR-SIS targets of disruptive behavior and academic engagement.

Findings yielded support for the *interpretation* of DBR-SIS scores as indicators of their respective areas. First, findings spoke to the test–retest reliability of each DBR-SIS target. Score stability was strongest for disruptive behavior and academic engagement, with unhappy proving less stable over time. With that said, though less stable, similar levels of stability have been documented in regards to alternative measures of depression, including the Child Depression Inventory (Nelson & Politano, 1990). As noted above, it is difficult to evaluate the extent to which stability in DBR-SIS scores across times should be anticipated or valued. Screening often takes place in the context of intervention efforts, which can alter student behavior and, thus, the rank ordering of students in terms of their behavior over time. Such changes in rank ordering will by definition attenuate correlations, suggesting that restricted correlations would be an intervention artifact and not measurement error. In the absence of information related to student intervention status within this study, future research remains necessary to clarify this issue.

Second, DBR-SIS score interpretation was also supported by evidence of convergent and discriminant validity. Each DBR-SIS target was found to be most strongly associated with its theoretically aligned TOCA-C scale (relative to alternative scales). That notable and statistically significant correlations were still noted for discriminant relations is to be expected, as research suggests even disparate areas of behavior are likely to be influenced by broader patterns of behavioral functioning (DiStefano, Greer, & Kamphaus, 2013; von der Embse, Pendergast, Kilgus, & Eklund, 2016). Nevertheless, that such strong associations were noted for theoretically aligned scales is encouraging and indicative of the potential for DBR-SIS targets to tap into their respective areas.

A review of correlational patterns indicated concurrent convergent relations (among data collected at the same time) exceeded predictive convergent relations (among data collected at different time points). This is to be expected given the logical assumption that measurements taken closer in time will be more related than

measurements separated in time. Yet, the robustness of this finding and interpretation was supported by the replication of concurrent relations within each time point, where the pattern of correlation findings and correlational magnitudes were found to be markedly similar across the two analyses.

Finally, the current findings also supported the *use* of DBR-SIS scores for universal screening purposes. AUC values spoke to the overall moderate to high diagnostic accuracy of each target in predicting its corresponding and theoretically convergent scale on the TOCA-C. ROC curve analyses also resulted in the selection of cut scores that could be used within screening for differentiating between at-risk and not at-risk students. The same cut score was selected for DBR-SIS DB and UN (i.e., 3 and 2, respectively) across each of the ROC curve analyses. The same DBR-SIS AE cut score was selected across the concurrent analyses (i.e., 6), whereas a slightly different score was selected within the predictive analysis (i.e., 7). Such consistency in cut score selection speaks to the robustness of cut score performance, thereby enhancing the confidence one might have in using the cut scores for applied decision making. Further confidence is gained by review of cut score performance. DBR-SIS DB and AE cut scores were found to consistently yield acceptable SE and SP. In contrast, where DBR-SIS UN exhibited acceptable SP, SE levels were found to consistently approximate but not meet the acceptability threshold.

As expected, NPV statistics were high while PPV statistics were low. These findings provide information regarding the potential applied implications of using these DBR-SIS targets for universal screening purposes. Specifically, NPV findings indicate that although the majority of students identified as not at risk would indeed be not at risk, on average, PPV findings suggest approximately only 50% of students identified as at risk would actually be at risk. Though such PPV findings were once again expected, they nevertheless speak to the potential for inaccurate decision making and, thus, inappropriate resource expenditure in providing students support they might not need.

Limitations and Future Research

Certain limitations to the investigation should be noted. First, the generalizability of the current findings is inherently limited given data were collected within a single Midwestern state and school year. Though this study builds upon previous multisite and longitudinal studies (e.g., Kilgus, Riley-Tillman, et al., 2014), future research inclusive of more diverse student samples and time points remains necessary. Second, the current findings are subject to mono-method and mono-informant biases, as both predictor and criterion scores represented teacher ratings. Such biases are likely to inflate correspondences between DBR-SIS and TOCA-C and, thus, estimates of DBR-SIS validity and diagnostic accuracy. Accordingly, readers are encouraged to interpret the findings with caution. Moving forward, it would be of interest to examine whether teacher DBR-SIS ratings predict criterion student self-report measures or a more comprehensive assessment inclusive of multiple methods (e.g., diagnostic interviews, parent and teacher rating scales).

Third, though not necessarily a limitation, it should be noted DBR-SIS data correspond to a single teacher rating within each time point. Prior DBR-SIS diagnostic accuracy studies have involved the collection of multiple data points (e.g., 5–10) within a time period (e.g., Kilgus et al., 2014). These values were then aggregated into a single summary statistic, with such collapsing supported by findings indicative of the time-series reliability of scores. It is suggested that future research compare the single and multiple rating approaches to determine whether they are associated with differential reliability, validity, or diagnostic accuracy. Fourth, though previous research has resulted in the development of efficacious protocols by which to train DBR-SIS users (Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012), such protocols were not used in this study. Future research should therefore employ these protocols while examining the extent to which they influence DBR-SIS diagnostic accuracy.

In the context of the broader literature, the current findings (a) further support existing evidence for reliability, validity, and diagnostic accuracy of the DB and AE targets, as well as (b) provide initial support for the novel UN target. Moving forward, there is a need for additional research regarding each of the targets under consideration. Researchers should further examine DBR-SIS test–retest reliability, while also considering additional forms (e.g., interrater). In addition, validity and diagnostic accuracy should be considered in relation to alternative outcomes. Such an outcome might include systematic direct observation, which has been used in prior DBR-SIS validity research (e.g., Chafouleas, Sanetti, et al., 2012; Smith, Eklund, & Kilgus, in press). Though the current study suggests the novel UN target is associated with more global ratings indicative of a student's behavior over 3 weeks (as indicated by the TOCA-C), the use of direct observation would indicate whether DBR-SIS estimates of behavior within a particular time and setting are associated with direct data collected within that same context. The examination of such association is particularly necessary in evaluating whether the UN target will be sensitive to subtle changes in behavior over time, as is necessary of a progress monitor. Furthermore, through direct observation, researchers could examine whether the DBR-SIS UN target is associated with a narrower criterion measure specifically representative of unhappy behavior. This would be accomplished by using direct observation to assess unhappy behavior using the same operational definition that teachers consider when completing DBR-SIS ratings. Finally, future research might also examine whether variations in the UN definition and corresponding examples (e.g., via the inclusion of additional, more observable behavioral examples), would influence UN target performance.

In addition, future research should both re-examine the performance of the DBR-SIS targets at the middle school levels, as well as at the high school and elementary levels. The latter is considered particularly relevant in the interest of determining how DBR-SIS contributes to the primary prevention of internalizing concerns. Finally, researchers should compare the DBR-SIS to other internalizing screeners in terms of validity, diagnostic accuracy, and feasibility. Such evidence would ultimately speak to the incremental value of DBR-SIS in the realm of internalizing screening.

References

Birmaher, B., Ryan, N. D., Williamson, D. E., Brent, D. A., Kaufman, J., Dahl, R. E., . . . Nelson, B. (1996). Childhood and adolescent depression:

A review of the past 10 years. Part I. *Journal of the American Academy of Child & Adolescent Psychiatry, 35,* 1427–1439. http://dx.doi.org/10.1097/00004583-199611000-00011

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.

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

Chafouleas, S. M., 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. http://dx.doi.org/10.1016/j.jsp.2011.11.007

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. http://dx.doi.org/10.1177/1534508409340391

Chafouleas, S. M., Sanetti, L. M. H., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change using Direct Behavior Rating Single-Item Scales. *Exceptional Children, 78,* 491–505. http://dx.doi.org/10.1177/001440291207800406

Cook, C. R., Rasetshwane, K., Sprague, J., Collins, T., Dart, E., Grant, S., & Truelson, E. (2011). Development and Validation of the Student Internalizing Behavior Screener: Examination of Reliability, Validity, and Classification Accuracy. *Assessment for Effective Intervention, 36,* 71–79. http://dx.doi.org/10.1177/1534508410390486

Costello, E. J., Copeland, W., & Angold, A. (2011). Trends in psychopathology across the adolescent years: What changes when children become adolescents, and when adolescents become adults? *Journal of Child Psychology and Psychiatry, 52,* 1015–1025. http://dx.doi.org/10.1111/j.1469-7610.2011.02446.x

Cyranowski, J. M., Frank, E., Young, E., & Shear, M. K. (2000). Adolescent onset of the gender difference in lifetime rates of major depression: A theoretical model. *Archives of General Psychiatry, 57,* 21–27. http://dx.doi.org/10.1001/archpsyc.57.1.21

DiStefano, C., Greer, F. W., & Kamphaus, R. W. (2013). Multifactor modeling of emotional and behavioral risk of preschool-age children. *Psychological Assessment, 25,* 467–476. http://dx.doi.org/10.1037/a0031393

Fergusson, D. M., & Woodward, L. J. (2002). Mental health, educational, and social role outcomes of adolescents with depression. *Archives of General Psychiatry, 59,* 225–231. http://dx.doi.org/10.1001/archpsyc.59.3.225

Fombonne, E., Wostear, G., Cooper, V., Harrington, R., & Rutter, M. (2001). The Maudsley long-term follow-up of child and adolescent depression. I. Psychiatric outcomes in adulthood. *The British Journal of Psychiatry, 179,* 210–217. http://dx.doi.org/10.1192/bjp.179.3.210

Forness, S. R. (2005). The pursuit of evidence-based practice in special education for children with emotional or behavioral disorders. *Behavioral Disorders, 30,* 311–330. http://dx.doi.org/10.1177/019874290503000406

Forness, S. R., Freeman, S. F. N., Paparella, T., Kauffman, J. M., & Walker, H. M. (2012). Special education implications of point and cumulative prevalence for children with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 20,* 4–18. http://dx.doi.org/10.1177/1063426611401624

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57,* 488–500. http://dx.doi.org/10.1177/001440299105700603

Grover, R. L., Ginsburg, G. S., & Ialongo, N. (2007). Psychosocial outcomes of anxious first graders: A seven-year follow-up. *Depression and Anxiety, 24,* 410–420. http://dx.doi.org/10.1002/da.20241

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J., III. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15,* 52–68. http://dx.doi.org/10.1037/h0088778

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36,* 582–600.

Kane, M. K. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73. http://dx.doi.org/10.1111/jedm.12000

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., Ormel, J., . . . Wang, P. S. (2009). The global burden of mental disorders: An update from the WHO World Mental Health (WMH) surveys. *Epidemiologia e Psichiatria Sociale, 18,* 23–33. http://dx.doi.org/10.1017/S1121189X00001421

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52,* 377–405. http://dx.doi.org/10.1016/j.jsp.2014.06.002

Kilgus, S. P., Reinke, W. M., & Jimerson, S. R. (2015). Understanding mental health intervention and assessment within a multi-tiered framework: Contemporary science, practice, and policy. *School Psychology Quarterly, 30,* 159–165. http://dx.doi.org/10.1037/spq0000118

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

Klein, D. N., Shankman, S. A., Lewinsohn, P. M., & Seeley, J. R. (2009). Subthreshold depressive disorder in adolescents: Predictors of escalation to full-syndrome depressive disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 48,* 703–710. http://dx.doi.org/10.1097/CHI.0b013e3181a56606

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation—Checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development, 42,* 15–30. http://dx.doi.org/10.1177/0748175609333560

Kovacs, M. (1992). *Manual for Children's Depression Inventory*. North Tonawanda, NY: Multi-Health Systems, Inc.

Lane, K. L., Oakes, W. P., Harris, P. J., Menzies, H. M., Cox, M., & Lambert, W. (2012). Initial evidence for the reliability and validity of the Student Risk Screening Scale for internalizing and externalizing behaviors at the elementary level. *Behavioral Disorders, 37,* 99–122. http://dx.doi.org/10.1177/019874291203700204

Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology, 45,* 163–191. http://dx.doi.org/10.1016/j.jsp.2006.11.005

Lewinsohn, P. M., Hops, H., Roberts, R. E., Seeley, J. R., & Andrews, J. A. (1993). Adolescent psychopathology: I. Prevalence and incidence of depression and other *DSM–III–R* disorders in high school students. *Journal of Abnormal Psychology, 102,* 133–144. http://dx.doi.org/10.1037/0021-843X.102.1.133

Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradović, J., Riley, J. R., . . . Tellegen, A. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology, 41,* 733–746. http://dx.doi.org/10.1037/0012-1649.41.5.733

McIntosh, K., Ty, S. V., & Miller, L. D. (2014). Effects of school-wide positive behavioral interventions and supports on internalizing problems current evidence and future directions. *Journal of Positive Behavior*

Interventions, 16, 209–218. http://dx.doi.org/10.1177/1098300713491980

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Survey Replication—Adolescent Suppl. (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry, 49,* 980–989. http://dx.doi.org/10.1016/j.jaac.2010.05.017

Miller, F. G., Cohen, D., Chafouleas, S. M., Riley-Tillman, T. C., Welsh, M. E., & Fabiano, G. A. (2015). A comparison of measures to screen for social, emotional, and behavioral risk. *School Psychology Quarterly, 30,* 184–196. http://dx.doi.org/10.1037/spq0000085

Nelson, R. J., Benner, G. J., Reid, R. C., Epstein, M. H., & Currin, D. (2002). The convergent validity of office discipline referrals with the CBCL-TRF. *Journal of Emotional and Behavioral Disorders, 10,* 181–188. http://dx.doi.org/10.1177/10634266020100030601

Nelson, W. M., & Politano, P. M. (1990). Children's depression inventory: Stability over repeated administration in psychiatric inpatient children. *Journal of Clinical Child Psychology, 19,* 254–256. http://dx.doi.org/10.1207/s15374424jccp1903_8

Pas, E. T., Bradshaw, C. P., & Mitchell, M. M. (2011). Examining the validity of office discipline referrals as an indicator of student behavior problems. *Psychology in the Schools, 48,* 541–555. http://dx.doi.org/10.1002/pits.20577

Pearcy, M. T., Clopton, J. R., & Pope, A. W. (1993). Influences on teacher referral of children to mental health services: Gender, severity, and internalizing versus externalizing problems. *Journal of Emotional and Behavioral Disorders, 1,* 165–169. http://dx.doi.org/10.1177/106342669300100304

Perou, R., Bitsko, R. H., Blumberg, S. J., Pastor, P., Ghandour, R. M., Gfroerer, J. C., . . . the Centers for Disease Control and Prevention (CDC). (2013). Mental health surveillance among children—United States, 2005–2011. *Morbidity and Mortality Weekly Report, 62,* 1–35.

Perroud, N., Aitchison, K. J., Uher, R., Smith, R., Huezo-Diaz, P., Marusic, A., . . . Craig, I. (2009). Genetic predictors of increase in suicidal ideation during antidepressant treatment in the GENDEP project. *Neuropsychopharmacology, 34,* 2517–2528. http://dx.doi.org/10.1038/npp.2009.81

Petscher, Y., Kim, Y. S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention, 36,* 158–166. http://dx.doi.org/10.1177/1534508410396698

Qualtrics. (2017). [Computer software]. Retrieved from http://www.qualtrics.com

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47,* 427–469. http://dx.doi.org/10.1016/j.jsp.2009.07.001

Rohrer, D., & Herman, K. C. (2014). *The Direct Behavior Rating Single-Item Scale (DBR-SIS) for internalizing behavior*. [Unpublished scale.] University of Missouri.

Slade, E. P. (2002). Effects of school-based mental health programs on mental health service use by adolescents at school and in the community. *Mental Health Services Research, 4,* 151–166. http://dx.doi.org/10.1023/A:1019711113312

Smith, R. L., Eklund, K., & Kilgus, S. P. (2018). Concurrent validity and sensitivity to change of Direct Behavior Rating Single-Item Scales (DBR-SIS) within an elementary sample. *School Psychology Quarterly, 33,* 83–93. http://dx.doi.org/10.1037/spq0000209

Spielberger, C. D. (1980). *Preliminary professional manual for the Test Anxiety Inventory (TAI*). Palo Alto, CA: Consulting Psychologists Press.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *The Canadian Journal of Psychiatry, 52,* 121–128. http://dx.doi.org/10.1177/070674370705200210

von der Embse, N. P., Pendergast, L. L., Kilgus, S. P., & Eklund, K. R. (2016). Evaluating the applied use of a mental health screener: Structural validity of the Social, Academic, and Emotional Behavior Risk Screener. *Psychological Assessment, 28,* 1265–1275. http://dx.doi.org/10.1037/pas0000253

von der Embse, N. P., Scott, E. C., & Kilgus, S. P. (2015). Sensitivity to change and concurrent validity of direct behavior ratings for academic anxiety. *School Psychology Quarterly, 30,* 244–259. http://dx.doi.org/10.1037/spq0000083

Wang, Z., Rohrer, D., Chuang, C., Fujuki, M., Reinke, W. M., & Herman, K. C. (2015). Five methods to score Teacher Observation of Classroom Adaptation Checklist and to examine group differences. *Journal of Experimental Education, 83,* 24–50. http://dx.doi.org/10.1080/00220973.2013.876230

Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology, 97,* 346–353. http://dx.doi.org/10.1037/0021-843X.97.3.346

Weist, M. D. (1999). Challenges and opportunities in expanded school mental health. *Clinical Psychology Review, 19,* 131–135. http://dx.doi.org/10.1016/S0272-7358(98)00068-3

Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology, 19,* 585–602. http://dx.doi.org/10.1007/BF00937993

Received September 7, 2017
Revision received March 17, 2018
Accepted March 20, 2018 ■