







Convergent validity of the Behavior Observation of Students in Schools for elementary school students with disruptive behaviors

Alexander Alperin  | Christopher M. Dudek  |
Linda A. Reddy  | Todd A. Glover  | Nicole B. Wiggs  |
Briana Bronstein 

Rutgers, The State University of New Jersey,
Piscataway, New Jersey, USA

Correspondence

Christopher M. Dudek, Graduate School of
Applied and Professional Psychology, Rutgers
University, 152 Frelinghuysen Rd,
Piscataway, NJ 08854, USA.

Email: cdudek@gsapp.rutgers.edu

Funding information

Institute of Education Sciences

Abstract

Systematic direct observations (SDOs) and behavior rating scales are integral to multimethod behavioral assessment approaches. The present study investigated the convergent validity of the Behavior Observation of Students in Schools (BOSS) form, a widely utilized SDO system, with the most frequently used teacher rating scales in schools, the Behavioral Assessment System for Children, Third Edition (BASC-3) and the Behavioral and Emotional Screening System, third edition (BESS-3). Specifically, the present study compared independent observer data collection of student on-task and disruptive behaviors on the BOSS with teachers' ratings of behavioral problems on the BASC-3 and BESS-3. Data come from a sample of 136 teachers and 349 students with or at risk of disruptive behavior disorders (DBDs). Pearson correlations were computed between BOSS and teacher ratings on the BASC-3 and BESS-3. Findings demonstrated significant, small to moderate correlations. Results indicate evidence of convergent validity for the BOSS and BASC3 assessments and highlight

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Psychology in the Schools* Published by Wiley Periodicals LLC.

multimethod behavioral assessment's utility in supporting students with or at risk of DBDs.

KEYWORDS

behavior assessment, BOSS, convergent validity, disruptive behavior, systematic direct observation (SDO)

1 | INTRODUCTION

Students with or at risk of disruptive behavior disorders (DBDs) pose complex educational and mental health concerns in schools. Displaying behaviors such as aggression, noncompliance (NC), and inattention, these students often receive diagnoses of attention deficit hyperactivity disorder, oppositional defiant disorder, and conduct disorder (Muratori et al., 2017). Significant school resources are allocated to address students with or at-risk for DBDs and these students represent the most common referrals to school-based teams, receiving approximately 30% of special education services nationwide (Allen, 2015; Pikard et al., 2018). Students with DBDs can negatively impact the learning and classroom environment for all students. Teachers, on average, report losing a significant amount of instructional time each week to address disruptive classroom behaviors (Education Advisory Board, 2019; Walker et al., 2004). Managing students' maladaptive behaviors can lead to teacher burnout and stress, and has demonstrated negative correlations with teacher retention and physical and emotional well-being (e.g., Armbruster et al., 2004; Bettini et al., 2019). The complex behavioral, emotional, and academic needs of these students pose significant risks for school and life success; thus receiving research-based interventions early in school settings are critical.

The selection and implementation of interventions that best meet the needs of students with DBDs are predicated on the systematic use of valid behavioral assessment to identify risk and areas in need for interventions. A common method for collecting behavior assessment data in schools is behavior-rating scales (Handler & DuPaul, 2005; Shapiro & Heick, 2004). Behavior rating scales offer school personnel several advantages, including (a) the quantification of behavior supported by reliability and validity data, (b) assessment of broad or narrow range of behaviors, (c) ease of administration and scoring, (d) multiple informant perspectives, and (e) may be compared with normative data samples for comparison (Gresham & Elliot, 2008; Hosp et al., 2003; McConaughy & Ritter, 2002). Abbreviated versions of rating scales offer utility for screening and ongoing progress monitoring (Gresham & Elliot, 2008; Kamphaus & Reynolds, 2007) and when combined with comprehensive rating scale systems, they can inform multiple steps throughout special education processes. Given these advantages, evidence from behavior rating scales is often used to make behavioral diagnoses, inform eligibility special programs criteria, plan for behavioral intervention and supports, and monitor intervention outcomes in schools.

In the context of disruptive and externalizing behaviors at school, classroom teachers are an important source of information on students' behavior functioning (American Psychiatric Association, 2013). Teachers often contribute their perspective to child study teams through behavior rating scales and clinicians make frequent use of teacher rating scales to assess students' problematic behavior and symptoms at school (Benson et al., 2019; Handler & DuPaul, 2005). Although familiar tools for classroom teachers, behavior rating scales (or any other rating scale) should not be the only tool used to collect data about students' behavior in the classroom due to their potential limitations. Minder et al. (2018) and Staff et al. (2021) highlight multiple validity issues affecting teacher behavior rating scales such as item misinterpretation, inaccurate recollection of events, halo effects, and state-based factors affecting the teacher informant. In particular, responder or rater bias, is a primary concern when examining teacher behavior rating scale data (Burns et al., 2003; DuPaul, 2003). For example, scholars (e.g., Green, 2005; Rocque, 2010) have raised concerns regarding teacher bias potentially influencing racial disparities in office discipline referrals and special education placement. Moreover, a recent systematic review in the ADHD literature

by Staff et al. (2021) highlighted a limitation within teacher behavior rating scale research; the vast majority of studies examining the validity of teacher ratings included parent informant ratings and other diagnostic assessments as part of their study design. Little research has examined teacher rating scale data about students' ADHD symptoms in isolation from parent ratings and other assessments (Staff et al., 2021).

With these limitations in mind, teacher behavior rating scales offer many benefits to school personnel, particularly their time efficiency, cost-effectiveness, and alignment with structured clinical interviews (Austerman, 2015). As such, teacher behavior rating scales will continue to contribute to special education processes in schools and research methodologies. Potential concerns about teacher rating scales validity and their limitations can be mitigated through multimethod behavior assessment approaches. A multimethod behavior assessment approach that includes direct behavioral observation is commonly regarded as best practice in schools (Individuals With Disabilities Education Act, 2004; Williford et al., 2015). Direct behavioral observations, which are more often referred to as systematic direct observation (SDOs) in the literature, are considered a "gold standard" of behavioral assessment in measuring students' classroom behavior (Jiang et al., 2019). In particular, SDOs are utilized in school-based behavior assessment approaches due to their specificity, operationalization of target behaviors, standardization, and efficiency (Salvia et al., 2010). For students with DBDs, direct observation systems offer the advantage of measuring discrete behaviors during particular times and places (i.e., classroom contexts) where students' disruptive behaviors are reported to occur frequently. Rating scales on the other hand, often utilize a summary judgment over a specified period of time (i.e., 1 week or 1 month). SDOs complement behavior rating scales as they are assumed to provide more specific, objective, and unique data (Hojnoski et al., 2020). SDOs involve systematic observer training that requires interobserver agreement criteria to be met, which in design aims to eliminate rater bias by operationalizing and objectively codifying observed behaviors (Steiner et al., 2013). Rating scales do not typically provide training and instead rely on the subjective judgment of the respondent and their understanding of the item. SDOs are idiographic in comparison to rating scale systems, which are often norm-based, thus allowing SDOs to add unique information about a child to a multimethod behavioral assessment approach.

As with any measurement approach, SDOs do also have limitations and disadvantages for use in school settings. Merrell (1993) and Steiner et al. (2013) identify six types of errors that can occur with SDOs including: poorly operationalized behaviors, situational specificity of behaviors, student reactivity affecting observations, behavior code selection errors by the observer, low interrater reliability, and observer bias. Despite aiming for rater objectivity, interobserver agreement reliability remains a paramount concern for SDO coding systems and it can be difficult for observers to reach agreement criterion. Compared with rating scales, SDOs may require significant amounts of time when training observers to reach criterion (e.g., 15–50 h depending on the assessment; Steiner et al., 2013). This is particularly problematic in school settings where training costs may prohibit year-over-year reliability training and assessments of school-based personnel conducting observations. To our knowledge, there is no comprehensive study or meta-analysis reviewing the interobserver agreement rates of SDOs between school-based personnel and research teams using the same SDO systems, thus it is unclear if observer type differences exist and if these differences change over time. As a result, initial implementation of SDOs can be costly and burdensome (Jiang et al. 2019), and the potential need each year for repeated observer training or refresher trainings to avoid observer drift can further increase costs.

Given best practice in behavioral assessment for children recommends multiple approaches and methodologies, it is common for both teacher behavior rating scales and student direct observation measures to be used in tandem in both practice and research (e.g., Riley-Tillman et al., 2008). Therefore, it makes sense to examine the relationship between teacher behavior rating scales and student direct observations. In the ADHD literature, this relationship has been studied across a large number of studies. A recent review by Minder et al. (2018) focused on the utility of SDO instruments in ADHD, found a mixed range of agreement ($r = .21-.93$) between behavioral observation instruments and teacher ratings, with mostly small to moderate ($r = .02-.50$) relationships present across studies. Expanding upon this line of inquiry, the aforementioned Staff et al.'s (2021) meta-analysis specifically focused on systematically reviewing studies examining the validity between teacher rating scales with semistructured clinical interviews and structured classroom observations. Similar to the earlier review by Minder et al., Staff et al. found

meta-analytic correlations between observational measures and teaching rating scales to be significant and range from small to moderate ($r = .12-.31$).

However, both the Minder et al. (2018) and Staff et al. (2021) reviews focused exclusively on ADHD studies and related variables with no discussion of how structured observations may relate to other DBDs and their symptomology. In comparison with ADHD literature, there exists fewer studies connecting teacher behavior rating scales with SDOs for DBDs and their related externalizing and internalizing symptoms. Specifically, there are a handful of studies (e.g., Chafouleas et al., 2005, 2012; Riley-Tillman et al., 2008; Smith et al., 2018; Steege et al., 2001) demonstrating correlations between direct behavior ratings and SDOs for disruptive behavior. Alternatively, SDOs have served as outcome measures to gauge the effectiveness of an intervention for disruptive behavior (e.g., Briesch & Daniels, 2013; Simonsen et al., 2011) as well as aggression (e.g., Kellner et al., 2008). Recently, Dart et al. (2019) conducted a systematic review of progress monitoring measures for internalizing symptoms and identified SDOs as a promising tool. While these research studies support the use of SDOs, most of the existing DBD research has examined SDOs in conjunction with brief, one to two-item behavior rating scales in contrast to larger more comprehensive rating scale systems.

There is a need for investigations that examine the convergent validity between commonly utilized student direct observation assessments and comprehensive, valid teacher behavioral rating scales in school settings specifically as it relates to DBDs and their symptoms. In addition to addressing validity questions, such investigations can offer insights into the utility of these assessments in school settings for identifying students at risk for or with DBDs and informing intervention processes. The present study addresses this gap by examining the convergent validity between two commonly used assessments in school settings: the Behavior Observation of Students in Schools (BOSS; Shapiro, 2004), a well-known student direct observation measure, and the Behavioral Assessment System for Children-3 (BASC-3 TRS [Teacher Rating Scale]; Kamphaus & Reynolds, 2007, 2015; Reynolds & Kamphaus, 2004) along with its abbreviated screener the Behavioral and Emotional Screening System, third edition (BESS-3), both of which are widely used comprehensive teacher behavior rating scales.

The BOSS (Shapiro, 2004) is a popular SDO assessment system used by school-based practitioners and researchers (e.g., Briesch & Daniels, 2013; Briesch et al., 2018; Cook et al., 2014). The BOSS measures the percent of time students are engaged in academic engagement behaviors and off-task behaviors using an interval and partial interval time sampling format. Jiang et al. (2019) highlight several advantages of the BOSS over other SDOs including evidence of treatment sensitivity, high levels of interobserver agreement, less training requirements, and peer comparison. Similarly, The BASC-3 TRS (child form), is a widely used teacher behavior rating form in schools for measuring student adaptive and maladaptive behaviors, particularly symptoms related to emotional and behavioral disorders. The BASC-3 TRS includes 156 items rated along a four-point Likert scale, that are summed to create five composite score indexes capturing adaptive skills, behavioral symptoms, externalizing problem behaviors, internalizing behaviors, and school-specific problems. A recent report by Benson et al. (2019) indicates the BASC-3 TRS is the most commonly used measure by school psychologists, which is not surprising given its long development history, comprehensive reliability analyses across versions, and utility with clinical diagnostic criteria.

Although not a direct comparison between the two measures, a small number of studies have examined convergent validity between the BOSS and rating scales measures. However, among these few studies findings have been mixed. For example, DuPaul and colleagues' (2004) investigation between the BOSS and the ADHD Rating Scale-IV and the Academic Competence Evaluation Scale yielded low and nonstatistically significant correlations. In contrast, Hosterman et al. (2008) found moderate correlations (r 's = .246-.491) between some BOSS categories and the Conners Teacher Rating Scale and ADHD Rating Scale-IV. Although they found correlational evidence for BOSS on-task categories, correlations were calculated separately based on the students' ethnicity and were not consistent across settings and groups. Hojnosi and colleagues (2020) presented limited concurrent validity evidence between a modified version of the BOSS and the Scale of Social Competence and School Adjustment for Preschool/Kindergarten (SSCSA; Missall, 2002). Moderate correlations (r 's = -.41 and .57) between the SCCA and two of the BOSS categories (i.e., off-task passive [OFT-P], AET) were found. These findings align with the small to moderate relationships between teacher rating

scales and SDOs that were identified via Minder et al. (2018) and Staff et al. (2021) meta-analyses for ADHD literature. In the context of examining students at risk for or with DBD, minimal studies have examined the BOSS and BASC-3 TRS together, and the same is true for the relationship between the BOSS and BESS-3 screener. A recent study by Jiang et al. (2019) utilized both the BOSS and the BASC-2 TRS alongside other measures of ODD and functional impairment to examine the relationship between engagement and ADH symptoms, as well as frequently co-occurring externalizing and internalizing disorders (e.g., depression, anxiety, ODD). Overall correlation magnitudes were small to moderate, mirroring Minder et al. (2018) and Staff et al.'s (2021) findings. Specific to the BOSS and BASC-2 TRS, Jiang et al. found significant but small to moderate negative relationships for teacher-rated depression on the BASC-2, ($-.35$) and the BOSS total engagement (TE), but no significant relationship was found for teacher-rated anxiety on the BASC-2 TRS and BOSS TE. Additionally, the authors found BOSS TE to significantly relate to ODD severity ($-.29$), and functional impairment ($-.21$), albeit correlational magnitude was small to moderate. To our knowledge, this is the only study explicitly utilizing the BOSS and BASC-TRs to examine externalizing and internalizing symptoms in addition to typical ADHD symptomology. However, Jiang et al. (2019) only examined two specific subscales of the BASC-2 TRS (i.e., anxiety and depression) and did not make use of the behavioral symptoms, externalizing behaviors, and school problems composites, which may be more reflective of students at risk for or with DBDs.

In light of Jiang et al.'s (2019) research, and the limited research documenting the relationship between DBD symptomology and SDOs, the current study sought to expand this work by comparing the convergence between the BOSS engagement and off-task behavior indices with the BASC-3 TRS composite scores and BESS-3 risk index for students at risk or with DBDs. Our research questions (RQs) were as follows:

- (1) What is the relationship between the BOSS on-task indices of active engaged time (AET), passive engaged time (PET), and total engaged time with BASC-3 TRS five composite scores (Adaptive Skills, Behavioral Symptoms Index, Externalizing Problems, Internalizing Problems, and School Problems)?
- (2) What is the relationship between the BOSS off-task behavior categories of off-task motor (OFT-M), off-task verbal (OFT-V), and OFT-P and the BASC-3 TRS five composite scores?
- (3) What is the relationship between the BOSS on-task and off-task indices and the BESS-3 Emotional Risk T-score?

Given the general consensus of small to moderate correlations between SDOs and teacher rating scales from meta-analyses (i.e., Minder et al., 2018; Staff et al., 2021), as well as similar correlational magnitude in related work (Jiang et al. 2019), we hypothesized correlational strength between BOSS indices and the BASC-3 TRS and BESS-3 would be similarly small to moderate in magnitude. For RQ1, it was hypothesized that the BOSS AET, PET, and TE would positively correlate with the BASC-3 Adaptive Skills composite due to their shared positive behavior valence. Similarly, we hypothesize that BOSS AET, PET, and TE would negatively correlate BASC-3 Behavioral Symptoms Index, Externalizing Problems, Internalizing Problems, and School Problems composites due to their difference in measuring positive versus negative behaviors. Conversely, for RQ2 we hypothesized the BOSS indices of OFT-M, OFT-V, and OFT-P would negatively correlate with the BASC-3 Adaptive Skills composite and positively correlate with the Behavioral Symptoms Index, Externalizing Problems, Internalizing Problems, and School Problems composites. For RQ3, it was hypothesized that the BOSS engagement indices of AET, PET, and TE would negatively correlate, and the BOSS disruptive behavior indices of OFT-M, OFT-V, and OFT-P would positively correlate with the BESS-3 Emotional Risk T-score.

2 | METHOD

Data in the current study are derived from 136 classroom teachers from 57 elementary schools who taught 349 students with or at-risk for DBDs that participated in a randomized controlled trial (RCT) focused on job-embedded coaching for paraprofessionals. The current investigation analyzed the relationship between available baseline

assessment data for the participating teachers and students. All study procedures were approved by the university's institutional review board (IRB). Participants provided informed consent and volunteered to participate in the study.

2.1 | Participants

The sample was comprised of 57 elementary schools (kindergarten through fifth grade) in the northeast United States and included participants from school years 2017–2020. Schools were located in urban and suburban areas. School economic status, as measured by the percentage of students receiving free and reduced lunch varied across the sample with an average 52% of students receiving free and reduced lunch (standard deviation [SD] = 31.41; range from 0% to 96%).

2.1.1 | Teachers

Teacher participant demographics are presented in Table 1. The present sample included 136 teachers that were between 24 and 69 years old (Mean [M] = 46.77, SD = 11.48). The vast majority of the teachers self-identified as female ($N = 132$, 96.21%). Teachers identified their race/ethnicity as either Hispanic/Latinx ($n = 21$, 15.44%), Black ($n = 31$, 22.79%), or White ($n = 71$, 52.21%), with 13 teachers not reporting on this variable ($n = 13$, 9.56%). In regard to education level, the majority of teachers had either a bachelor's degree ($n = 66$, 48.52%) or a graduate degree ($n = 47$, 34.55%), whereas a small number had only some college experience ($n = 4$, 2.94%), and some teachers chose not to report on this variable ($n = 19$, 13.97%). A large portion of teachers possessed more than 10 years of teaching experience ($n = 57$, 41.91%), with smaller portions of teachers possessing between 5 and 10 years' experience ($n = 25$, 18.38%), between 2 and 5 years' experience ($n = 15$, 11.02%), and less than 2 years' experience ($n = 16$, 11.76%); a small portion of teachers also chose not to report on this variable ($n = 23$, 16.17%).

2.1.2 | Students

Student participant demographics are presented in Table 1. The 349 students participating in this study were on average 7.53 years old (5–12 years; SD = 1.97) and the majority were male ($n = 266$, 77.33%). Grade level distribution for the 349 students included 91 (26.15%) kindergarteners, 51 (14.66%) first graders, 37 (10.63%) second graders, 56 (41.17%) third graders, 63 (16.09%) fourth graders, and 49 (14.08%) fifth graders.

2.2 | Measures

2.2.1 | BOSS

The BOSS utilizes an interval and partial interval method for measuring students' academic engagement (i.e., time spent on the academic task at hand) and off-task behaviors (i.e., disruptive behaviors). Observations are brief (e.g., 30 min) and collected data are reported as a percentage of intervals observed for each behavior, from which a percentage of time can be inferred. Unique to the BOSS, is the division of on-task behavior codes into AET and PET, which more specifically define student academic engagement compared to previous definitions and studies (Greenwood, 1996). The BOSS measures three different types of off-task behaviors: OFT-M, which is defined as physical behaviors that pull students' focus away from academic instruction (i.e., walking around the classroom); OFT-V, verbal expression not relevant to the lesson content (i.e., discussing television with a peer); and OFT-P,

TABLE 1 Participant demographic information.

Teacher demographics		n = 136
Age		M = 46.77 SD = 11.48 Range = 24–69
Gender		
Female		132 (97.06%)
Male		4 (2.94%)
Race/ethnicity		
Hispanic/Latinx		21 (15.44%)
Black or African American		31 (22.79%)
White		71 (52.21%)
Not reported		13 (9.56%)
Education level		
Some college		4 (2.94%)
Bachelor's degree		66 (48.52%)
Graduate degree		47 (34.55%)
Not reported		19 (13.97%)
Years of teaching experience		
Less than 2 years		16 (11.76%)
Between 2 and 5 years		15 (11.02%)
Between 5 and 10 years		25 (18.38%)
More than 10 years		57 (41.91%)
Not reported		22 (16.91%)
Student demographics		n = 349
Age		M = 7.53 SD = 1.97 Range = 5–12
Gender		
Female		132 (97.06%)
Male		4 (2.94%)
Grade level		
Kindergarten		91 (26.15%)
1st grade		51 (14.66%)
2nd grade		37 (10.63%)
3rd grade		56 (41.17%)
4th grade		63 (16.09%)
5th grade		49 (14.08%)

Abbreviations: M, mean; SD, standard deviation.

defined as behaviors that pull student focus from academic instruction that are not physical or verbal in nature (i.e., looking out the window).

The BOSS form was modified to meet the behavioral assessment needs of the overall RCT targeting disruptive behaviors in classrooms (see Table 2 for modified definitions). It is commonplace to adjust which codes are used with the BOSS for research purposes (Briesch et al., 2018), and the present BOSS modifications closely matched the adaptation of the BOSS used by DuPaul et al. (2004) in their examination of ADHD predictors. Specifically, the form was modified to incorporate additional behavior definitions and examples that reflect disruptive, aggressive, and defiant behaviors students may express in classrooms. For the larger RCT and the current study, on-task and off-task behaviors were coded across six behavior categories/indices and the definitions for each along with examples are provided in Table 2. On-task behaviors included active engagement (AE; e.g., answering a teacher's question or actively writing), passive engagement (PE; e.g., listening to teacher or reading quietly). A composite TE, was created by adding the AE and PE totals together. Off-task behaviors included inappropriate physical (IP; e.g., fidgeting, out of seat, hitting a peer), inappropriate verbal (IV; e.g., humming, calling out of turn, using profanity), and Disruptive Academic (DA; e.g., off-task, daydreaming). Like DuPaul et al. (2004), an NC code (e.g., purposefully not following a directive) was added to better understand when students refused to follow a teacher directive. Unlike the traditional BOSS form, the current study did not measure the quantity of teacher-directed instruction intervals observed (DuPaul et al., 2004). To reduce confusion between PE and DA, observers were instructed to only measure DA when the student was deemed off-task for more than 3 consecutive seconds. Further, PE codes were able to be changed during the interval if the student was observed to be daydreaming within 3 s of the momentary time interval. These correction procedures reduced overlap in codes between PE and DA.

Each observation was 15 min in length, and included 60 intervals (15 s each). At the beginning of the interval, observers utilized momentary time sampling to code for student academic engagement (i.e., on-task behavior), specifically the presence of AE or PE. During the remainder of each interval, the partial interval recording method was used to code the four behavior categories of IP, IV, NC, and DA. For the NC code, if the student was given a directive by the teacher or paraprofessional and purposely did not comply with the request by the end of the next 15 s interval, the behavior was coded as NC. After the 15-min observation was completed, behavior codes were tallied across intervals to generate each index score and the TE score was generated by combining AE and PE categories into one code.

Evidence supporting psychometric validity of the BOSS has been reported across several studies. Interobserver agreement ranges between 90% and 100% (Ota & DuPaul, 2002; Shapiro, 2004) and strong kappa coefficients have been found (DuPaul et al., 2004) for pairs of observers co-observing with the BOSS. Studies have demonstrated discriminant validity and treatment sensitivity between typically developing children and children with ADHD (DuPaul et al., 2004; Ota & DuPaul, 2002).

2.2.2 | BASC-3 TRS

The BASC-3 TRS (Child Form), designed for students 6–11 years old, was utilized to capture teachers' perceptions of student behavioral and emotional difficulties in the classroom. The BASC-3 TRS encompasses 156 items and describes observable positive and negative student behaviors that may occur in the classroom. The 156 items are organized into five composite scores: (1) Adaptive Skills (i.e., adaptability, social skills, functional communication, leadership, and study skills); (2) Behavioral Symptoms Index (i.e., hyperactivity, aggression, depression, attention problems, atypicality, and withdrawal); (3) Externalizing Problems (i.e., hyperactivity, aggression, and conduct problems); (4) Internalizing Problems (i.e., anxiety, depression, and somatization); and (5) School Problems (i.e., learning problems, attention problems). Teachers rate the frequency of observed student behavior items through a four-point Likert scale ($N = \text{never}$, $S = \text{sometimes}$, $O = \text{often}$, $A = \text{almost always}$). The items are then summed according to their representative scale, which yields a raw score for each scale that is converted to a T -score based on the norming samples.

TABLE 2 Modified BOSS behavioral definitions and examples.

Behavior	Operational definitions	Examples
Active Engagement (AE)	Times when the student is actively attending to the assigned work; on-task and actively participating	Writing; reading aloud; raising a hand to answer a teacher's question; talking to the teacher about the assigned material; talk to a peer about the assigned material; looking up a word in a dictionary; typing essay on computer
Passive Engagement (PE)	Times when the student is passively attending to the assigned work; on task and passively participating	Listening to a lecture; looking at an academic worksheet; reading assigned material silently; looking at the blackboard during teacher instruction listening to a peer respond to a question, watching a video, viewing lesson on the smartboard
Inappropriate Physical Behavior (IP)	A forceful movement directed at another person, either directly or by utilizing a material object as an extension of the hand A forceful movement directed at an inanimate object or inflicts physical damage on an object Physical behavior that interferes with or disrupts classroom functioning and/or makes it difficult for others to perform their work	Hitting, biting, kicking, pinching, raising a clenched fist or open hand and swinging toward another, throwing objects toward another person Damaging property (tearing papers) or objects (ripping book bags); kicking a bookcase; throwing an object forcefully onto the floor Taking objects from others without asking permission; excessive or inappropriate motor activity either in or out of his/her seat; out of seat; lack of body control; moving around the room; leaving an assigned area and/or leaving the group, tapping peers, making inappropriate gestures, drawing on assignment, playing with objects at desk
Inappropriate Verbal Behavior (IV)	Verbalizations that are abusive or threatening and directed toward other people; all negative, noncontact communication Verbal behavior that interferes with or disrupts classroom functioning and/or makes it difficult for others to perform their work	Verbal bullying, verbal threats, tattling, teasing, name-calling; cursing Initiating conversations during quiet work periods; calling out in class; making noises that distract others (e.g., animal noises, grunts); verbal interruption; talk outs; inappropriate elevated voice level; lack of emotional control (e.g., crying)
Noncompliance (NC)	If (a) the child is <i>given an instruction/direction</i> by an aide or teacher and <i>purposely</i> makes no effort to comply with the request by the end of the next observation interval (i.e., within 15 s), or (b) the <i>teacher reprimands the child or reminds</i> him/her to follow earlier instructions	Not following directions; not completing or starting assigned work
Disruptive Academic Behavior (DA)	For at least 3 consecutive seconds: Attending to any stimulus or activity other than the one assigned; not directing focus toward the teacher during presentation of a lesson; "does nothing"	Off-task; staring blankly/daydreaming; looking out of the window when directed to complete work, looking at teacher while she is talking to another student

Note: BOSS behavior codes and definitions were adapted and modified from Shapiro (2011) and Sheridan et al. (2012).

The 3rd edition of the BASC-3 TRS has strong internal consistency (i.e., .92–.97), fair to good test–retest reliabilities for (i.e., .77, .91), and weak to fair interrater reliabilities (i.e., .37–.73). As aforementioned, the BASC-3 has demonstrated convergent correlations with its earlier predecessor the BASC-2 (Reynolds & Kamphaus, 2004), as well as the Achenbach System of Empirically Based Assessment (ASEBA) (Achenbach & Rescorla, 2001), the Conners 3rd Edition (Conners, 2008), and the Autism Spectrum Rating Scales (Goldstein & Naglieri, 2010). For the ASEBA, correlations ranged from .27 to .76, whereas correlations with the Conners 3rd edition ranged between .29 and .90.

2.2.3 | BESS-3

The BESS-3 Teacher Form (TF), used for children 6–12 years old, is a brief emotional and behavioral screener that measures the risk of student behavioral and emotional problems (Kamphaus & Reynolds, 2015). It contains 20 items that are scored using a four-point Likert frequency rating scale (0 = *never* to 3 = *almost always*), with higher scores signifying a greater risk for student behavior and emotional problems. The ratings from the 20 items are combined to generate a single *T*-score ($M = 50$, $SD = 10$), which then classifies students' behavioral risk into three categories: normal ($T < 59$), elevated risk ($60 \leq T < 69$) or extreme risk ($70 \leq T$). In the present study, the vast majority of students had *T*-scores in the elevated to extreme risk ranges.

The BESS-3 has demonstrated acceptable reliability indices for split-half reliability (.96), test–retest reliability (.91), and acceptable internal consistency (.83) (Kamphaus & Reynolds, 2007). Convergent validity with other teacher behavior rating scales such as the ASEBA (r 's = .71–.77), Conner's rating scales (r 's = .51–.78), Children's Depression Inventory ($r = .51$) and the Vineland Adaptive Behavior Scales (r 's = .32–.69) have been documented in BESS-3 validation research (Kamphaus & Reynolds, 2007, 2015). Convergent validity between the BESS-3 total risk score and school report-card outcomes have also been established (Eklund et al., 2009).

2.3 | Procedures

Data were collected during the larger RCT's baseline period (before intervention, fall) each year (2017–2020). Districts and schools were recruited via email correspondence and in person presentations. Participating teachers were recruited via in-person recruitment meetings, followed by an informed consent process in accordance with IRB policies. Participating students and parent/guardian approval were recruited through an IRB-approved consent process that utilized an opt-out procedure. A study notification letter was sent home to all parents/guardians informing them about the study and provided instructions on how to contact classroom teachers and study authors if they *did not want* their child to participate in the study.

To participate in the larger RCT, teachers nominated 3–5 students with or at-risk for DBDs. For each student, teachers completed a corresponding BESS-3 TF. If the teacher ratings of the student produced a *T*-score of 60 or greater, students were eligible to participate in the study. For classrooms where more than three students had a *T*-score greater than 60, the three students with the highest *T*-scores were selected. Students diagnosed with developmental disabilities were not eligible for study participation.

Once the final participating students from each classroom were selected via the BESS-3 nomination process, teachers then scheduled BOSS observations with study coordinators. For BOSS, students were observed three times within a 10-day period by independent observers using the modified BOSS (i.e., 45 min total per student). Teachers provided study coordinators with observation times when students' disruptive behaviors were most prevalent. Observations occurred during whole-group, small-group, or individualized instruction when the classroom teacher and their paraprofessional were present. Following the BOSS observations, teachers completed BASC-3 TRS forms within 2 weeks.

Independent observers were trained in the BOSS through 4 h of didactic instruction, video examples, discussion, practice, and feedback from the trainer. Observers were required to achieve a reliability criterion of 80% or above to perform student observations. Independent observers were blind to study condition.

Interobserver agreement observations were conducted as part of the larger RCT study that the current convergent validity study draws data from. For interobserver agreement, two independent observers observed the same student on the same day and time using the BOSS measure. Interobserver agreement was calculated by comparing (i.e., Pearson correlations) the total scores for each BOSS index between observer pairs for each interobserver agreement observation conducted. The current study evidenced high levels of agreement between observer pairs with r 's of .92 to .80 across the six modified BOSS indices. The agreement coefficients in present study are consistent in magnitude with previous literature documenting BOSS observer agreement.

2.4 | Data analytics

For the BASC-3 TRS and BESS-3 TF, all data were entered into the Pearson QGlobal electronic data management system. The QGlobal system performed the automated tasks of generating scale scores, composite scores, and T -scores according to the norming samples of the assessments. Descriptive statistics were subsequently calculated across students for the BASC-3 TRS composite T -scores and the BESS-3 TF Emotional Risk T -score. In regard to the modified BOSS, baseline observation scores were averaged together for each student to produce an average baseline BOSS score per category (e.g., 0.27 for IP) per student. Next, descriptive statistics were generated for BOSS averaged scores across the student sample for each BOSS category.

Pearson correlation coefficients were computed between each student's averaged baseline BOSS score and each student's five composite scores on the BASC-3 TRS. The same process was repeated for the correlations between students' averaged baseline BOSS scores and the BESS-3 TF score for each student. Correlations between .10 and .20 were considered small, between .30 and .40 were medium, .50 and .60 large, and .70 and .80 were very large (Cohen, 1992).

3 | RESULTS

3.1 | BOSS

Descriptive statistics for the seven BOSS indices are provided in Table 3. BOSS scores evidenced variability in each of the indices, with behaviors observed to occur at a minimum of 0% of the time and a maximum of 98% of the time. Across observations, students appeared to have a similar level of observed AE ($M = 27\%$) and PE ($M = 32\%$), with an average TE level of 59% ($SD = 0.18$). The PE behavior was the most frequently observed behavior followed by AE. For off-task behaviors, IP was observed on average 25% of the time ($SD = 0.16$), and was followed by DA with an average occurrence of 16% of the time ($SD = 0.12$) and IV with an average occurrence of 15% of the time ($SD = 0.13$). The least observed behavior was NC, occurring only 2% of the time ($SD = 0.05$). In general, off-task behaviors occurred individually less than on-task behaviors, however, IP and AE were observed at fairly similar rates (i.e., 25% and 27% respectively).

3.2 | BASC-3 and BESS-3

Table 3 includes descriptive statistics for the teacher ratings of student behaviors on the BASC-3 TRS and BESS-3 TF. For the BASC-3 TRS, all five composites presented T -scores with scores between a minimum of 40 and a maximum of 115. The two highest teacher-rated composites were the Externalizing Problems ($M = 70.83$,

TABLE 3 BOSS descriptive statistics for the BOSS, BASC-3, and BESS-3.

	M	SD	Range
BOSS categories			
% Active Engagement (AE)	0.27	0.15	0.00–0.76
% Passive Engagement (PE)	0.32	0.15	0.02–0.79
% Total Engagement (TE)	0.59	0.18	0.02–0.98
% Inappropriate Physical (IP)	0.25	0.16	0.00–0.95
% Inappropriate Verbal (IV)	0.15	0.13	0.00–0.81
% Noncompliance (NC)	0.02	0.05	0.00–0.45
% Disruptive Academic (DA)	0.16	0.12	0.00–0.64
BESS Emotional Risk score	70.56	8.65	45–93
BASC-3 Composite scores			
Adaptive Skills	35.42	6.865	18–56
Behavioral Symptoms Index	70.77	12.04	41–106
Externalizing Problems	70.83	15.07	41–115
Internalizing Problems	59.05	14.55	39–106
School Problems	64.38	9.78	42–85

Abbreviations: BASC, Behavioral Assessment System for Children; BESS, Behavioral and Emotional Screening System; BOSS, Behavior Observation of Students in Schools; M, mean; SD, standard deviation.

SD = 15.07) and Behavioral Symptoms Index ($M = 70.57$, $SD = 12.61$) composites, which evidenced comparable means and SDs. The third highest composite, School Problems, had an average score of 64.38 ($SD = 9.08$). This was followed by the Internalizing Problems ($M = 58.55$, $SD = 15.49$) composite, and the lowest-rated composite was Adaptive Skills ($M = 35.22$, $SD = 7.35$). Teacher ratings of student behavioral and emotional risk on the BESS-3 TF evidenced an average Emotional Risk T -score of 70.56 ($SD = 8.65$), with a score range of 45–93. The mean Emotional Risk T -score indicates that the current student sample was rated in the extreme risk category.

3.3 | Correlations between BOSS and the BASC-3 and BESS

Table 4 presents Pearson correlations between the BOSS on-task and off-task behaviors and teacher ratings of student emotional and behavioral difficulties on the BASC-3 TRS and BESS-3 TF. Each of the BOSS indices evidenced significant correlations with the BASC-3 TRS composites. In general, correlations were in the hypothesized direction and magnitude, such that less on-task and greater disruptive behaviors on the BOSS were associated with greater teacher-rated behavioral and emotional problems on the BASC-3.

For RQ1, the BOSS categories of on-task behavior evidenced significant small correlations with the majority of BASC-3 TRS Composites. AE demonstrated a significant small positive correlation with the BASC-3 Adaptive Skills composite ($r = .23$) and significant negative small correlations with the Externalizing Problems ($r = -.27$), Behavioral Symptoms Index ($r = -.24$), and School Problems ($r = -.16$) composites. PE had significant small negative correlations with the Externalizing Problems ($r = -.16$) and Behavioral Symptoms Index ($r = -.16$) composites. TE exhibited a significant small positive correlation with the Adaptive Skills composite ($r = .27$), significant moderate negative correlations with the Externalizing Problems ($r = -.36$) and Behavioral Symptoms Index ($r = -.33$) composites, and

TABLE 4 Pearson correlations between BOSS, BASC, and BESS scores.

BOSS Indices	BESS Emotional Risk	BASC Adaptive Skills	BASC Behavioral Symptoms Index	BASC Externalizing Problems	BASC Internalizing Problems	BASC School Problems
% Active Engagement	-.28**	.23**	-.24**	-.27**	-.07	-.16**
% Passive Engagement	-.14**	.09	-.16**	-.16**	-.09	-.09
% Total Engagement	-.35**	.27**	-.33**	-.36**	-.14*	-.20**
% Inappropriate Physical	.26**	-.22**	.24**	.30**	.04	.14*
% Inappropriate Verbal	.21**	-.18**	.21**	.35**	.01	.16**
% Noncompliance	.21**	-.09	.21**	.23**	.08	.05
% Disruptive Academic	.19**	-.19**	.13*	.11*	.01	.10

Abbreviations: BASC, Behavioral Assessment System for Children; BESS, Behavioral and Emotional Screening System; BOSS, Behavior Observation of Students in Schools.

* $p > .05$; ** $p > .01$.

significant small negative correlations with the School Problems ($r = -.20$) and Internalizing Problems ($r = -.14$) composites.

For RQ2, the BOSS categories for disruptive behavior primarily demonstrated significant small correlations with the BASC-3 TRS Composites. IP had a significant moderate positive correlation with the Externalizing Problems composite ($r = .30$), significant small positive correlations with the Behavioral Symptoms Index ($r = .24$) and School Problems composites ($r = .14$), and a significant small negative correlation with the Adaptive Skills composite ($r = -.22$). IV evidenced a significant moderate positive correlation with the Externalizing Problems composite ($r = .35$), significant small positive correlations with the Behavioral Symptoms Index ($r = .21$) and School Problems ($r = .16$) composites, and a significant small negative correlation with the Adaptive Skills composite ($r = -.18$). NC exhibited significant small positive correlations with the Externalizing Problems ($r = .23$) and Behavioral Symptoms Index ($r = .21$) composites. DA had a significant small positive correlation with the Behavioral Symptoms Index ($r = .13$) and Externalizing Problems ($r = .11$) composites, and a significant small negative correlation with the Adaptive Skills composite ($r = -.19$).

For RQ3, each of the BOSS indices evidenced significant small to moderate correlations with the BESS-3 TF Emotional Risk *T*-score. Significant negative correlations were found between the BESS-3 score and BOSS on-task behaviors, and significant positive correlations were found with the BOSS off-task behaviors. Specifically, TE ($r = -.35$) demonstrated significant moderate negative correlations with the BESS-3 score and AE ($r = -.28$) and PE ($r = -.14$) had significant small negative correlations. The BOSS off-task behaviors of IP ($r = .26$), IV ($r = .21$), NC ($r = .21$), and DA ($r = .19$) had significant small positive correlations with the BESS-3 score. Collectively, results suggest that observed decreases in student on-task behaviors and observed increases in student off-task behaviors were associated with increases in teacher-perceived emotional and behavioral risk.

4 | DISCUSSION

Findings from the current study demonstrate meaningful relations (i.e., small to moderate correlations) between the BOSS observational measure and the BASC-3, BESS-3 screener form and the full BASC-3 TRS form. Correlational results suggest that increased student behavioral symptoms and risk as rated by teachers was associated with less observed on-task behaviors and more observed off-task behaviors of students in elementary schools. Results of this study contribute to the limited convergent validity research supporting the use of teacher-rated behavior rating scales independently from parent-rated behavior rating scales, and in conjunction with a student direct observational measure. Additionally, we provide evidence that adds to the convergent validity of BOSS in schools and offer insights for school practitioners in using multimethod behavior assessments.

4.1 | Relationship between BOSS on-task scores and BASC-3 composite scores

Regarding the first RQ, significant relations in the expected directions were found between the BOSS on-task behavior indices (i.e., AE, PE, and TE), and the BASC-3 five composite scores. The positively valenced constructs of AE and TE correlated in the same direction as the positively valenced Adaptive Skills construct on the BASC-3, whereas PE did not evidence a significant relationship. In regard to the BASC-3 four negative symptoms indices (i.e., Behavioral Symptoms, Externalizing Problems, Internalizing Problems, and School Problems), the BOSS AE index had more robust and significant quantity of correlations (i.e., significant findings with all but Internalizing Problems composite) compared with the PE (only two small significant correlations). The TE was found to be significantly related with the four negative symptoms indexes on the BASC-3, with correlations ranging between small and moderate. In general, results in this study offer initial evidence of convergent validity supporting the BOSS on-task behavior indices for disruptive behaviors. Student engagement is associated with academic competence and school success, especially for younger students (Hojnoski et al., 2020). Thus, it would make sense for engagement

behaviors on BOSS to be related with other positive student outcomes, such as the BASC3 Adaptive Skills composite. In contrast, the negative disruptive behaviors on the BOSS (off-task behavior indices of IP, IV, NC, and DA) can detract from and interfere with a students' level of engagement. Subsequently, we would expect to observe lower levels of engagement in students who are at increased risk for DBDs, given the complexity of the characteristics associated with DBDs (Hinshaw, 1992). A student's level of engagement in the classroom may indicate the progress or effectiveness of a specific intervention, which underscores the importance of sensitive, reliable, and valid tools such as the BOSS.

The difference in correlational findings between AE and PE with the BASC-3 may be due to the BOSS' ability to discriminate between different types of student engagement. The distinction between AE and PE on the BOSS is a unique feature compared to other SDO measures that capture engagement as one global construct. Since the AE category measures overt behaviors (e.g., raising hand), it may be easier for observers to capture observed engagement in students compared with PE (e.g., silently attending). The nuances between AE and PE might also be explained by the classroom environment and teacher instructional demands, which impact whether students are required to be actively or passively engaged based on lesson format. Interestingly, TE evidenced significant small to moderate negative correlations with all BASC-3 indices, which were more robust than AE and PE alone. Further, TE was the only category on the BOSS to have a significant correlation with the BASC-3 Internalizing Problem composite, which suggests that it might also have utility in indicating the presence of internalizing symptoms in younger students (Jiang et al., 2019).

4.2 | Relationship between BOSS off-task scores and BASC-3 composite scores

For RQ2, correlations between the BOSS disruptive behavior categories of IP, IV, NC, and DA and the BASC-3 TRS composites were mostly in the expected direction and consistent with our hypotheses. The behaviors of IP (r 's = $-.22$ to $.24$) and IV (r 's = $-.18$ to $.35$) had significant small to moderate correlations with all BASC-3 indices except the Internalizing Problems composite. IP and Verbal behaviors are overt and easily observable behaviors by both teachers and observers, and are indicative of DBDs. Thus, we would expect to see these behaviors correlate with other measures of overt negative behavior that may impact student learning and be of concern to teachers, like the BASC-3 Behavioral Symptoms and Externalizing Problems indices. Similarly, IP and IV were the only negative behaviors on the BOSS to correlate with the School Problems Composite, suggesting these types of overt behaviors may be more problematic for school settings. Beyond their overt nature, the IP and IV behaviors may also cause the most classroom disruption for teachers and the entire class, not just the individual student displaying the behavior, and may stand out in the teachers' mind when completing rating scale assessments. For example, IP and IV behaviors may demonstrate a low frequency of occurrence, but high severity of disturbance, which may adversely disrupt the flow of classroom instruction depending on their severity. In contrast, the BASC-3 Internalizing Problems composite assesses students' inwardly directed symptomatology and may be more challenging for teachers to identify and accurately rate (Dart et al., 2014). It is possible that the students in this study had higher rates of internalizing symptoms than was captured by the teacher rating scales.

The BOSS categories of NC (r 's = $-.09$ to $.23$) and DA (r 's = $-.19$ to $.13$) had small, significant correlations with the Adaptive Skills, Behavior Symptoms, and Externalizing Problems composites, but had nonsignificant correlations with the Internalizing Problems and School Problems composites. NCs were dependent on contextual factors in the classroom (e.g., educator providing a command to the student) and can be considered an act of direct defiance by the student (e.g., not taking out textbook). Thus, the inward-directed symptomatology measured in the Internalizing Problems and School Problems composites are of a different nature than those captured by NC. For DA, it is somewhat unexpected that the correlations with Internalizing Problems and Schools Problems composites were nonsignificant. Similar to PE, it could be that since this category captures negative passive (e.g., daydreaming) behaviors, the independent observers may have been unable to track them as easily as the other categories, which are more active overt behaviors. As previously mentioned, there is also a general difficulty with accurately measuring internalizing problems (Bradshaw et al., 2008; Dart et al., 2014).

The strength of the correlations was generally more robust when considering the relationship between the BOSS and the BASC-3 TRS Externalizing Problems composite. As the BOSS was used to measure student disruptive and externalizing behaviors in the classroom, it was not surprising that the convergence would be highest with the Externalizing Problems composite. However, the strength of correlations between the two measures in the current study may have been attenuated by the restricted range of observed disruptive behavior. Although the current study included students at risk for, or with disruptive behaviors, the overall observed disruptive behavior rates are potentially low in comparison to other studies. As such, the potential for negative skew in the observed disruptive behaviors on the BOSS may have reduced overall correlational magnitude with the BASC3 forms (Kendall & Stuart, 1958). Despite this, correlational magnitude in the current study was in line with findings from Minder et al. (2018) and Staff et al. (2021), as well as our hypothesized magnitude. Overall, correlational appear to provide additional evidence for the utility of the BOSS for measuring student disruptive/externalizing behaviors in the classroom.

Finally, for the third RQ, the BOSS indices evidenced significant, small to moderate correlations with the BESS-3 TF Emotional Risk score in the expected directions (i.e., negative correlations for BOSS Engagement categories and positive correlations for BOSS disruptive behaviors indices). The results of the present study support the use of the BESS-3 screener in conjunction with the BOSS. In fact, the results of the present study give evidence for the use of the BESS-3 TF Emotional Risk score to initially identify students in the classroom, who can then be observed with the BOSS for gathering additional evidence supporting the risk for behavior symptoms. Combined, the multimethod approach of a brief behavioral symptoms screener and valid SDO system has the potential to inform referrals, child study teams, and special services in schools.

Hosterman et al. (2008) state that prior research is understudied and inconclusive regarding the convergence of teacher rating scales with classroom observations. Despite this, it is generally considered best school-based practice to utilize classroom observations with teacher rating scales as part of the behavioral assessment battery (Jiang et al., 2019). The results of the present study demonstrate the validity of using the BOSS alongside a prevalently-used teacher rating scale, the BASC-3 TRS. Specifically, our findings indicate that school practitioners can be reasonably confident that the BASC-3 TRS Externalizing Problems composite in combination with the BOSS will provide a comprehensive assessment of student disruptive behaviors in the classroom.

Despite its frequent utilization in school-based practice and research, the existent psychometric evidence for the BOSS as it relates to DBDs (excluding ADHD) is sparse and inconclusive (Moffett & Morrison, 2020). Prior research has mainly focused on using the BOSS in combination with teacher rating scales of ADHD symptomatology. Specifically, both DuPaul et al. (2004) and Hosterman et al. (2008) found some significant correlations (e.g., low to moderate) between the BOSS and ADHD Rating Scale-IV. While the results of the current investigation are consistent with the findings of DuPaul et al. (2004) and Hosterman et al. (2008), our study adds to this work by studying convergent validity of the BOSS, BASC3, and BESS-3 with a sample of students with and at risk for DBDs in K-5 grade settings. Additionally, compared with the ADHD Rating Scale-IV, which assesses students for ADHD, the BASC-3, and BESS-3 were designed to create a broader behavioral profile of the student. It is also important to note that the DuPaul et al. (2004) investigation and the Hosterman et al. (2008) concentrated on students with ADHD. Compared with both of these studies, the present investigation supports the utility of the BOSS with students at risk for or with disruptive behaviors.

It is also important to compare our findings with those presented in Hojnoski and colleagues (2020) recent investigation, which examined the psychometrics of the BOSS in relation to measuring engagement and disruptive behaviors in preschool settings as measured by the SSCSA for Preschool/Kindergarten (Missall, 2002). Hojnoski and colleagues (2020) found that the SSCSA Total score had a moderate, positive, and statistically significant correlation ($r = .57$) with the BOSS index of AE and a moderate, negative, and statistically significant correlation ($r = -.41$) for passive interfering behaviors. In our study, AE had small, significant correlations with the BESS-3 TF Emotional Risk score ($r = -.28$) and BASC-3 TRS composites of Adaptive Skills ($r = .23$), Behavioral Symptoms Composite ($r = -.24$), Externalizing Problems ($r = -.27$), and School Problems ($r = -.16$). Similarly, DA behavior (congruent to passive interfering behaviors in Hojnoski et al., 2020) had small, significant correlations with the BESS-3 TF Emotional Risk

score ($r = .19$) and BASC-3 TRS composites of Adaptive Skills ($r = -.19$), Behavioral Symptoms Composite ($r = .13$), and Externalizing Problems ($r = .11$). Unlike Hojnosi and colleagues (2020) study, our findings yielded more significant correlations, but did not have substantially greater robustness in terms of correlational strength (e.g., the highest r value was $-.36$ in our study). This may be due to Hojnosi and colleagues (2008) focusing on preschool children and modifying the BOSS accordingly whereas the grade-level focus in the current study was kindergarten through fifth grade.

4.3 | Implications for school practice

Findings from this study offer suggestions for school-based practice. When selecting, implementing, tailoring, and evaluating interventions for students with disruptive behaviors, it is crucial to collect accurate behavioral assessment data. Best practice in behavioral assessment of students involves multimethod and multi-informant approaches. Given SDOs reputation as a “gold standard” for behavioral assessment in classrooms (Jiang et al., 2019) and the frequent use of teacher behavior rating scales in school psychology practice (Benson et al., 2019), the odds are high that both SDOs and teacher behavior rating scales will be used to facilitate identification, intervention planning, progress monitoring, and evaluation of students with DBDs. Individually, each methodology possesses advantages and disadvantages, thus best practice strongly advises against relying on a single methodology. However, the combined concurrent use of both SDOs and teacher behavior rating scales as part of screening, diagnostic, and intervention processes offers tremendous advantages for overcoming the limitations within each methodology and painting a more comprehensive view of students' behavior.

For example, SDOs offer the benefit of measuring student behavior discretely during specific times, settings, and contexts. Additionally, behavioral definitions can be operationalized to specific situational circumstances in the classroom such as in the case of low-frequency or covert behaviors (Fabiano et al., 2004). This allows for greater flexibility and adaptation in capturing data relevant to identifying and quantifying students' problematic behavior, especially when it is highly situational. In contrast, teacher behavior rating scales may lack this level of specificity as they often involve a retrospective reflection across time periods (e.g., 1 week, 1 month) and classroom contexts and involve a summary judgment of a behavior versus objectively quantified recorded instances of the behavior. However, the same temporal and contextual specificity that favor the use of SDOs are also a potential weakness. Low-frequency, high-severity behavioral occurrences in classrooms may not be easily captured by SDOs despite their ability to specifically operationalize behavior during specific times and contexts; quite simply, they may not be observed. Teacher behavior rating scales on the other hand, may be able to more accurately assess the conditions, behavioral symptoms, and severity of such occurrences. Moreover, SDOs may not be able to capture internalizing symptoms and behaviors, as well as data related to student affect. Rating scales on the other hand, particularly multi-informant, may provide an avenue to capture such information.

SDOs like the BOSS can provide objective and unique data and can be used in conjunction with other behavioral assessment methods, like rating scales, to capture a more holistic picture of student behavior. SDOs complement behavior rating scales as they are assumed to provide more specific, objective, and unique data (Hojnosi et al., 2020) that does not come from potentially biased sources, such as parents or classroom teachers. Well-developed SDOs require systematic training of observers and standardized procedures for implementation that permit the examination of interrater reliability in training and in practice (Steiner et al., 2013), whereas multi-informant rating scale systems do not typically train raters, nor include objective item definition and interpretation. This can lead to misunderstanding and misinterpretation, especially when we consider that teachers are not trained as clinical diagnosticians. However, rating scales offer the benefit of cost-effectiveness and feasibility. In regard to SDOs, the costs of training and the amount of time required for observers to reach agreement criteria and become proficient in using an SDO may be prohibitive for uptake by schools, especially if refresher trainings, drift assessments, and new staff require trainings each year. Although they make use of standardized approaches to

codifying behavior, SDOs are idiographic in comparison to rating scale systems, which are often norm-based, thus allowing SDOs to add unique information about a child to a multi-method behavioral assessment approach. On the other hand, teacher behavior ratings are useful precisely because they are norm-based and offer the benefit of comparing a specific child to relevant typical norms.

The current study's findings highlight the potential utility of combining a brief behavior screener with an SDO system to potentially identify students at behavioral and emotional risk. Alternatively, after screening processes have determined a risk is present, SDOs like the BOSS can be used in conjunction with more comprehensive measures like the BASC-3 to provide a detailed and accurate depiction of the student's behavior in the classroom. Both examples provide support for how SDOs like the BOSS can be used within an MTSS or RTI approach to screen, monitor, and evaluate the progress of students. Furthermore, a wide variety of school personnel (e.g., paraprofessionals, teachers, school psychologists) can serve as observers for the BOSS. Valid tools such as the BOSS should be considered more often by school practitioners when conducting behavior assessment for student referrals for disruptive behaviors (Shapiro & Heck, 2004).

4.4 | Limitations

There are limitations with the current study. First, this study was part of a larger RCT that was delivered primarily in elementary general education classrooms in a northeastern state in the United States. The current study did not control for or use hierarchical modeling to account for regional, school, grade level, and classroom effects. Therefore, results may not generalize to other grade levels (e.g., high school) or regions of the country. Second, observers were extensively trained in the BOSS and were independent to study condition and the school context. In practice, school personnel may not have as much competency in the BOSS and are generally more embedded (e.g., relationships with students that they are observing) in the school system compared with the independent observers employed in this study. Third, the teachers in the current study and the larger RCT were not blind to the students in the study. The students observed were initially selected for displaying disruptive behaviors by the classroom teachers as part of the larger RCT screening process. As such, teachers' knew in advance which students were to be observed with the BOSS and subsequently may have been biased in completing the BASC-3 TRS ratings. Fourth, the presence of independent observers in the classroom may have impacted teacher, paraprofessional, and student behavior and interactions in the classroom. Fifth, the BOSS used in this investigation was modified for a larger paraprofessional behavior coaching RCT. Specifically, the BOSS was modified to include disruptive, aggressive, and defiant behaviors, while the measurement of direct instruction time (or instructional type) was removed. This prevented the current study from exploring how classroom factors, such as lesson type or activity type, may have impacted disruptive behavior rates and their relationship with teacher-rated behaviors. While the BOSS is often modified depending on the research being conducted (e.g., Wood et al., 2016), it is possible that the results from this investigation might have been different had the instrument been unchanged. Finally, data analyzed in this study was from the beginning of the school year and did not capture whether the convergence between BOSS and teacher rating scales increased, decreased, or stayed constant throughout the year.

4.5 | Future research

Findings from this investigation offer avenues for future research. Specifically, replication studies are needed to further substantiate the BOSS's reliability and validity with different student populations, classroom settings (e.g., special education), grade levels (e.g., middle school), schools (e.g., private, public), and regions in the country. As the BOSS codes used vary across research projects, investigations should examine how specific combinations of codes impact its psychometric properties. Studies are needed to explore the convergence of the BOSS and other SDOs

with other prominent teacher rating scales such as the Social Skills Improvement System (Gresham & Elliot, 2008). Moreover, our study utilized independent observers, but often schools will have various school personnel collect SDO data (e.g., school psychologists, interventionists, paraprofessionals). Research should explore whether the type of observer (e.g., independent, teacher paraprofessional) influences the psychometrics of the BOSS.

4.6 | Conclusion

The current research examined the convergence between the BOSS and the BASC-3 TRS and BESS-3 TF for student with and at risk for DBDs in K-5 grade schools. Overall, findings from the current investigations provide evidence of convergent validity for the BOSS and utility of BOSS in measuring the classroom behaviors of elementary school students with or at risk of DBDs. The BOSS appears to be an efficient, specific, standardized SDO that school practitioners could use when conducting behavior assessment for student referrals for disruptive behaviors.

ACKNOWLEDGMENTS

This work was supported by the US Department of Education—Institute of Education Sciences NCSEER efficacy project (awarded to Rutgers, The State University of New Jersey) under Grant # R324A170069. The positions and opinions expressed in this article are solely those of the authors.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. Data used in the current study are available upon request to the corresponding author.

ETHICS STATEMENT

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

ORCID

Alexander Alperin  <https://orcid.org/0000-0002-5872-1076>

Christopher M. Dudek  <https://orcid.org/0000-0002-8338-4489>

Linda A. Reddy  <https://orcid.org/0000-0001-8314-2810>

Todd A. Glover  <https://orcid.org/0000-0001-7100-8139>

Nicole B. Wiggs  <https://orcid.org/0000-0002-5417-5864>

Briana Bronstein  <https://orcid.org/0000-0003-0040-3574>

REFERENCES

- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. University of Vermont, Research Center for Children, Youth, & Families.
- Allen, K. (2015). Externalizing disorders: Assessment, treatment, and school-based interventions. In R. Flanagan, K. Allen, & E. Levine (Eds.), *Cognitive and behavioral interventions in the schools: Integrating theory and research into practice* (pp. 161–180). Springer.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>

- Armbruster, P., Sukhodolsky, D., & Michalsen, R. (2004). The impact of managed care on children's outpatient treatment: A comparison study of treatment outcome before and after managed care. *American Journal of Orthopsychiatry*, 74(1), 5–13. <https://doi.org/10.1037/0002-9432.74.1.5>
- Austerman, J. (2015). ADHD and behavioral disorders: Assessment, management, and an update from DSM-5. *Cleveland Clinic Journal of Medicine*, 82(2–7). <https://doi.org/10.3949/ccjm.82.s1.01>
- Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology*, 72, 29–48.
- Bettini, E., Gilmour, A. F., Williams, T. O., & Billingsley, B. (2019). Predicting special and general educators' intent to continue teaching using conservation of resources theory. *Exceptional Children*, 86, 310–329. <https://doi.org/10.1177/0014402919870464>
- Bradshaw, C. P., Koth, C. W., Bevans, K. B., Jalongo, N., & Leaf, P. J. (2008). The impact of school-wide positive behavioral interventions and supports (PBIS) on the organizational health of elementary schools. *School Psychology Quarterly*, 23(4), 462–473. <https://doi.org/10.1037/a0012883>
- Briesch, A. M., & Daniels, B. (2013). Using self-management interventions to address general education behavioral needs: Assessment of effectiveness and feasibility. *Psychology in the Schools*, 50(4), 366–381. <https://doi.org/10.1002/pits.21679>
- Briesch, A. M., Volpe, R. J., & Floyd, R. G. (2018). *School-based observation: A practical guide to assessing student behavior*. Guilford Publications.
- Burns, G. L., Walsh, J. A., Gomez, R., & de Moura, M. A. (2003). Understanding source effects in ADHD rating scales: Reply to DuPaul (2003). *Psychological Assessment*, 15, 118–119. <https://doi.org/10.1037/1040-3590.15.1.118>
- Chafouleas, S. M., McDougal, J. L., Riley-Tillman, T. C., Panahon, C. J., & Hilt, A. M. (2005). What do daily behavior report cards (DBRCs) measure? An initial comparison of DBRCs with direct observation for off-task behavior. *Psychology in the Schools*, 42, 669–676.
- 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(4), 491–505.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Conners, C. K. (2008). *Conners* (3rd ed.). Multi-Health Systems.
- Cook, C. R., Collins, T., Dart, E., Vance, M. J., McIntosh, K., Grady, E. A., & DeCano, P. (2014). Evaluation of the class pass intervention for typically developing students with hypothesized escape-motivated disruptive classroom behavior. *Psychology in the Schools*, 51(2), 107–125. <https://doi.org/10.1002/pits.21742>
- Dart, E. H., Arora, P. G., Collins, T. A., & Doll, B. (2019). Progress monitoring measures for internalizing symptoms: A systematic review of the peer-reviewed literature. *School Mental Health*, 11(2), 265–275.
- Dart, E. H., Collins, T. A., Klingbeil, D. A., & McKinley, L. E. (2014). Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review*, 43(4), 367–384.
- DuPaul, G. J. (2003). Assessment of ADHD symptoms: Comment on Gomez et al. (2003). *Psychological Assessment*, 15, 115–117. <https://doi.org/10.1037/1040-3590.15.1.115>
- DuPaul, G. J., Volpe, R. J., Jitendra, A. K., Lutz, J. G., Lorah, K. S., & Gruber, R. (2004). Elementary school students with AD/HD: Predictors of academic achievement. *Journal of School Psychology*, 42(4), 285–301. <https://doi.org/10.1016/j.jsp.2004.05.001>
- Education Advisory Board (2019). *Breaking bad behavior: The rise of classroom disruptions in early grades and how districts are responding*. <http://pages.eab.com/rs/732-GKV-655/images/BreakingBadBehaviorStudy.pdf>
- Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *The California School Psychologist*, 14(1), 89–95. <https://doi.org/10.1007/bf03340954>
- Fabiano, G. A., Pelham, W. E., Manos, M. J., Gnagy, E. M., Chronis, A. M., Onyango, A. N., Lopez-Williams, A., Burrows-MacLean, L., Coles, E. K., Meichenbaum, D. L., Caserta, D. A., & Swain, S. (2004). An evaluation of three time-out procedures for children with attention-deficit/hyperactivity disorder. *Behavior Therapy*, 35, 449–469. [https://doi.org/10.1016/S0005-7894\(04\)80027-3](https://doi.org/10.1016/S0005-7894(04)80027-3)
- Goldstein, S., & Naglieri, J. A. (2010). *Autism Spectrum Rating*. Multi-Health Systems Inc.
- Green, T. D. (2005). Promising prevention and early intervention strategies to reduce overrepresentation of African American students in special education. *Preventing School Failure*, 49, 33–41.
- Greenwood, C. R. (1996). The case for performance-based instructional models. *School Psychology Quarterly*, 11(4), 283–296.
- Gresham, F. M., & Elliot, S. N. (2008). *Social skills improvement system (SSIS): Rating scales manual*. PsychoCorp.
- Handler, M. W., & DuPaul, G. J. (2005). Assessment of ADHD: Differences across psychology specialty areas. *Journal of Attention Disorders*, 9(2), 402–412.

- Hinshaw, S. P. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin*, 111(1), 127–155. <https://doi.org/10.1037/0033-2909.111.1.127>
- Hojnoski, R. L., Margulies, A. S., Barry, A., Bose-Deakins, J., Sumara, K. M., & Harman, J. L. (2008). Analysis of two early childhood education settings: Classroom variables and peer verbal interaction. *Journal of Research in Childhood Education*, 23(2), 193–209.
- Hojnoski, R. L., Missall, K. N., & Wood, B. K. (2020). Measuring engagement in early education: Preliminary evidence for the behavioral observation of students in schools–early education. *Assessment for Effective Intervention*, 45(4), 243–254. <https://doi.org/10.1177/1534508418820125>
- Hosp, J. L., Howell, K. W., & Hosp, M. K. (2003). Characteristics of behavior rating scales. *Journal of Positive Behavior Interventions*, 5(4), 201–208. <https://doi.org/10.1177/10983007030050040301>
- Hosterman, S. J., DuPaul, G. J., & Jitendra, A. K. (2008). Teacher ratings of ADHD symptoms in ethnic minority students: Bias or behavioral difference? *School Psychology Quarterly*, 23(3), 418–435. <https://doi.org/10.1037/a0012668>
- Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).
- Jiang, Y., Capriotti, M., Beaulieu, A., Rooney, M., McBurnett, K., & Pfiffner, L. J. (2019). Contribution of the behavioral observation of students in schools to ADHD assessment. *School Mental Health*, 11(3), 464–475. <https://doi.org/10.1007/s12310-019-09313-5>
- Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2 behavioral and emotional screening system manual*. Pearson.
- Kamphaus, R. W., & Reynolds, C. R. (2015). *Behavior Assessment System for Children–Third Edition (BASC-3): Behavioral and emotional screening system (BESS)*. Pearson Assessment.
- Kellner, M. H., Bry, B. H., & Salvador, D. S. (2008). Anger management effects on middle school students with emotional/behavioral disorders: Anger log use, aggressive and prosocial behavior. *Child & Family Behavior Therapy*, 30(3), 215–230.
- Kendall, M. G., & Stuart, A. (1958). *The advanced theory of statistics*. Griffin.
- McConaughy, S. H., & Ritter, D. R. (2002). Best practices in multidimensional assessment of emotional or behavioral disorders. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1303–1320). National Association of School Psychologists.
- Merrell, K. W. (1993). Using behavior rating scales to assess social skills and antisocial behavior in school settings: Development of the school social behavior scales. *School Psychology Review*, 22(1), 115–133. <https://doi.org/10.1080/02796015.1993.12085641>
- Minder, F., Zuberer, A., Brandeis, D., & Drechsler, R. (2018). A review of the clinical utility of systematic behavioral observations in attention deficit hyperactivity disorder (ADHD). *Child Psychiatry & Human Development*, 49, 572–606. <https://doi.org/10.1007/s10578-017-0776-2>
- Missall, K. N. (2002). Reconceptualizing school adjustment: A search for intervening variables (Doctoral Dissertation, University of Minnesota, 2002). *Dissertation Abstracts International*, 63(5-A), 1712.
- Moffett, L., & Morrison, F. J. (2020). Off-task behavior in kindergarten: Relations to executive function and academic achievement. *Journal of Educational Psychology*, 112(5), 938–955. <https://doi.org/10.1037/edu0000397>
- Muratori, P., Milone, A., Manfredi, A., Polidori, L., Ruglioni, L., Lambruschi, F., Masi, G., & Lochman, J. E. (2017). Evaluation of improvement in externalizing behaviors and callous-unemotional traits in children with disruptive behavior disorder: A 1-year follow up clinic-based study. *Administration and Policy in Mental Health and Mental Health Services Research*, 44(4), 452–462. <https://doi.org/10.1007/s10488-015-0660-y>
- Ota, K. R., & DuPaul, G. J. (2002). Task engagement and mathematics performance in children with attention-deficit hyperactivity disorder: Effects of supplemental computer instruction. *School Psychology Quarterly*, 17(3), 242–257. <https://doi.org/10.1521/scpq.17.3.242.20881>
- Pikard, J., Roberts, N., & Groll, D. (2018). Pediatric referrals for urgent psychiatric consultation: Clinical characteristics, diagnoses and outcome of 4 to 12 year old children. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 27(4), 245–251.
- Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavioral Assessment Scale for Children* (2nd ed.). Pearson Assessments.
- 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 data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions*, 10(2), 136–143.
- Rocque, M. (2010). Office discipline and student behavior: Does race matter? *American Journal of Education*, 116(4), 557–581.
- Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment: In special and inclusive education* (11th ed.). Wadsworth.
- Shapiro, E. S. (2004). *Direct observation: Manual for the behavioral observation of students in schools (BOSS)*. Pearson.
- Shapiro, E. S. (2011). *Academic skills problems* (4th ed.). Guilford.

- Shapiro, E. S., & Heck, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, 41(5), 551–561. <https://doi.org/10.1002/pits.10176>
- Sheridan, S. M., Bovaird, J. A., Glover, T. A., Garbacz, S. A., Witte, A., & Kwon, K. (2012). A randomized trial examining the effects of conjoint behavioral consultation and the mediating role of the parent–teacher relationship. *School Psychology Review*, 41(1), 23–46. <https://doi.org/10.1080/02796015.2012.12087374>
- Simonsen, B., Myers, D., & Briere, III D. E. (2011). Comparing a behavioral check-in/check-out (CICO) intervention to standard practice in an urban middle school setting using an experimental group design. *Journal of Positive Behavior Interventions*, 13(1), 31–48.
- 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(1), 83–93.
- Staff, A. I., Oosterlaan, J., Van der Oord, S., Hoekstra, P. J., Vertessen, K., de Vries, R., van den Hoofdakker, B. J., & Luman, M. (2021). The validity of teacher rating scales for the assessment of ADHD symptoms in the classroom: A systematic review and meta-analysis. *Journal of Attention Disorders*, 25(11), 1578–1593. <https://doi.org/10.1177/1087054720916839>
- Steege, M. W., Davin, T., & Hathaway, M. (2001). Reliability and accuracy of a performance-based behavioral recording procedure. *School Psychology Review*, 30(2), 252–261.
- Steiner, N. J., Sidhu, T., Rene, K., Tomasetti, K., Frenette, E., & Brennan, R. T. (2013). Development and testing of a direct observation code training protocol for elementary aged students with attention deficit/hyperactivity disorder. *Educational Assessment, Evaluation and Accountability*, 25(4), 281–302. <https://doi.org/10.1007/s11092-013-9166-x>
- Walker, H. M., Ramsey, E., & Gresham, F. M. (2004). *Antisocial behavior in school: Evidence-based practices*. Wadsworth.
- Williford, A. P., Wolcott, C. S., Whittaker, J. V., & Locasale-Crouch, J. (2015). Program and teacher characteristics predicting the implementation of Banking Time with preschoolers who display disruptive behaviors. *Prevention Science*, 16, 1054–1063.
- Wood, B. K., Hojnoski, R. L., Laracy, S. D., & Olson, C. L. (2016). Comparison of observational methods and their relation to ratings of engagement in young children. *Topics in Early Childhood Special Education*, 35(4), 211–222. <https://doi.org/10.1177/0271121414565911>

How to cite this article: Alperin, A., Dudek, C. M., Reddy, L. A., Glover, T. A., Wiggs, N. B., & Bronstein, B. (2023). Convergent validity of the Behavior Observation of Students in Schools for elementary school students with disruptive behaviors. *Psychology in the Schools*, 60, 4039–4060. <https://doi.org/10.1002/pits.22983>