Journal of Behavioral Education
https://doi.org/10.1007/s10864-021-09458-x

ORIGINAL PAPER

Psychometric Properties of the Behavior Assessment System for Children Student Observation System (BASC-3 SOS) with Young Children in Special Education

Ellyn M. Schmidt[1] · W. Andrew Rothenberg[1,2] · Bridget C. Davidson[1,3] · Miya Barnett[4] · Jason Jent[1] · Heleny Cadenas[1] · Corina Fernandez[1] · Eileen Davis[1]

Accepted: 13 October 2021

Abstract

Measuring classroom behavior among young children is important to guide assessment and intervention decisions, yet there is limited literature on appropriate direct observation tools for this purpose. This article describes the psychometric properties of the Behavior Assessment System for Children, Student Observation System (BASC-3 SOS) with 135 children ages 20 to 67 months ($M = 35$ months, 64% Latinx, 78% with an established developmental disability) and their teachers ($N = 36$) as part of a larger randomized control trial of a teacher training intervention. Inter-rater reliability on individual BASC-3 SOS behaviors ranged from poor to good. Correlations between BASC-3 SOS scores across time indicated low to moderate developmental test–retest reliability. Significant correlations between BASC-3 SOS scores and teacher ratings provided evidence for convergent, divergent, and predictive validity. Differences between BASC-3 SOS scores for children with versus without disabilities supported the tool's discriminant validity. There were no significant pre- to post-treatment changes in BASC-3 SOS scores. Overall, results provide mixed evidence for the psychometric properties of the BASC-3 SOS when used with young, diverse children with and without disabilities. Implications for clinical and research purposes are discussed.

Keywords Systematic direct observation · Classroom observation tool · Developmental disabilities · Early special education · Teacher–Child Interaction Training

✉ Ellyn M. Schmidt
exs1247@med.miami.edu

Extended author information available on the last page of the article

Published online: 01 November 2021

Introduction

Child externalizing behavior problems can emerge as early as toddlerhood (Carter et al., 2003) and are a common concern in early childhood education settings. For example, teacher ratings in Head Start classrooms indicated that externalizing behaviors, such as hyperactivity, impulsive behavior, and physical aggression, were among the most commonly observed problems (Cai et al., 2004). Additionally, rates of behavior problems are higher among children with developmental delays than among typically developing children (Emerson & Einfeld, 2010). Clinicians, researchers, and school staff need tools for validly and reliably measuring disruptive behavior in young children, including those with delays and disabilities. Such tools are important for identifying children in need of intervention, determining appropriate treatment targets, and monitoring children's response to interventions. Higher externalizing behaviors predict lower active participation in school activities and lower attendance (Olivier et al., 2020), thus well-planned and carefully monitored interventions to decrease specific problematic behaviors can have far-reaching impacts for children's engagement in school activities.

One option for measuring child behavior in early education settings is teacher interviews, which yield summary information about problem behaviors from a teacher's perspective, though this information is subjective and not quantifiable. Teacher rating scales that represent constructs such as hyperactivity, inattention, and aggression (e.g., Behavior Assessment System for Children–Third Edition [BASC-3; Reynolds & Kamphaus, 2015], Sutter–Eyberg Student Behavior Inventory, Revised [SESBI-R; Eyberg & Pincus, 1999]) are another option. A strength of rating scales is strong evidence for reliability and validity of many scales (e.g., SESBI-R: Querido & Eyberg, 2003), while an important limitation is that they represent a teacher's perspective and thus may be biased (Briesch et al., 2018). Another approach is direct behavior ratings (DBR) of specific target behaviors over brief time periods, such as a classroom activity. DBRs are simple and brief, and they are highly flexible in terms of the behaviors rated, the type of scale (e.g., yes/no, Likert scale), and the interval of time over which behavior is rated. However, DBRs have lower interobserver agreement than other forms of behavior measurement, and as with rating scales, DBRs are impacted by teacher bias (Briesch et al., 2010). Further, because DBRs typically include only a few behaviors, they may be more appropriate for measuring behaviors of an individual child rather than across groups of children.

Clinicians, researchers, and school staff can also directly observe children. Narrative observations that describe what the observer sees yield detailed information but are prone to overinterpretation and confirmation bias (Hintze et al., 2002). A more objective and structured approach is systematic direct observation (SDO), in which specific behaviors that have been operationally defined are measured using standardized procedures during predetermined places and times (Salvia & Ysseldyke, 2004; Hintze et al., 2002). Importantly, SDOs yield scores based on the observed behaviors throughout the observation, with scores expected to be the same across observers (Salvia & Ysseldyke, 2004; Hintze et al., 2002). SDOs are typically conducted over 10–30 min and can quantify behavior in terms of features such as frequency,

duration, and time sampling interval recording (e.g., partial interval, in which the behavior is coded if it occurs at least once in the interval; whole interval, in which the behavior is coded only if it occurs for the duration of the interval; and momentary time sampling, in which the behavior is coded if it occurs at a designated point in the interval). While training to use SDOs can be time consuming and SDOs only capture a brief sample of behavior, SDO data are more objective than teacher ratings, DBRs, or narrative observations (Briesch et al., 2018).

There are a variety of SDOs for measuring classroom behavior of school-age children, some of which have strong psychometric properties (see Briesch et al., 2018 for review). However, there are far fewer SDOs developed or recommended for children 5 years and younger. Using tools designed only for school-age children is not recommended given that the behaviors, tasks, and adult–child interactions relevant to early education contexts are not necessarily consistent with those relevant in later grades (Bramlett & Barnett, 1993; Hojnoski et al., 2020). For example, appropriate behavior in group activities in elementary settings typically involves remaining seated and speaking only when called on, whereas appropriate behavior during group activities in preschools often includes standing, dancing, and choral responding.

Given the aforementioned differences between elementary and early education contexts, some researchers have designed and used SDOs specific to young children. First, the Preschool Observation Code (POC; Bramlett & Barnett, 1993) is a 10-min SDO where behaviors including disruptive behavior (e.g., throwing objects), child compliance, and play engagement (i.e., orientation toward activity) are marked using momentary time sampling and frequency counts. To our knowledge, there is only one published study examining the psychometrics of the POC, which reported only mixed evidence for interobserver agreement and did not report demographics of their sample (Bramlett & Barnett, 1993). Second, the Revised School Observation Coding System (REDSOCS; Ginn et al., 2009) uses a 10-min observation to record six behaviors: inappropriate behavior (e.g., aggression), noncompliance, off-task behavior (e.g., out of seat), appropriate behavior, compliance, and on-task behavior. Three studies have focused on the psychometrics of the REDSOCS with primarily White samples of preschoolers. Within these studies, the REDSOCS demonstrates evidence for convergent, divergent, and discriminant validity, but limited and inconsistent evidence for treatment sensitivity, variable estimates of interobserver agreement, and no evidence for test–retest reliability (Bagner et al., 2010; Fawley et al., 2020; Jacobs et al., 2000). A third SDO for young children is the Behavioral Observation of Students in Schools-Early Childhood (BOSS-EE; Hojnoski et al., 2020), adapted from the BOSS school-age tool. The BOSS-EE involves a 10–20-min observation with momentary time sampling used to code active and passive engagement and partial interval recording used to code interfering behaviors (motor, verbal, passive). There is only one study of BOSS-EE, which provided support for content and concurrent validity using a predominantly English-speaking, 40% White, 33% special education sample of 43 children (Hojnoski et al., 2020). However, evidence for interobserver agreement and test–retest reliability was mixed, and predictive validity and treatment sensitivity were not explored.

A final SDO for children from 2 to 21 years, is the BASC-3, Student Observation System (BASC-3 SOS; Reynolds & Kamphaus, 2015). This tool includes a Likert scale rating of the frequency of each behavior across the 15-min observation (Part A) and a momentary time sampling method for recording occurrences of behaviors at the end of 30-s intervals (Part B). Eleven Problem Behaviors (e.g., Inappropriate Movement) and four Adaptive Behaviors (e.g., Response to Teacher/Lesson) are coded. While there are no published studies on psychometrics of the BASC-3 SOS, two studies indicate good interobserver agreement (Pearson correlation coefficients 0.69–1.0; percentage of agreement 0.81–1.0) for the BASC-2 SOS Part B (Lett & Kamphaus, 1997; Margiano et al., 2009), which is the same as the BASC-3 SOS except that it lacks the Inappropriate Interactions category. One of these studies included three male fourth-grade students who were Hispanic, Asian, and White (Margiano et al., 2009), while the other study included a primarily male and White sample of children with a mean age of approximately 8 years (Lett & Kamphaus, 1997). Regarding discriminant validity, BASC-2 SOS Part B scores distinguished children with versus without attention-deficit/hyperactivity disorder (ADHD; Lett & Kamphaus, 1997). There is also evidence for treatment sensitivity of the BASC-2 SOS (Margiano et al., 2009). The only psychometric information for the BASC-3 SOS comes from the manual, which reported fair to good interobserver agreement for Part A (Fleiss Kappa 0.44–1.0), with stronger reliability reported for Part B (Fleiss Kappa 0.62–1.0) when multiple observers rated a single fourth-grade male student whose race/ethnicity was not reported (Reynolds & Kamphaus, 2015). No information has been reported on the test–retest reliability or convergent, divergent, or predictive validity of this tool, and there is no published data on the psychometrics of this tool for early childhood populations or children diverse in terms of race/ethnicity.

The SDO literature emphasizes that tools appropriate for research must have evidence for interobserver agreement, test–retest reliability, treatment sensitivity, and content, concurrent/convergent, and predictive validity (Hintze, 2005). Clearly, further research is necessary for any SDO for young children to be deemed evidence based. Further, there is limited research on any of these SDOs with diverse samples or children receiving special education. Of the available SDOs, there are several benefits of the BASC-3 SOS. First, it measures a wider variety of problem behaviors, yielding more detailed information that can better inform intervention planning. Second, this tool codes multiple adaptive behaviors not captured by other SDOs, such as positive peer interactions and transition movements (i.e., moving appropriately between activities). Scores thus represent a fuller picture of both problem behaviors and strengths. Third, the BASC-3 SOS includes options for Likert ratings (Part A) and momentary time sampling (Part B) of the same behaviors whereas other tools select one approach. Finally, the BASC-3 SOS was designed for use in combination with other components of the BASC-3 (e.g., rating scales), which may make the full BASC-3 system a desirable choice for practitioners and researchers alike. However, the SDO literature currently recommends only using the BASC-3 SOS for descriptive purposes rather than for assessment, progress monitoring, or intervention evaluation given limited psychometric study of this tool (Briesch et al., 2018).

Current Study

Previous investigations of the BASC-3 SOS, as with many other SDO tools, have not focused on children diverse in terms of race/ethnicity and disability. Using tools with diverse populations when their psychometric properties have not been established in these populations is problematic given that there may be cultural and context-specific expectations of child behavior that impact whether scores from the tool are meaningful (Bulotsky-Shearer et al., 2013; Salvia & Ysseldyke, 1991). Thus, further investigation into the psychometrics of the BASC-3 SOS in samples of young children from racial/ethnic minority backgrounds who have disabilities is necessary. Accordingly, this project explored the extent to which the BASC-3 SOS captures variability in classroom behavior of young, primarily Latinx children in a special education setting, and also explored multiple psychometric properties of this tool. Establishing such psychometric evidence would allow clinicians, researchers, and school staff to use this tool as part of assessment and intervention evaluation processes with confidence in the data that the tool provides. We collected BASC-3 SOS data in the context of a larger randomized control trial (RCT) that evaluated the impact of a teacher behavior management intervention. We hypothesized that individual Problem Behaviors would occur relatively infrequently given findings of other SDOs (Hojnoski et al., 2020; Jacobs et al., 2000). We also hypothesized that Adaptive Behavior would be more frequent in small than large group settings and Problem Behavior would be more frequent in large than small group settings based on findings with the BOSS-EE (Hojnoski et al., 2020).

We hypothesized moderate to good inter-rater reliability, given the operational behavior definitions provided by the tool paired with our systematic training of coders. Regarding test–retest reliability, we hypothesized weak to moderate and statistically significant correlations between behavior scores across time points for children who did not receive a behavioral intervention, based on findings of test–retest reliability for a similar SDO (Hojnoski et al., 2020).

In terms of convergent validity, we hypothesized positive, moderate, and statistically significant correlations between BASC-3 SOS Problem and Adaptive Behaviors and teacher reports of behaviors falling into these categories at the same time point. In the area of divergent validity, we hypothesized a negative correlation between BASC-3 SOS Problem Behavior scores and teacher-reported social–emotional strength, given that these tools measure opposite constructs. Similarly, we hypothesized negative correlations between BASC-3 SOS Adaptive Behaviors and teacher-reported problem behavior. In terms of predictive validity, we hypothesized positive, moderate, and statistically significant correlations between Adaptive and Problem Behavior scores at our first time point with teacher reports of behaviors falling into these categories at our third time point.

We also hypothesized that Adaptive Behavior scores would be higher and Problem Behavior scores would be lower among children without disabilities compared to children with disabilities, which would indicate discriminant validity of the BASC-3 SOS. Finally, with regard to treatment sensitivity, we explored changes in BASC-3 SOS Adaptive and Problem Behavior scores following participation in universal Teacher–Child Interaction Training (TCIT-U), a teacher-coaching

intervention that focuses on teachers' use of positive interaction and effective behavior management strategies with their students. Teachers participated in in-person didactic workshops (i.e., lecture, discussion, worksheets, live and video demonstrations, and role-plays) to learn these strategies and subsequently participated in live coaching in their classroom to receive feedback and support to effectively implement these strategies. Coaching occurred approximately twice per week for 14–16 weeks (see Davidson et al., (2021) for further description of TCIT-U implementation). Two RCTs have demonstrated improvements in child classroom behavior from pre to post TCIT-U implementation according to teacher ratings (Budd et al., 2016; Davidson et al., 2021); thus, we hypothesized that BASC-3 SOS Problem scores would decrease and Adaptive Behavior scores would increase from before to after TCIT-U implementation in the intervention group.

Method

Participants

Data come from a larger RCT with a waitlist control group to investigate outcomes of TCIT-U in early special education programs (12 classrooms across 3 schools) in a large, southeastern United States metropolitan area (Davis et al., 2021). There were no exclusion criteria for teachers or students. For example, elevated problem behavior was not required. Teachers ($N=36$) were all female and predominately Latinx (89%). Of the 135 students in the sample, 67% were male and 64% were Latinx. In terms of the intersection between race and ethnicity, 11% were Non-Latinx White, 46% were Latinx White, 15% Non-Latinx Black/African American, 5% Latinx Black/African American, 7% Non-Latinx Asian, 1% Latinx Asian, 2% Non-Latinx Other, and 13% Latinx Other. The mean age was 35 months (range 20–67 months, SD = 9.4) and 78% had an established disability. Specifically, 22% of children had a diagnosis of Autism Spectrum Disorder, 11% of children had a hearing impairment, and the remaining 34% had an unspecified area of delay/disability. Children's primary language was 25% English, 25% Spanish, 19% English and Spanish, 2% English and Creole, 3% other, and 27% unknown. Students were receiving services as usual within their classroom as well as individualized services through Individuals with Disabilities Education Act (IDEA) Part C, such as specific teaching practices (e.g., visual supports) and/or specific school-based services (e.g., speech and language therapy, occupational therapy, Behavior Intervention Plans). Fifteen percent of children had an elevated SESBI-R score at Time 1. Teachers and parents of students provided informed consent.

Measures

BASC-3 SOS

The BASC-3 SOS is a systematic direct observation tool designed to capture a variety of classroom behaviors (Reynolds & Kamphaus, 2015). This tool includes two procedures for measuring child behavior (Parts A and B). Using the Likert scale

procedure in Part A, the observer rates the frequency with which each behavior occurred across the full observation period, with response options of *never observed, sometimes observed*, or *frequently observed*. Using the momentary time sampling procedure in Part B, the observer records whether or not each behavior occurred during the last 3 s of every 30-s interval over the course of a 15-min observation period. Data from only Part B of the BASC-3 SOS were analyzed for the current study because (a) the Part B time sampling procedure provides a wider range of possible scores, (b) stronger inter-rater reliability has been documented with Part B as opposed to Part A (Reynolds & Kamphaus, 2015), and (c) previous research studies utilizing the BASC-3 SOS have used Part B (Lett & Kamphaus, 1997; Margiano et al., 2009).

The BASC-3 SOS includes four Adaptive Behaviors: (a) Response to Teacher/Lesson (e.g., answers teacher's questions appropriately), (b) Peer Interaction (e.g., talking appropriately with peers), (c) Work on School Subjects (e.g., independent engagement in an activity), and (d) Transition Movement (e.g., lining up to leave the classroom), and eleven Problem Behaviors: (a) Inappropriate Interactions (e.g., distracting others by imposing on their personal space), (b) Inappropriate Movement (e.g., fidgeting in seat), (c) Inattention (e.g., looking at objects unrelated to activity), (d) Inappropriate Vocalization (e.g., making disruptive noises), (e) Somatization (e.g., complaining of stomach ache), (f) Repetitive Motor Movement (e.g., rocking back and forth), (g) Aggression (e.g., hitting), (h) Self-Injurious Behavior (e.g., banging head on hard surface), (i) Inappropriate Sexual Behavior (e.g., touching others inappropriately), (j) Bowel/Bladder Problems (e.g., urinating in clothing), and (k) Other (i.e., behaviors not captured in the other categories). Only Response to Teacher and Work on School Subjects are mutually exclusive codes. Because some examples accompanying behavior definitions provided in the BASC-3 SOS manual are more applicable to school-age children than to early childhood populations, the research team maintained a document of examples of child behaviors often observed in this sample that fit within each BASC-3 SOS category (see supplemental material). For example, for preschoolers, additional behaviors that constitute examples of Work on School Subjects include independent engagement in play at a sand table or eating a snack independently.

Though 15-min observations were attempted, some observations were of shorter duration due to activities ending or children leaving the classroom. Thus, the percentage of observed intervals during which each behavior occurred was the variable used for analyses. For example, if a child was observed for 13 min, this observation would contain 26 intervals rather than 30. Thus, if this child engaged in aggression during 5 observed intervals, the child's score for Aggression would be 19.2% (i.e., 5 divided by 26). We calculated composite Adaptive and Problem Behavior scores by summing the percentage of intervals during which each behavior belonging to these categories occurred (e.g., a child whose Problem Behaviors were Inattention in 10% of intervals and Aggression in 5% of intervals would have a Problem Behavior composite of 15%). This procedure limits the number of variables in analyses, thereby reducing multicollinearity, and is described in the BASC-3 SOS manual as one way to interpret overall child behaviors.

Sutter–Eyberg Student Behavior Inventory—Revised (SESBI-R)

The SESBI-R is a 38-item questionnaire assessing maladaptive behaviors (e.g., non-compliance) in 2–16-year-olds observed by the teacher at school during the past week (Eyberg & Pincus, 1999). The Intensity Scale assesses the frequency with which each behavior was observed, with a 7-point scale ranging from (1) *Never* to (7) *Always*. This scale yields a total raw score, converted to normed *T* scores with scores ≥ 60 indicating clinically significant behavior concerns. The SESBI-R has demonstrated high test–retest reliability and treatment sensitivity, in addition to high discriminant and predictive validity (Querido & Eyberg, 2003). Internal consistency for Intensity items in the current sample was high (baseline $\alpha = 0.96$, post $\alpha = 0.97$, follow-up $\alpha = 0.97$). The Intensity *T*-scores were used in the current study.

Devereux Early Childhood Assessment, Second Edition (DECA)

The DECA is a teacher rating scale that measures within-child protective factors related to resilience. In this study, the 36-item Toddlers Record Form (MacKrain, LeBuffe, & Powell, 2007) was used for students ages 18–36 months while the 38-item Preschoolers Second Edition (LeBuffe & Naglieri, 2012) was used for students ages 3–5 years. Both yield a Total Protective Factor score (TPF) comprised of three subscales which assess students' social–emotional strengths: Initiative (i.e., child's ability to independently meet needs through thoughts and actions), Self-Control (i.e., child's ability to use healthy strategies to express emotions and manage behavior), and Attachment (i.e., child's ability to initiate and maintain positive relationships). Teachers rated the frequency of the observed behavior during the past four weeks using a five-point Likert scale. The DECA was nationally normed and the TPF demonstrated good reliability and validity in a diverse sample of young children (Bulotsky-Shearer et al., 2013). The TPF *T*-Scores were used in this study.

Procedure

All procedures were approved by the university Institutional Review Board. The trial was registered at clinicaltrials.gov (NCT04000230). A baccalaureate research associate (not masked to allocation) coordinated assessments at baseline (i.e., Time 1), post-training (i.e., Time 2: 14–16 weeks after baseline following implementation of TCIT-U in intervention classrooms), and follow-up (i.e., Time 3: 4–6 weeks after TCIT-U ended), which involved distributing hard copies of the SESBI-R and DECA, and filming videos for BASC-3 SOS coding. There were six children ages 20–23 months, thus below the age range for SESBI-R and BASC-3 SOS. Data were analyzed for these children given that they were in the same classrooms with similar behavioral expectations as their slightly older peers. Lead teachers received gift cards ($100) after completing questionnaires for their participating students at each time point.

BASC-3 SOS

The research associate filmed 15-min videos of each child at each time point of the study. Five undergraduate research assistants and a doctoral graduate student masked to group (intervention vs. waitlist) and time (pretreatment vs. post-treatment vs. follow-up) coded these videos in randomized order using the BASC-3 SOS initially in Noldus The Observer XT 14, a software program to record behavioral observation data (Noldus Information Technology, 2017) and subsequently in Excel due to the necessity of working remotely without access to the Noldus software due to COVID-19. Importantly, coding in Excel followed the same procedure and yielded the same data as the original coding approach in Noldus. Coders were trained by reviewing the operational definitions of BASC-3 SOS behaviors and examples of these behaviors, observing a trained coder to learn the coding procedure, and then coding practice videos prior to coding a reliability set of three videos. New coders were required to meet $\geq 80\%$ inter-rater agreement (as measured by coefficient kappa) with a trained BASC-3 SOS coder on each behavior before coding participant videos. A working document of examples of each behavior was maintained throughout the coding process to promote consistent coding of situations that occurred in early childhood classrooms (see supplemental material).

Results

BASC-3 SOS Descriptive Statistics

Means, standard deviations, ranges, medians, and skew for each behavior as well as the composite Adaptive and Problem Behavior Scores indicated the extent to which the BASC-3 SOS captures variability in behavior in young children with and without disabilities. Scores for both the intervention and waitlist groups at baseline (i.e., Time 1) were examined (see Table 1). Children frequently demonstrated Adaptive Behaviors, with Response to Teacher/Lesson or Work on School Subjects occurring in an average of 74% of observed intervals. All Problem Behaviors demonstrated positive skew, as each behavior was infrequently observed, with Inattention (25.4% of intervals) and Inappropriate Movement (20.5% of intervals) being the most commonly observed. Table 2 displays composite Adaptive and Problem Behavior mean scores in large group (e.g., circle time or morning meeting) and small group (e.g., sand table or pretend play area with a subgroup of children) activities. Mean Problem Behavior Composite Scores were significantly higher in large ($M = 54.1$, $SD = 5.7$) than small group settings ($M = 34.2$, $SD = 6.3$; $t(54) = 2.34$, $p = 0.023$), while mean Adaptive Behavior Composite Scores were marginally, but not significantly higher in small ($M = 84.2$, $SD = 3.5$) than large group settings ($M = 74.7$, $SD = 3.9$; $t(54) = -1.77$, $p = 0.082$).

Table 1 Descriptive statistics for individual behaviors and composite scores from BASC-3 SOS

Behavior	Mean (SD)[a]	Range[a]	Median[a]	Skew[a]	Interobserver agreement[b, c] (95% CI)	Test–retest times 1–2[d]	Test–retest times 2–3[e]	Test–retest times 1–3[f]
Adaptive Behavior Composite	76.8 (21.6)	10–113	80.0	−0.83		0.52**	.28*	.32*
Response to Teacher/Lesson	62.0 (26.1)	0–100	66.7	−0.4	0.60 (0.43–0.73)			
Work on School Subjects	12.0 (19.9)	0–87	0.0	2.0	0.82 (0.72–0.88)			
Transition Movement	2.4 (5.2)	0–30	0.0	3.3	0.26 (0.03–0.47)			
Peer Interaction	2.7 (4.6)	0–31	0.0	2.9	0.51 (0.32–0.67)			
Problem Behavior Composite	50.3 (37.8)	0–173	40.0	0.81		0.32*	−.01	.32*
Inappropriate Movement	20.5 (21.2)	0–93	13.3	1.2	0.68 (0.53–0.79)			
Aggression	0.3 (1.3)	0–10	0.0	5.3	0.28 (0.05–0.48)			
Inattention	25.4 (19.9)	0–87	20.0	1.0	0.73 (0.59–0.82)			
Inappropriate Peer Interactions	2.3 (4.4)	0–27	0.0	3.0	0.00 (−0.23–0.23)			
Inappropriate Vocalization	2.5 (5.5)	0–31	0.0	3.2	0.64 (0.47–0.76)			
Self-Injurious Behavior	1.4 (5.5)	0–37	0.0	4.9	0.69 (0.54–0.80)			
Repetitive Motor Movement	4.6 (7.8)	0–53	0.0	3.3	0.32 (0.09–0.52)			
Somatization	0.8 (2.2)	0–10	0.0	2.9	0.14 (−0.10–0.36)			
Inappropriate Sexual Behavior	0.1 (0.5)	0–3	0.0	5.8	NA[g]			
Bowel/Bladder Problem	0.1 (0.5)	0–3	0.0	5.8	NA[g]			
Other	0.5 (1.5)	0–10	0.0	3.9	0.58 (0.40–0.72)			

SD standard deviation, *CI* confidence interval, *NA* not applicable

*$p<0.05$; ** $p<0.01$

[a]$n=107$, scores represent percentage of intervals during which behaviors occurred; [b]$n=70$; [c]Interobserver agreement measured by ICC; [d]$n=49$; [e]$n=53$; [f]$n=49$; [g]Not applicable because behavior was not observed in subset of observations that was used to calculate interobserver agreement

Table 2 Psychometric properties of BASC-3 SOS based on group size

	Adaptive behavior composite[a]		Problem behavior composite[a]	
	Large group	Small group	Large group	Small group
Mean (SD)[b]	74.4 (21.7)	84.2 (17.2)	54.1 (31.9)	34.2 (31.5)
Developmental test–retest				
Time 1–Time 2	0.91* ($n=5$)	0.07 ($n=6$)	0.27 ($n=5$)	0.50 ($n=6$)
Time 2–Time 3	0.58^ ($n=10$)	−0.49 ($n=5$)	0.42 ($n=10$)	0.45 ($n=5$)
Time 1–Time 3	0.68* ($n=9$)	−0.13 ($n=6$)	0.53 ($n=9$)	−0.41 ($n=6$)
Convergent and divergent validity				
SESBI	−0.48** ($n=31$)	−0.21 ($n=31$)	0.25 ($n=31$)	0.36^ ($n=31$)
DECA	0.48** ($n=31$)	0.26 ($n=31$)	−0.18 ($n=31$)	−0.47* ($n=31$)

^$p<0.10$; *$p<0.05$; **$p<0.01$

[a]Scores represent percentage of intervals during which behaviors occurred

[b]Values pertain to Time 1 data

Inter-Rater Reliability

Twenty percent of BASC-3 SOS observations were randomly selected to be double coded to assess inter-rater reliability. We calculated intraclass correlation coefficients (ICCs) for each individual behavior using the two-way random effects, single-rater, consistency ICC (Koo & Li, 2016). ICCs were interpreted as: < 0.5 = poor reliability, 0.50.75 = moderate reliability, 0.75–0.9 = good reliability, and > 0.9 = excellent reliability (Koo & Li, 2016). ICCs for each behavior are displayed in Table 1 and indicate moderate to good reliability for all behaviors with the exceptions of Transition Movement, Inappropriate Interaction, Somatization, Repetitive Motor Movement, and Aggression. Additionally, Inappropriate Sexual Behavior and Bowel/Bladder Problems were never observed. Thus, these 7 behaviors were excluded from calculations of Composite Adaptive and Problem Behavior Scores.

Developmental Test–Retest Reliability

Correlations between Composite Adaptive and Problem Behavior Scores for the waitlist group between time points indicated test–retest reliability. Correlations were interpreted as: $r<0.3$ = weak, $0.3>r<0.5$ = moderate, and $r>0.5$ = strong (Cohen, 1988). Correlations of Adaptive Behavior Composites were significant between all time points (Time 1–2 $r=0.52$, $p<0.01$, $n=49$); Times 2–3 $r=0.28$, $p<0.05$, $n=53$; Times 1–3 $r=0.32$, $p<0.05$, $n=49$), while correlations of Problem Behavior Composites were significant between Times 1 and 2 ($r=0.32$, $p<0.05$, $n=49$) and between Times 1 and 3 ($r=0.32$, $p<0.05$, $n=49$), but not between Times 2 and 3 ($r=-0.01$, $p=0.94$, $n=53$). Because child behavior may differ across activities, we ran correlations between time points for which children were observed in the same type of activity (i.e., large or small group). Though this subsample of children was small, results indicate that the Adaptive Behavior Composite had stronger test–retest reliability in large than in small groups. For Problem Behaviors, most correlations were stronger when the

child was observed in the same setting at both time points compared to when setting was not consistent (see Table 2).

Convergent and Divergent Validity

Convergent validity was examined with correlations between baseline BASC-3 SOS Composite Adaptive Behavior and baseline DECA TPF Scores as well as correlations between baseline BASC-3 SOS Composite Problem Behavior and baseline SESBI-R Scores. Divergent validity was examined by calculating correlations between baseline BASC-3 SOS composite Adaptive Behavior and baseline SESBI-R scores as well as correlations between baseline BASC-3 SOS Composite Problem Behavior and baseline DECA TPF Scores. These correlations were statistically significant, in the expected direction, and weak to moderate in strength (convergent validity: Adaptive Behavior Composite and DECA TPF $r = 0.37$, $p < 0.001$, Problem Behavior Composite and SESBI-R $r = 0.38$, $p < 0.001$; divergent validity: Adaptive Behavior Composite and SESBI-R $r = -0.29$, $p < 0.01$, Problem Behavior Composite and DECA TPF $r = -0.43$, $p < 0.001$). Additionally, correlations between Adaptive Behavior Composite Scores and both the SESBI-R and the DECA TPF Scores in large group settings were stronger than in small group settings. Correlations between Problem Behavior Composite Scores and both the SESBI-R and DECA TPF Scores in small group settings were stronger than in large group settings (see Table 2).

Predictive Validity

Predictive validity was examined with correlations between baseline BASC-3 SOS Composite Behavior scores and Time 3 DECA TPF and SESBI-R Scores. Only children in the waitlist group were include in this analysis, as no changes in behavior between time points were hypothesized for these children (i.e., predictive validity would be unaffected by intervention effect in this waitlist group). The correlation between BASC-3 SOS Composite Adaptive Behavior Scores at baseline and Time 3 DECA TPF Scores for the waitlist group was positive, moderate, and statistically significant ($r = 0.52$, $p < 0.001$). Similarly, the correlation between BASC-3 SOS Composite Problem Behavior Scores at baseline and Time 3 SESBI-R Scores for the waitlist group was positive, moderate, and statistically significant ($r = 0.40$, $p < 0.01$).

Discriminant Validity

T-tests were used to compare average Composite Behavior Scores between children with and without identified disabilities as a measure of discriminant validity. In this sample, children with disabilities engaged in significantly more frequent Problem Behavior ($M = 55.5$, SD = 4.3) compared to those without disabilities ($M = 33.6$, SD = 5.9; $t(51.1) = -2.99$, $p = 0.004$), as well as significantly less frequent Adaptive

Behaviors ($M = 74.5$, SD = 2.5) compared to those without disabilities ($M = 84.2$, $SD = 3.5$; $t(105) = 2.0$, $p = 0.048$).

Treatment Sensitivity

ANOVAs comparing baseline (i.e., Time 1) BASC-3 SOS Composite Scores with post intervention Composite Scores (i.e., Time 2) for children in the intervention group and the waitlist group served as a measure of treatment sensitivity. Neither ANOVA revealed a main effect of time (Adaptive Behavior: $F(1, 95) = 0.12$, $p = 0.73$; Problem Behavior: $F(1, 95) = 0.87$, $p = 0.35$) or an interaction between time and group membership (Adaptive Behavior: $F(1, 95) = 0.01$, $p = 0.92$; Problem Behavior: $F(1, 95) = 0.03$, $p = 0.87$).

Discussion

Identifying evidence-based assessments for classroom behavior of young children, including those from racial/ethnic minority backgrounds and those with developmental concerns, is critical to accurately identify those in need of interventions, determine appropriate treatment targets, and monitor response to interventions. This is especially important for young children with developmental delays because they are at higher risk for behavior problems (Emerson & Einfeld, 2010), and if inappropriate behaviors are left untreated, children are at risk for more serious conduct problems later in development (McMahon & Frick, 2005). This study extends the literature on SDO tools by examining the psychometric properties of the BASC-3 SOS, one tool that is often used clinically but that has limited research on its psychometric properties. Our results provide mixed support for the use the BASC-3 SOS with young children diverse in terms of race/ethnicity and disability given that not all behaviors were coded reliably and that we found mixed evidence for various areas of reliability and validity.

Overall, observations using the BASC-3 SOS revealed low frequencies of Problem Behaviors across children in our sample, with individual behaviors occurring in 0–25% of observed intervals. Inattention and Inappropriate Movement were the most commonly observed problem behaviors, while Aggression, Somatization, Inappropriate Sexual Behavior, and Bowel/Bladder Problems were observed in fewer than 1% of intervals, indicating that these particular behavior variables may have limited utility when the BASC-3 SOS is used with young children. The overall low observed rates of Problem Behaviors among young children without clinically significant behavior problems align with other SDOs such as the REDSOCS (i.e., observed in 5–17% of intervals; Jacobs et al., 2000) and the BOSS-EE (i.e., observed in 0–17% of intervals; Hojnoski et al., 2020). In our sample, Problem Behaviors occurred more frequently in large groups and Adaptive Behaviors occurred more frequently in small groups, a trend that is consistent with research using the BOSS-EE (Hojnoski et al., 2020).

Of note, another potential contributing factor to the low frequency of problem behaviors pertains to the use of a momentary time sampling procedure. Specifically, the BASC-3 SOS manual specifies that behaviors be coded in a 3-s momentary time period at the end of each 30-s interval. Thus, if a behavior occurs only during the first 27 s of that interval, it is not measured with this approach. While momentary time sampling is considered to yield the most accurate estimate of some behaviors (Hintze et al., 2002), it may yield inaccurate estimates of discrete behaviors that occur infrequently. The use of partial interval sampling would be one approach to capture any instances of discrete behaviors during the observational period and thus yield a potentially more accurate estimate of behavior. Alternatively, shorter interval lengths (e.g., 15 s) would result in more opportunities to record behaviors and fewer instances of behaviors occurring but not being captured due to a procedural aspect of the coding system. Indeed, Briesch and colleagues (2017) found that shorter interval lengths yielded more dependable estimates of behavior than longer intervals. The use of momentary time sampling for measuring discrete behaviors within the BASC-3 SOS is thus an important limitation of this tool that researchers and practitioners should consider when selecting an SDO for their purpose.

In the current study, inter-observer agreement was moderate to good for the Adaptive Behaviors of Response to Teacher/Lesson, Work on School Subjects, and Peer Interaction, as well as for the Problem Behaviors of Inattention, Inappropriate Movement, Inappropriate Vocalization, Self-Injurious Behavior, and Other behaviors. ICCs were lower than anticipated for Somatization, Repetitive Motor Movement, Inappropriate Peer Interaction, Aggression, and Transition Movement. Low inter-observer reliability for some Problem Behaviors has also been reported using other SDOs for young children, with one reason being the low frequency of these behaviors impacting reliability calculations (Bagner et al., 2010; Hojnoski et al., 2020; Koo & Li, 2016). Two other factors that may have contributed to some lower than anticipated ICCs are insufficient initial/ongoing coder training or a lack of clarity in the definitions of behaviors. Our team used a running list of examples of each behavior category as a way to further operationally define these behaviors in this population (see supplemental material). It would be helpful for the BASC-3 SOS developers to expand their manual to provide more examples of behaviors that fit each of their categories for children in different age groups to promote highly reliable coding. Although the BASC-3 SOS manual asserts that the tool can be used without the extensive training that other SDOs often require (Lett & Kamphaus, 1997), the low inter-observer reliability for some behaviors in our study suggests that further training may be necessary. For example, observers may need to meet reliability criteria on more than 3 videos (the number required for the current study), check reliability against a reliable coder periodically, and participate in booster training to avoid observer drift. Given the limitations on clinician's time, this level of training may be a barrier to using the BASC-3 SOS in practice.

Consistency in scores across time is another important indicator of the quality of an SDO (Hintze, 2005). Our results provide mixed evidence for test–retest reliability of Adaptive and Problem Behavior Scores, as we found significant positive correlations between most time points for children in our waitlist group, and these relationships were of a similar magnitude as reported for the BOSS-EE (Hojnoski et al.,

2020). An exception to this finding was that Problem Behaviors between Times 2 and 3 were not significantly correlated, with one possible reason being that the activity setting was not standardized across observations. Many SDOs suggest that children be observed in the same setting across observations to obtain a reliable estimate of their behavior under particular conditions (Briesch et al., 2018), an approach that was not feasible in the current study. Our exploratory analyses indicated that test–retest reliability was the strongest for Adaptive Behaviors in large group settings. In small group settings, Adaptive Behaviors were not positively correlated across time points. Problem Behaviors were more strongly correlated when children were observed in the same setting (i.e., small group at both time points) as opposed to different settings (i.e., small group at Time 1, large group at Time 2) for most comparisons. However, the small sample size for these exploratory analyses limits conclusions that can be drawn. Also of note, observations were separated by several weeks, a period of time across which behavior may be less stable in comparison to a few days. This is one potential reason that test–retest reliability was lower than anticipated in some cases.

Evidence for convergent validity of the BASC-3 SOS was demonstrated by positive, moderate, and statistically significant correlations between adaptive behaviors as measured by teacher ratings on the DECA-TPF and BASC-3 SOS Adaptive Behavior Composite Scores, as well as between problem behaviors as measured by teacher ratings on the SESBI-R and BASC-3 SOS Problem Behavior Composite Scores. The magnitude of these correlations was similar to estimates of convergent validity when comparing teacher ratings to other SDOs; the correlation coefficients in the present study were 0.37 and 0.38 while REDSOCS and BOSS-EE correlations have ranged from 0.16 to 0.55 (Bagner et al., 2010; Hojnoski et al., 2020; Jacobs et al., 2000). Our results thus provide evidence that the BASC-3 SOS yields meaningful estimates of Adaptive and Problem Behavior of young children, including those with disabilities.

Regarding divergent validity, we found negative, small to moderate, and statistically significant correlations between appropriate behaviors as measured by teacher ratings on the DECA-TPF and BASC-3 SOS Problem Behavior Composite scores, as well as between appropriate behaviors as measured by teacher ratings on the SESBI-R and BASC-3 SOS Problem Behavior Composite scores. These correlations are of similar magnitude to correlations between BOSS-EE classroom engagement behavior and teacher-rated inappropriate behavior (i.e., −0.29 and −0.43 in this study, −0.41 in Hojnoski et al., 2020). Our results indicate that the BASC-3 SOS does not positively relate to variables with which theoretically it should not.

Predictive validity is critical to understand whether a measure is able to forecast a criterion at a later time point (Hintze, 2005). This psychometric property has not been reported for other SDO tools for young children. In the current study, correlations between BASC-3 SOS Adaptive Behavior Composites and teacher-rated positive behavior several months later were positive, statistically significant, and moderate in strength. Further, correlations between BASC-3 SOS Problem Behavior Composites and teacher-rated classroom problem behavior several months later were also positive, statistically significant, and moderate in strength. These findings

strengthen the utility of the BASC-3 SOS, as they indicate that scores on this tool are meaningfully related to classroom behavior in the future.

Higher BASC-3 SOS Problem Behavior Scores and lower Adaptive Behavior Scores among children with disabilities compared to those without disabilities provides initial evidence for the discriminant validity of this tool when used with young children. Importantly, we did not find ceiling or floor effects in either of these populations, providing initial evidence that the BASC-3 SOS functions well when used with both populations while also discriminating between them in the hypothesized direction. Given that our sample of children without disabilities was relatively small, further exploration of the psychometrics of this tool in larger samples of children is important to deepen an understanding of the usefulness and appropriateness of this tool with different populations. Similarly, research into whether the BASC-3 SOS discriminates between young children with and without ADHD, as has been demonstrated for school-age children (Lett & Kamphaus, 1997), is an important future direction.

The only psychometric property that we did not find any evidence for was treatment sensitivity; there was not a significant decrease in BASC-3 SOS Problem Behavior or a significant increase in Adaptive Behavior among children who received TCIT-U, an intervention aiming to promote positive behavior. This finding aligns with a recent investigation of TCIT-U using the REDSOCS (Fawley et al., 2020). One possibility is that SDOs, and the BASC-3 SOS in particular, may not capture changes in response to interventions implemented with non-clinical samples. In our sample, at baseline, Adaptive Behavior frequency was high and Problem Behavior frequency was low, and only 15% of children showed clinical elevation in problem behavior on the SESBI-R, leaving little room for improvement. TCIT-U led to significant but small improvements in child behavior in this sample for the treatment group over the control group based on teacher ratings (Davis et al., 2021), suggesting that TCIT-U did have a measurable impact on the classroom behavior of students. Thus, perhaps behavioral improvement of this small magnitude is not detected through SDOs. Other studies that found evidence for treatment sensitivity of SDOs have used samples in which rates of disruptive behavior were higher prior to intervention (e.g., Bagner et al., 2010). Of note, in a study using the REDSOCS, child noncompliance did not demonstrate sensitivity to treatment even in their clinical sample, which the authors attribute to low baseline frequency (Bagner et al., 2010), supporting this idea that SDOs may not capture changes in behavior when the behavior occurs infrequently to begin with. In contrast, it is possible that the BASC-3 SOS may demonstrate treatment sensitivity following Tier 3 behavioral supports among children exhibiting severe externalizing behavior prior to intervention. Future research utilizing single case research designs may be helpful to explore whether the BASC-3 SOS captures changes over time when Tier 3 interventions are implemented with individual students. A final possibility is that the 14- to 16-week time period between baseline and post-intervention in the current study may not be a sufficient time period to confer observable intervention effects of TCIT-U on child behavior using an SDO.

Study Strengths

This study addresses gaps in the literature on SDOs for young children generally and on the BASC-3 SOS in particular in several ways. A previous study of the BASC-2 SOS focused primarily on composite behaviors and did not include sufficient information about individual behaviors (Lett & Kamphaus, 1997). Examining inter-observer agreement of individual behaviors is important because these scores can be particularly useful for identifying intervention targets, thus understanding if they can be coded reliably is critical. We explored the frequencies and inter-observer reliability across each individual behavior and used this information to create more meaningful composite scores. Further, including multiple time points and collecting data in the context of an RCT allowed for examination of test–retest reliability, treatment sensitivity, and predictive validity, which have seldom been explored in this area of literature. This study thus provides new and critical information about the strengths and weaknesses of the BASC-3 SOS when used for clinical and research purposes. Additionally, observations conducted in both small and large group settings enabled initial exploration of the psychometrics of the BASC-3 SOS in activities that may include differing behavioral expectations and thus differing behaviors.

One strength of our sample is the representation of racial/ethnically diverse young children with various disabilities, which is unique within the SDO literature, as many studies have included primarily White samples of children with either no disability/disorder or with ADHD (e.g., Bagner et al., 2010; Jacobs et al., 2000; Lett & Kamphaus, 1997). In contrast, only 11% of children in our study were Non-Latinx White, and 78% had a delay/disability such as autism, hearing impairment, or language delay. If SDOs are to be used with diverse populations, it is critical that psychometrics are explored in samples that reflect this diversity in order to ensure that scores are meaningful indicators of child behavior (Bulotsky-Shearer et al., 2013; Salvia & Ysseldyke, 1991). Given that some psychometric properties were weaker than those previously reported in our primarily Latinx sample (e.g., inter-observer reliability), future research should explore the psychometrics of the BASC-3 SOS in older Latinx children to shed light on whether this difference is more related to the race/ethnicity of children or to their age. Second, our sample included children with and without elevated disruptive behavior, allowing us to explore the extent to which this SDO is sensitive enough to capture instances of behavior among general samples of young children.

Limitations and Future Directions

Regarding limitations, first, Problem Behaviors were not coded in a mutually exclusive manner; for example, hitting a child while yelling would be coded as Aggression, Inappropriate Vocalization, and Inappropriate Interaction. A benefit of this approach is that it captures the full range of problem behavior; however, this approach may lead to overestimates of a child's total problem behavior. The BASC-3 SOS manual does not explicitly indicate whether behaviors should or should not be coded mutually exclusively, but our team's understanding based on example

behaviors provided for each category was that this is how the tool was designed to be used. Clarification in the BASC-3 SOS manual would be helpful regarding this aspect of coding. Of note, our Composite Adaptive and Problem Behavior Scores only included behaviors that were reliably coded according to ICCs. This approach minimizes measurement error and thus yields meaningful composites, but our composites do not include all behaviors due to some having low ICCs. While the BASC-3 SOS indicates that users can sum all behaviors to obtain overall estimates of adaptive and problem behaviors, our results suggest that this approach may not be appropriate when using the tool with preschool-age children.

The BASC-3 SOS was developed for a wide age range (i.e., 2–21 years), but whether or not the tool captures behaviors that have relevance for the early childhood age group should be further explored (i.e., content validity). Some behaviors never or very seldomly occurred in our sample and some had low inter-rater reliability. Our team maintained a working document of behavior examples to clarify how to apply behavioral definitions to our population. A systematic approach to content validity is warranted, which could include surveying experts and practitioners about the operational definitions of each behavior, as other researchers have done (e.g., Hojnoski et al., 2020). Removal of variables that are deemed not to be relevant for this population would also have implications for other psychometric properties, as low frequency of behaviors can negatively impact inter-rater reliability estimates.

Another limitation of this study is that we were only able to conduct one observation of each child at each time point. There is a paucity of data-informed guidance regarding the number and length of observations needed to obtain reliable and valid measures of child behavior (McMahon & Frick, 2005). Some SDO tools recommend three observations of a child at the same time point in a particular setting (Briesch et al., 2018), thus future studies of the BASC-3 SOS could follow this recommendation to further explore reliability and validity. Alternatively, Generalizability theory (G theory) could be used to systematically determine the number of observations that would be required for the BASC-3 SOS to yield reliable estimates of behavior (Hintze, 2005).

It is also of note that children in this study were not always observed during the same type of activity at each time point. Our exploratory analysis of behavior in small versus large group settings as well as data from the BOSS-EE (Hojnoski et al., 2020) suggest that young children's behavior may differ across activities. For example, a child's behavior during a small group free play activity may not necessarily be representative of how that same child would behave in a large group pre-academic activity. Future studies of the BASC-3 SOS should include larger samples of children observed in the same types of activities across time points.

Implications

With few SDOs available to measure children's behavior in early education settings, identifying psychometrically sound tools is an important goal. Our results indicate that many of the behaviors measured by the BASC-3 SOS have utility for young children, while some (i.e., Transition Movement, Aggression, Inappropriate Peer Interaction, Somatization, Bowel/Bladder Problems, Inappropriate Sexual Behavior,

Repetitive Motor Movement) were observed very infrequently and less reliably in this sample. However, our study did not examine whether these behavior codes may be relevant to assessment and intervention planning for children with specific behavior concerns. By providing evidence for inter-observer and test–retest reliability as well as convergent, divergent, discriminant, and predictive validity of the BASC-3 SOS when including reliably measured behaviors, this study builds evidence for the psychometric merit of this tool for some uses in early childhood settings that include children receiving special education. In particular, the BASC-3 SOS may enhance a multi-method school psychological assessment that also includes teacher- and parent-reported rating scales, direct assessment, and clinical interviews to thoroughly understand a child's areas of behavioral difficulty and strengths. For behavior analysts, the BASC-3 SOS may be helpful in an initial observation of a student to ensure that the occurrence/frequency of a wide variety of problem behaviors are considered before treatment plans and progress monitoring tools that target individual behaviors are designed. At the same time, the lack of evidence for treatment sensitivity in our study indicates that the BASC-3 SOS may not be useful for evaluating overall intervention effectiveness or for progress monitoring. Of note, the quantity of behaviors included in the BASC-3 SOS may preclude its clinical utility for progress monitoring an individual student's specific problem behaviors. Further exploration of the reliability and validity of the BASC-3 SOS as outlined previously will be important to build on our findings and to further clarify the appropriate clinical and research uses of this tool.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10864-021-09458-x.

Acknowledgements The authors would like to acknowledge the children, teachers, and other school staff who participated in this study, as well as our undergraduate research assistants.

Funding This research was generously supported by a grant from The Children's Trust (Miami, FL). The funder had no involvement in study design, collection, analysis or interpretation of data, writing of the report, or decision to submit the article for publication.

Declarations

Conflict of interest The authors declare they have no relevant financial or non-financial interests.

Consent to Participate Teachers and parents of students provided written informed consent.

Ethical Approval All procedures were approved by the Institutional Review Board at the University of Miami and were in line with the tenants of the Declaration of Helsinki.

References

Bagner, D. M., Boggs, S. R., & Eyberg, S. M. (2010). Evidence-based school behavior assessment of externalizing behavior in young children. *Education & Treatment of Children, 33*(1), 65–83. https://doi.org/10.1353/etc.0.0084

Bramlett, R. K., & Barnett, D. W. (1993). The development of a direct observation code for use in preschool settings. *School Psychology Review, 22*(1), 49–62.

Briesch, A., Volpe, R. J., & Floyd, R. G. (2018). *School-Based Observation: A Practical Guide to Assessing Student Behavior*. The Guilford Press.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, 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*(3), 408–421.DIO: https://doi.org/10.1080/02796015.2010.12087761

Briesch, A., Ferguson, T., Daniels, B., Volpe, R. J, & Feinberg, A. (2017). Examining the influence of interval length on the dependability of observational estimates. *School Psychology Review, 46*(4), 426–432. https://doi.org/10.17105/SPR-2016-0006.V46-4

Budd, K. S., Garbacz, L. L., & Carter, J. S. (2016). Collaborating with public school partners to implement Teacher-Child Interaction Training (TCIT) as universal prevention. *School Mental Health, 8*(2), 207–221. https://doi.org/10.1007/s12310-015-9158-8

Bulotsky-Shearer, R. J., Fernandez, V. A., & Rainelli, S. (2013). The validity of the Devereux Early Childhood Assessment for culturally and linguistically diverse Head Start children. *Early Childhood Research Quarterly, 28*(4), 794–807. https://doi.org/10.1016/j.ecresq.2013.07.009

Cai, X., Kaiser, A. P., & Hancock, T. B. (2004). Parent and teacher agreement on Child Behavior Checklist items in a sample of preschoolers from low-income and predominantly African American families. *Journal of Clinical Child and Adolescent Psychology, 33*(2), 303–312. https://doi.org/10.1207/s15374424jccp3302_12

Carter, A. S., Briggs-Gowan, M. J., Jones, S. M., & Little, T. D. (2003). The Infant-Toddler Social and Emotional Assessment (ITSEA): Factor structure, reliability, and validity. *Journal of Abnormal Child Psychology, 31*(5), 495–514. https://doi.org/10.1023/A:1025449031360

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

Davidson, B. C., Davis, E., Cadenas, H., Barnett, M., Sanchez, B. E. L., Gonzalez, J. C., & Jent, J. (2021). Universal Teacher-Child Interaction Training in early special education: A pilot cluster-randomized control trial. *Behavior Therapy, 52*(2), 379–393. https://doi.org/10.1016/j.beth.2020.04.014

Davis, E. M., Schmidt, E. M., Rothenberg, W. A., Fernandez, C., Davidson, B., Barnett, M., Garcia, D., Jent, J. (2021). *Universal Teacher-Child Interaction Training in early special education: A cluster randomized control trial*. Manuscript submitted for publication.

Emerson, E., & Einfeld, S. (2010). Emotional and behavioural difficulties in young children with and without developmental delay: A bi-national perspective. *Journal of Child Psychology and Psychiatry, 51*(5), 583–593. https://doi.org/10.1111/j.1469-7610.2009.02179.x

Eyberg, S., & Pincus, D. (1999). *Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory-Revised: Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Fawley, K., Stokes, T., Rainear, C., Rossi, J., & Budd, K. (2020). Universal TCIT improves teacher–child interactions and management of child behavior. *Journal of Behavioral Education, 29*, 635–656. https://doi.org/10.1007/s10864-019-09337-6

Ginn, N., Seib, A., Boggs, S. R., & Eyberg, S. M. (2009). Manual for the revised edition of the school observation coding system (REDSOCS).

Hintze, J. (2005). Psychometrics of direct observation. *School Psychology Review, 34*(4), 507–519. https://doi.org/10.1080/02796015.2005.12088012

Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2002). Best practices in the systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (4th ed., pp 993–1006). Bethesda, MD: National Association of School Psychologists.

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

Jacobs, J. R., Boggs, S. R., Eyberg, S. M., Edwards, D., Durning, P., Querido, J. G., et al. (2000). Psychometric properties and reference point data for the revised edition of the school observation coding system. *Behavior Therapy, 31*(4),695–712. https://doi.org/10.1016/S0005-7894(00)80039-8

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

LeBuffe, P. & Naglieri, J. (2012). *Devereux Early Childhood Assessment for Preschoolers Technical Manual*. Lewisville, NC: Kaplan.

Lett, N. J., & Kamphaus, R. W. (1997). Differential validity of the BASC student observation system and the BASC teacher rating scale. *Canadian Journal of School Psychology, 13*(1), 1–14. https://doi.org/10.1177/082957359701300101

Mackrain, M., Lebuffe, P., & Powell, G. (2007). *Devereux Early Childhood Assessment for Infants and Toddlers Technical Manual*. Lewisville, NC: Kaplan.

Margiano, S. G., Kehle, T. J., Bray, M. A., Nastasi, B. K., & DeWees, K. (2009). Examination of the effects of self-modeling on autobiographical memory. *Canadian Journal of School Psychology, 24*(3), 203–221. https://doi.org/10.1177/0829573509343096

McMahon, R. J., & Frick, P. J. (2005). Evidence-based assessment of conduct problems in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 477–505. https://doi.org/10.1207/s15374424jccp3403_6

Noldus Information Technology. (2017). *Noldus The Observer XT 14*. Wageningen.

Olivier, E., Morin, A. J. S., Langlois, J., Tardif-Grenier, K., & Archambault, I. (2020). Internalizing and externalizing behavior problems and student engagement in elementary and secondary school students. *Journal of Youth and Adolescence, 49*, 2327–2346. https://doi.org/10.1007/s10964-020-01295-x

Querido, J. G., & Eyberg, S. M. (2003). Psychometric properties of the Sutter-Eyberg Student Behavior Inventory-Revised with preschool children. *Behavior Therapy, 34*(1), 1–15. https://doi.org/10.1016/S0005-7894(03)80018-7

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior Assessment System for Children–Third Edition, Student Observation System*. Bloomington, MN: Pearson.

Salvia, J., & Ysseldyke, J. E. (1991). *Assessment* (5th ed.). Houghton Mifflin.

Salvia, J., & Ysseldyke, J. E. (2004). Assessment in special and inclusiveeducation (9th ed.). Princeton, NJ:Houghton Mifflin.

Authors and Affiliations

Ellyn M. Schmidt[1] · W. Andrew Rothenberg[1,2] · Bridget C. Davidson[1,3] · Miya Barnett[4] · Jason Jent[1] · Heleny Cadenas[1] · Corina Fernandez[1] · Eileen Davis[1]

W. Andrew Rothenberg
war37@miami.edu

Bridget C. Davidson
dr.bridgetdavidson@gmail.com

Miya Barnett
mbarnett@ucsb.edu

Jason Jent
jjent@med.miami.edu

Heleny Cadenas
helenycadenas@gmail.com

Corina Fernandez
cxf475@med.miami.edu

Eileen Davis
exm305@miami.edu

[1] Department of Pediatrics, Mailman Center for Child Development, University of Miami Miller School of Medicine, 1601 NW 12th Avenue, Miami, FL 33136, USA

[2] Duke University Center for Child and Family Policy, Durham, NC 27708, USA

[3] Pediatric Psychology Associates, Aventura, FL, USA

[4] University of California, Santa Barbara, CA, USA