



Development and Initial Psychometrics of a Generic Treatment Integrity Measure Designed to Assess Practice Elements Targeting Social, Emotional, and Behavioral Outcomes in Early Childhood Settings

Bryce D. McLeod¹ · Kevin S. Sutherland² · Michael Broda² · Kristen L. Granger² · Ruben G. Martinez¹ · Maureen A. Conroy³ · Patricia A. Snyder³ · Michael A. Southam-Gerow¹

Accepted: 14 September 2021
© Society for Prevention Research 2021

Abstract

Though treatment integrity measurement is important for research intended to promote social and behavioral outcomes of children at risk for emotional and behavioral disorders (EBDs) in early childhood settings, measurement gaps exist in the field. This paper reports on the development and preliminary psychometric assessment of the treatment integrity measure for early childhood settings (TIMECS), an observational measure designed to address existing measurement gaps related to treatment integrity with tier 2 interventions in the early childhood field. To assess the preliminary score reliability (interrater) and validity (construct, discriminant) of the TIMECS, live observations ($N=650$) in early childhood classrooms from 54 teachers (92.6% female, 7.4% male; 61.1% White) and 91 children (M age = 4.53 years, $SD = .44$; 45.1% female, 54.9% male; 45.1% Black) at risk for EBDs were scored by 12 coders using the TIMECS and an observational measure designed to assess teacher–child interactions. Teachers also self-reported on the quality of the teacher–child relationship. Interrater reliability (intraclass correlation coefficients, ICC [2,2]) for the quantity (i.e., adherence) item scores had a mean of .81 ($SD = .07$; range from .68 to .95), and the quality (i.e., competence) item scores had a mean of .69 ($SD = .08$; range from .52 to .80). Scores on the TIMECS Quantity and Quality items and scales showed evidence of construct validity, with the magnitude of the correlations suggesting that the quantity and quality items assess distinct components of treatment integrity. A TIMECS quantity scale also showed promise for intervention evaluation research by discriminating between teachers who had and had not been trained in a specific evidence-based intervention targeting social and behavioral skills in early childhood. The findings support the potential of the TIMECS to assess treatment integrity of teacher-delivered practices designed to address child social and behavioral outcomes of children at risk for EBDs in early childhood settings.

Keywords Treatment integrity · Teacher implementation · Practice elements · Early childhood

In response to the complex and often intensive needs of young children at risk for emotional and behavioral disorders (EBDs), teachers implement targeted or individualized

evidence-based interventions (EBIs) to help promote social-emotional competencies. Assessing and reporting on the treatment integrity with which these EBIs are delivered is important for intervention evaluation and implementation research (Proctor et al., 2011; Sutherland et al., 2013). Treatment integrity (also called fidelity; Sutherland et al., 2013) in the education literature refers to the degree to which an EBI is delivered as intended and is often noted to be composed of five components: adherence, dosage, differentiation, competence (or quality), and responsiveness (Sanetti et al., 2020; Sutherland et al., 2013). Adherence refers to the extent to which a teacher delivers practices prescribed by the EBI protocol. Dosage denotes the extent to which a child

✉ Bryce D. McLeod
bmcLeod@vcu.edu

¹ Department of Psychology, Virginia Commonwealth University, 806 West Franklin Street, PO Box 842018, Richmond, VA 23284-2018, USA

² School of Education, Virginia Commonwealth University, Richmond, VA, USA

³ Anita Zucker Center for Excellence in Early Childhood Studies, University of Florida, Gainesville, FL, USA

is exposed to a practice contained within the EBI protocol. Differentiation denotes the extent to which a teacher delivers practices proscribed by the EBI protocol (e.g., treatment purity). Competence refers to the level of skill and degree of responsiveness demonstrated by a teacher when delivering the practices prescribed by the EBI. And responsiveness captures the extent to which a child engages in and receives the key ingredients that are part of the EBI. Each of these components captures a unique aspect of treatment integrity that is important to assess in intervention evaluation and implementation research.

Intervention and Implementation Research

For intervention research, assessment of treatment integrity can help researchers interpret study findings in relation to child social and behavioral outcomes. If an EBI outperforms a comparison group, then it is critical to ensure that the findings can be attributed to the EBI. The measurement of adherence and competence can thus aid interpretation of study findings by establishing the extent to which an EBI was delivered and how well it was delivered. Conversely, it is important to determine whether “failure” to produce a desired outcome is due to an EBI. If the EBI does not produce a desired outcome, yet adherence and competence are sufficient, then the EBI may not be effective in a specific context (Sutherland et al., 2013).

Treatment integrity is also important to assess as an outcome in implementation research (Proctor et al., 2011). Poor social and behavioral outcomes produced by an EBI that has been effective in other settings may indicate a need to adapt teacher training or consultation approaches. Adherence and competence can be assessed in order to gauge the success of implementation strategies that involve teacher professional development, including coaching efforts (Sutherland et al., 2013). Thus, the success of implementation strategies on teacher behavior can be gauged via the assessment of treatment integrity.

Though critical to intervention evaluation and implementation research (Sutherland et al., 2013), the science and measurement of treatment integrity are underdeveloped in the education field (Sutherland et al., 2013). A review of education research spanning early childhood to high school settings revealed that about 75% of studies reporting on practices or interventions delivered in school settings include treatment integrity data (Sanetti et al., 2020). Most studies that assess integrity focus on adherence, which leaves competence, dosage, differentiation, and responsiveness understudied (Sanetti et al., 2020). In fact, multiple components of treatment integrity are rarely assessed for EBIs (for an exception see Sheridan et al., 2014). In terms of methods, most treatment integrity measures in education research rely upon

observational methods (> 60%), with self-report and permanent product review the second and third most common methods used (Sanetti et al., 2020). Treatment integrity measures vary widely in the scoring strategies used to generate scores, ranging from simple checklists (Sutherland & Snyder, 2007) to the use of Likert-type ratings (e.g., Sutherland et al., 2014). Thus, although each treatment integrity component is important to assess for intervention evaluation and implementation research, there is variation in what components are assessed and how.

Considerations for the Development of Treatment Integrity Measures

Several barriers to measuring treatment integrity in education exist, including a lack of theory and specific guidelines for assessing integrity (Sutherland et al., 2013). And of the measures that exist, many are not suitable for intervention evaluation and implementation because they (a) only assess adherence to a specific EBI, (b) are not designed to assess variability in the delivery of practices (e.g., use a checklist format; Sutherland & Snyder, 2007), and (c) do not have established score reliability and validity. To advance the field, this study reports on the initial score reliability and validity of an observational measure that fills a prior gap in the availability of measures for assessing treatment integrity.

In developing and validating treatment integrity measures for intervention evaluation and implementation research conducted in early childhood settings, it is important to use observational methods to assess both adherence and competence. Observational methods are considered the “gold standard” as they provide an objective assessment (Sanetti et al., 2020; Sutherland et al., 2013), and other methods (e.g., self-report measures) have showed poor correspondence with observer-rated measures (see Caron et al., 2019; Chapman et al., 2013). Adherence, typically assessed using observational or self-report methods, is established as an important aspect of integrity as it estimates the extent to which an EBI is delivered as designed (Sanetti et al., 2020). However, competence may be more important to assess as it shows higher-order skills and a level of mastery of the practices contained within an EBI (Sutherland et al., 2013). Also, evidence suggests that teacher competence may be important in promoting positive child social-emotional outcomes (Sutherland et al., 2018). For gauging the success of teacher training and coaching efforts, assessing competence helps establish whether teachers may have mastered the delivery of the practices (e.g., skillfulness in delivery of each practice, responsiveness to child needs) contained within an EBI (e.g., Sutherland et al., 2018). Despite the importance of assessing competence, only a few observational and self-report measures designed to assess the quality of

teacher-delivered practices exist (Sanetti et al., 2020). Observational measures that can provide objective assessments of both adherence and competence are thus needed to support intervention evaluation and implementation research.

The assessment of adherence and competence is also important to efforts to link these components to child social and behavioral outcomes. An assumption of EBIs is that high quality delivery of the practices account for the influence of those programs on child outcomes (Durlak et al., 2011; Wanless & Domitrovich, 2015). Yet to our knowledge only one study in the early childhood literature has evaluated integrity-outcome relations (see Sutherland et al., 2018). Findings from the broader education literature are mixed. Some studies have found significant integrity-outcome associations (e.g., Abry et al., 2013; Sutherland et al., 2018), and other studies have found no significant associations (e.g., Berry et al., 2016). These varied findings mirror those from the mental health field where two meta-analyses have indicated that the integrity-outcome association is virtually zero (Collyer et al., 2019; Webb et al., 2010). Observational integrity measures designed to assess integrity-outcome relations are thus needed to evaluate if adherence and competence predict child outcomes in early childhood settings.

A second issue to consider is whether to develop integrity measures for a specific EBI or for use more broadly. Recognizing the need for novel measurement approaches that are pragmatic, researchers have developed integrity measures that are designed to be more flexible than traditional measures (e.g., Garland et al., 2010; McLeod et al., 2018). Items on these measures are not tied to specific EBIs, but are designed to assess practice elements (i.e., discrete practices) found across a variety of EBIs (Chorpita & Daleiden, 2009; McLeod et al., 2017). For example, “praise” (i.e., verbal statements of approval), tangible reinforcement (i.e., tangible item delivered contingently on an appropriate response), and emotion regulation (i.e., strategies to identify or help support children’s emotions) are the three most common practice elements used to address social-emotional competencies in the early childhood literature (see McLeod et al., 2017). Many early childhood settings deliver more than a single EBI, necessitating the use of multiple measures to estimate adherence and competence. However, a “generic” measure that is not tied to a specific EBI may permit stakeholders to utilize a single measure to gauge integrity for multiple EBIs (Sutherland et al., 2013). A generic measure may also allow researchers to describe the practices delivered by teachers in business as usual classrooms, which can aid interpretation of study findings that utilize these classrooms as a comparison condition.

Currently, observational measures that assess integrity focus on specific EBIs for young children. For example, the *Teaching Pyramid Observation Tool* (Hemmeter et al., 2018) is used to assess the adherence and competence with which teachers deliver social and behavioral practices associated

with the Pyramid Model, a multi-tiered EBI. As another example, the *BEST in CLASS Adherence and Competence Scale* (Sutherland et al., 2014) is used to assess the adherence and competence of the practices found in the BEST in CLASS intervention (Conroy et al., 2019; Sutherland et al., 2018). While these measures have strengths, neither focuses on the practices teachers use with specific children who are at risk for EBDs to support their social and behavioral learning. To advance, the field needs an observational measure that can assess teacher delivery of practices to young children at-risk for EBDs.

Current Study

This study reports on the development and preliminary psychometric evaluation of scores on the treatment integrity measure for early childhood settings (TIMECS), a measure designed to assess the quantity (i.e., adherence) and quality (i.e., competence) of teacher-delivered practices that focus on social and behavioral skills in early childhood classrooms. Our purpose was to develop a measure that could be used to support the evaluation and implementation of EBIs. We thus developed items that (a) capture practice elements (i.e., “discrete clinical technique used as part of a larger intervention plan”; Chorpita & Daleiden, 2009, p. 560) found across EBIs that target child social and behavioral outcomes (McLeod et al., 2017) and (b) gauge the quantity and quality of practice element delivery.

To guide the development and psychometric evaluation of the TIMECS items and scales, we used a multicomponent measurement model (Hogue et al., 2008; McLeod et al., 2018) that included the quantity of practice delivery, the quality of practice delivery, and relational factors (quality of teacher–child interactions and the relationship). This model has been used to guide the evaluation of construct validity (convergent, discriminant) of scores on integrity measures in the mental health field (e.g., Hogue et al., 2008). To establish score validity, we hypothesized that the item and scale scores would demonstrate convergence and divergence with related and distinct domains (quantity, quality, teacher–child relationship) across different methods (observer, self) consistent with patterns of associations seen in previous integrity research (e.g., Hogue et al., 2008; Sutherland et al., 2014). Specifically, we hypothesized that the TIMECS scores would evidence construct validity (convergent, discriminant) score validity by demonstrating large, but not redundant ($r > .70$; Kline, 1979), correlations between corresponding quantity and quality items (Sutherland et al., 2014). We hypothesized that the scores would evidence discriminant validity by demonstrating small to moderate correlations with observational and teacher-report measures of the teacher–child relationship (Hogue et al., 2008; Sutherland et al., 2014), as these

represent distinct constructs. We also evaluated whether targets of measurement (child, teacher, group, coder, time) accounted for systematic variation in TIMECS Quantity and Quality scales as child and teacher characteristics have been found to influence the quantity and quality of practice delivery differently (McLeod et al., 2018; Sutherland et al., 2018). If different targets of measurement account for diverse proportions of variance across the Quantity and Quality scales, this would provide evidence that the scales assess distinct integrity components. Finally, we evaluated the potential for using the TIMECS to assess the adherence and competence of an EBI by ascertaining if scales composed of items that map onto the content of an EBI could detect expected differences between teachers trained to deliver an EBI and teachers delivering business as usual (i.e., discriminative validity; see McLeod et al., 2018).

Method

Participants and Settings

Participants for this study were children and their teachers in early childhood classrooms in a Southeastern state. Classrooms were located in urban and suburban communities and were either federal- (e.g., Head Start) or state-funded early childhood programs. The teachers used a wide range of curricula, such as the High Scope Early Childhood Curriculum (Schweinhart & Weikart, 1997). Some classrooms also implemented targeted social, emotional, and behavioral learning curricula. The early childhood classrooms varied in their settings, with some part of elementary schools and others in community settings (e.g., community centers). On average, there were 17.26 ($SD = 3.54$) children and 2.09 ($SD = 0.29$) adults per classroom.

Our goal was to assess teacher delivery of practices targeted at children at risk for EBDs (i.e., *focal* children). Multiple measures were used to screen child participants for inclusion in the study. The focal children ranged in age from 3 to 5 years and were identified using the first two stages of the *Early Screening Project* (ESP; Walker et al., 1995). First, teachers identified up to 5 children who demonstrated the most severe and chronic problem behaviors in their classrooms. Once children were nominated, caregiver consent was sought. Next, teachers completed the second stage of the ESP, the Externalizer Teacher Questionnaire, for those children with caregiver consent to confirm risk for EBDs. Last, the *Battelle Developmental Inventory, Second Edition Screener* (Newborg, 2005) was used for each child, and if a child demonstrated average or above average cognitive abilities they were retained in the sample. Children with the two most extreme scores on the ESP with average or above average cognitive abilities were retained in the sample as the

focal children. The screening process resulted in the inclusion of 54 teacher participants and 91 children.

Teacher Participants The 54 teachers (94.4% 26–55+ years old; 92.6% female, 7.4% male; 61.1% White, 35.2% Black, 3.7% multiracial) averaged 7.69 ($SD = 7.98$) years of teaching in early childhood classrooms (see Table 1) and had the following educational backgrounds: 38.9% Bachelors; 48.1% Master's degree, 13.0% other degree. Seventeen teachers had previously received training and coaching in the BEST in CLASS intervention (Conroy et al., 2019), a tier 2 EBI that targets the reduction of problem behaviors demonstrated by young children at-risk for EBDs. These teachers had been trained to deliver BEST in CLASS intervention practices (see Conroy et al., 2019) via a 6-hour professional development workshop and received 14 weeks of coaching in a previous year. Both training and coaching focused upon six learning modules: (a) Rules, Expectations, and Routines; (b) Behavior Specific Praise; (c) Precorrection and Active Supervision; (d) Opportunities to Respond and Instructional Pacing; (e) Instructive and Corrective Feedback; and (f) Home-School Communication. A final module, Linking and Mastery, supported teachers in delivering practices efficiently with proficiency.

Child Participants We collected data from 91 children (54.9% male) who averaged 4.53 years old ($SD = .44$), 45.1% Black, 8.8% White, 1.1% Native American/American Indian, 1.1% Asian/Pacific Islander, 5.5% multiracial or other race, 45.0% race not reported, and 7.7% Latinx ethnicity. Thirty children were included in a classroom in which the teacher had previously been trained in BEST in CLASS. See Table 1 for a breakdown of demographic characteristics.

Development of the Treatment Integrity Measure for Early Childhood Classrooms

The TIMECS is a 21-item observational rating scale designed to assess the quantity and quality of teacher-delivered practices that foster positive social and behavioral outcomes for children at risk for EBD in early childhood settings. A three-step process based on the development of previous exemplar treatment integrity coding systems was used to develop the TIMECS (Hogue et al., 2008; Sutherland et al., 2014).

Step 1: Item Development Our goal was to develop a measure to assess integrity of practices found in EBIs designed to target social and behavioral skills, rather than the integrity of a specific EBI. To accomplish this goal, we followed methods used in the mental health field to distill practice elements from interventions (see Garland et al., 2010). We searched the early childhood literature to identify EBIs, interventions,

Table 1 Teacher demographic and training data; child demographic data

Teacher variables	<i>M (SD) or %</i>		<i>F or χ^2</i>
	BiC (<i>n</i> = 17)	BAU (<i>n</i> = 37)	
Sex			
Female	94.1	91.9	
Male	5.9	8.1	
Age			
18–25	–	2.7	7.96
26–35	64.7	29.7	
36–45	5.9	21.6	
46–55	23.5	21.6	
> 55	5.9	18.9	
Prefer not to answer	–	5.4	
Race			
Black	35.3	35.1	0.47
White	64.7	59.5	
Native American/American Indian	–	–	
Asian/Pacific Islander	–	–	
Multiracial	–	5.4	–
Highest level education			
High school diploma	5.9	2.7	4.75
Associates degree	5.9	2.7	
Bachelor's degree	23.5	45.9	
Master's degree	52.9	45.9	
Doctoral degree	5.9	–	
Other	5.9	2.7	
Years teaching	9.32 (6.12)	14.68 (10.34)	3.90
Years teaching preschool	7.06 (5.43)	7.99 (8.99)	0.49
Child variables			
	BiC (<i>n</i> = 30)	BAU (<i>n</i> = 61)	
Age	4.31 (0.57)	4.65 (0.31)	0.04
Female	46.7	54.30	0.36
Race			
Black	53.3	41.0	3.22
White	3.3	11.5	
Asian/Pacific Islander	–	1.6	
Native American/American Indian	–	1.6	
Other/multiracial	6.7	4.9	
Unknown/not reported	36.7	39.4	
Ethnicity			
Latinx	–	11.5	3.88*

Analysis of variance was conducted with continuous variables, and chi-square analyses were conducted with categorical variables

BiC BEST in CLASS, BAU business as usual

* $p < .05$; ** $p < .01$

and practices that had been evaluated in randomized group designs, quasi-experimental designs, and single-case experimental designs. Fifty articles were identified and an iterative process was used by the research team to distill the practice

elements from the interventions and practices. To ensure content validity, the practice elements were then reviewed by experts in the early childhood field. In all, 21 practice elements that could be used in an observational system were identified (see McLeod et al., 2017 for a description) that were designed to be independent such that no two items overlapped. Once the 21 practice elements were identified, we wrote two definitions for each practice; one for Quantity (adherence) and one for Quality (competence). See Table 2 for the items and definitions.

Step 2: Scoring Strategy We selected a scoring strategy for the Quantity items designed to capture the breadth and depth of practice delivery that has been used in previous exemplar integrity coding systems found in the mental health field (e.g., Hogue et al., 2008). This “extensiveness” scoring strategy involves ratings designed to measure the degree to which teachers use specific practices during an observation. Coders estimate the extent to which teachers engage in each practice during the entire observation using a 7-point Likert-type scale with the following anchors: 1 = *not at all*, 3 = *somewhat*, 5 = *considerably*, and 7 = *extensively*. Extensiveness ratings comprised two key components: thoroughness and frequency. Thoroughness refers to the depth, complexity, or persistence with which the teacher engages in each practice. Thoroughness is determined by: (a) the concentration of effort or commitment the teacher puts into the practice, (b) the detail in which the teacher describes the rationale for the practice, (c) the intensity of the practice, (d) the extent to which the teacher follows-through with the practice, or (e) the extent to which the practice is pursued intensively across an observation. Frequency refers to the number of times throughout the observation that a given practice is delivered (regardless of the thoroughness of the practice). Both thoroughness and frequency are considered in making a rating on each item; therefore, extensiveness ratings provide quantity, or dosage, information about each practice.

For the Quality items, we chose a “competence” scoring strategy used in exemplar competence coding systems developed for mental health interventions (e.g., Hogue et al., 2008; McLeod et al., 2018) that estimates the technical quality of practice elements (skillfulness) and their timing and appropriateness for the given child and situation (responsiveness). To gauge competence, coders are asked to make ratings on a 7-point Likert-type scale with the following anchors: 1 = *very poor*; 3 = *acceptable*; 5 = *good*; 7 = *excellent*. Coders are asked to consider the extent to which a teacher demonstrated four dimensions: (a) expertise, commitment, motivation; (b) clarity of language; (c) appropriate timing of the practice (responsiveness); and (d) ability to read and respond to where the child appears to be (responsiveness). Each dimension is considered when making a rating for each item.

Table 2 TIMECS Quantity and Quality item scores and interrater reliability

Item name	Definition	Quantity				Quality			
		<i>N</i>	Mean (<i>SD</i>)	Range	ICC	<i>N</i>	Mean (<i>SD</i>)	Range	ICC
Social skills	Teacher provides instruction on strategies that can facilitate positive social interactions with their peers or adults	650	1.66 (.85)	5.00	.88	248	4.44 (.87)	4.50	.77
Emotion regulation	Teacher provides instruction focused on identifying, labeling, or regulating emotions	650	1.33 (.68)	4.00	.89	125	4.30 (1.01)	5.00	.74
Problem solving	Teacher provides instruction designed to generate solutions to social, behavioral, or pre-academic problems	650	1.13 (.42)	4.50	.83	47	4.47 (.91)	4.00	.74
Promoting behavioral competence	Instruction that focuses on promoting positive behavior (e.g., engagement) during instructional activities	650	3.95 (1.05)	5.50	.80	635	4.52 (.97)	6.00	.80
Teacher–child relationship	Teacher behavior that conveys warmth, closeness, and interest when listening to and interacting	650	2.49 (1.25)	6.00	.87	422	4.84 (.91)	5.00	.77
Rules	Teacher teaches the rules and behavioral expectations of the classroom	650	1.43 (.77)	4.00	.90	149	4.86 (.90)	4.50	.71
Narrating	Teacher provides verbal description of behavior	650	1.46 (.66)	3.00	.80	170	4.60 (.91)	3.50	.64
Supportive listening	Teacher actively demonstrates understanding of the topic	650	1.73 (.94)	5.00	.83	239	4.63 (1.04)	4.50	.75
Choices	Teacher provides an opportunity to select between two or more options	650	1.11 (.32)	2.50	.68	43	4.21 (.82)	3.00	.76
Monitoring	Teacher actively monitors	650	4.80 (.92)	5.00	.69	643	4.78 (.90)	5.00	.69
Modeling	Teacher demonstrates, or has a peer demonstrate, a specific behavioral or pre-academic skill to promote learning	650	2.14 (1.09)	4.50	.81	331	4.56 (.77)	4.00	.59
Rehearsal	Teacher encourages practice of a behavioral skill (e.g., during interactions with peers)	650	1.20 (.50)	3.50	.80	56	4.59 (.81)	3.50	.58
Pre-correction	Teacher uses prompts prior to the occurrence of a behavior to remind of appropriate behavior and correct responding (e.g., reminding of rules, expectations)	650	1.58 (.73)	3.50	.77	202	4.68 (.84)	5.00	.56
Opportunities to respond	Teacher uses questions or prompts (i.e., gestural, verbal, visual, physical) that seek an active, observable, and specific response	650	4.67 (1.03)	6.00	.72	644	4.75 (.93)	5.00	.78
Visual cueing	Teacher uses visual cues to prompt for appropriate behavioral responses or consequences	650	1.77 (.93)	4.00	.79	248	4.42 (.82)	4.00	.58
Premack principle	Teacher uses a more reinforcing behavior (e.g., playtime) to reinforce less probable behaviors (e.g., lesson time)	650	1.15 (.41)	2.50	.80	58	3.73 (.86)	3.50	.66
Tangible reward	Teacher gives a tangible/representative reward in response to an appropriate social, emotional, or behavioral response	650	1.11 (.39)	2.50	.89	44	4.53 (1.00)	4.00	.76
Time-out	Teacher removes a child from a preferred activity for a specified period of time following a problem behavior	650	1.04 (.25)	3.00	.95	17	4.00 (1.19)	3.50	.68
Praise	Teacher provides positive verbal statements of approval in response to an appropriate social, emotional, behavioral, or pre-academic response	650	2.87 (1.08)	4.50	.82	533	4.52 (.90)	4.50	.73
Error correction	Teacher provides corrective feedback following an incorrect response or undesirable behavior	650	2.51 (1.07)	5.50	.79	449	4.20 (.93)	5.50	.70
Instructive feedback	Teacher provides extra instructional information while responding to correct response or appropriate behavior	650	1.51 (.71)	3.00	.76	179	4.68 (.74)	3.50	.52

TIMECS treatment integrity measure for early childhood classrooms, ICC intraclass correlation coefficient

Observations focused on teacher behavior directed toward a single focal child and were between 30 and 60 min in duration. The observations occurred during a teacher-led instructional (e.g., circle time, small group) activity, a child-led activity (e.g., center time), or transitions. The observations could be composed of more than one instructional context; however, most of the observations (e.g., at least 30 min) had to be composed of teacher-directed instructional time where the teacher was engaged with the focal child or a group of children that the focal child was part of. Observations had to be at least 30 min. Coders took notes during each observation. At the end of an observation, coders scored the 21 items for quantity and quality. If an item was not observed, a rating of “1” was given for quantity and the corresponding quality item was not scored.

Step 3: Scoring Manual Once the first steps were completed, we produced the TIMECS scoring manual. To promote interrater reliability and reduce coder inference, the scoring manual provided scoring procedures, a definition of each item, examples of extensiveness and competence for every item, exemplars, and guidelines for making item distinctions for both the quantity and quality items (see Hogue et al., 1996). Coders then used the scoring manual to code the practices delivered by teachers previously trained to deliver BEST in CLASS ($N=169$ observations) and teachers in business as usual ($N=125$) classrooms. Coder feedback was used to revise the TIMECS scoring manual.

Measures for Validity Analyses

The following self- and observer-report measures of the teacher-student relationship were collected to help establish the discriminant validity of the TIMECS item and scale scores.

Student–Teacher Relationship Scale Student–teacher relationship scale (STRS; Pianta & Hamre, 2001) is a teacher-report measure designed to assess the quality of teacher-student relationships. The 15 items are measured on a 5-point Likert-type scale from 1 (*definitely does not apply*) to 5 (*definitely applies*). The STRS has been used in studies of preschool children (e.g., Howes & Ritchie, 1999), demonstrated score validity with low-income samples (Hamre & Pianta, 2001), and evidenced predictive score validity with regard to academic and social functioning (Hamre & Pianta, 2001). The STRS Closeness and Conflict subscales were used in the present study, and internal consistency for the subscales was acceptable (Cronbach’s $\alpha=.78$ and $.89$, respectively).

Observational Teacher–Child Interaction Scale Observational teacher–child interaction scale (OTCIS) consists of six items that assesses child behavior directed towards the

teacher and the quality of teacher–child interactions. Sample items include: “The extent to which the child demonstrates that s/he likes or cares for the teacher”, and “The extent to which the child appears anxious, uncomfortable, or awkward when interacting with the teacher”. Coders observe teacher–child interactions and rate each item on a 6-point scale ranging from 0 (*not at all*) to 5 (*a great deal*). Scores on the OTCIS were produced by averaging across coders to produce item scores for each observation. The scale score was produced by averaging across items; higher scores represented more positive teacher–child interactions. The scale has demonstrated evidence of score reliability and validity (McLeod et al., 2021). For the current sample, reliability estimates were based on 650 classroom observations. Interrater reliability was $ICC(2,2)=.83$, and α was $.80$.

Coding and Observation Sampling Procedures

Four doctoral students in clinical psychology, three doctoral students in education, and five data staff (M age = 25.92 years, $SD=3.90$; 83.0% female; 58.0% White, 25.0% Latino, 8.0% Asian; 8.0% multiracial) were trained by the first and second authors to use the TIMECS and OTCIS over a 2-month period. The coders did not have previous experience using observational treatment integrity measures. Training started with didactic instruction and discussion of the scoring manual and coding exercises designed to test and expand understanding of each item. Next, coders engaged in independent coding of recordings, and results were discussed in weekly meetings. Coders then conducted independent live coding in early childhood classrooms. Finally, coders independently coded 40 10-minute videos, and reliability for each coder was assessed against master codes. In total, training took approximately 70 hours.

Independent coding commenced once a coder met “good” score reliability on each item ($ICC[2,2] \geq .60$, Cicchetti, 1994). We sought to code eight observations for each child. Coders were assigned to observations using a balanced incomplete block design (Fleiss, 1981). The order of observations was randomly assigned, stratifying for classroom and time. Observations ($N=650$) were an average of 40.56 min ($SD=11.08$). Two observers were sent to each class, and the observers were instructed to sit at the rear of the class and not interact with the teacher, teacher assistant, children, or each other. Following observations, the coders scored the TIMECS (integrity) and OTCIS (teacher–child interactions).

Data Analytic Strategy

Our data analytic strategy was based on practices common for the evaluation of the score reliability and validity of integrity measures (e.g., Hogue et al., 2008; Sutherland et al., 2014) that focuses on item performance, interrater reliability, and construct validity.

Interrater Reliability: TIMECS Quantity and Quality Items We examined descriptive statistics and interrater reliability for the TIMECS Quantity and Quality items. We hypothesized that TIMECS items would demonstrate “good” interrater reliability ($ICC [2,2] \geq .60$; Cicchetti, 1994). Interrater reliability was calculated for each item using the ICC (Shrout & Fleiss, 1979). The reliability coefficients represent the model ICC (2,2) based on a two-way random effects consistency model, which provides an estimate of the ratio of the true score variance to total variance. Thus, these ICCs provide a reliability estimate of the mean scores of all coders considered for each TIMECS item and allow for generalizability to other samples.

Construct Validity: TIMECS Quantity and Quality Item Scores We evaluated whether the TIMECS Quantity and Quality item scores demonstrated evidence of construct validity. These analyses focused on the magnitude of the correlations among scores on the TIMECS Quantity items, TIMECS Quality items, and the OTCIS. TIMECS item scores were generated by averaging scores across the two coders. Mean scores were used, as opposed to scores produced by a single coder, to reduce measurement error by removing differences across coders (McLeod et al., 2018). We hypothesized that a large correlation (i.e., convergent validity) would be observed between corresponding TIMECS Quantity and Quality items, as these items assess the quantity and quality of the same practice. We hypothesized that the correlations between the TIMECS Quantity and Quality items and the OTCIS would be small in magnitude (i.e., discriminant validity), as these measures are designed to assess distinct constructs. The correlations were interpreted following Rosenthal and Rosnow’s (1984) guidelines: r is “small” if 0.10–0.23, “medium” if 0.24–0.36, and “large” if > 0.36 . Follow-up contrasts evaluated if the correlations among the TIMECS items (Quantity, Quality) were significantly stronger than the correlations between the TIMECS items and the OTCIS. For the contrasts, standardized contrasts were calculated from the difference in the mean correlation coefficients (Hedges, 1994). The significance of each contrast was determined by dividing the contrast value by the square root of the pooled variance, which produces a z -score.

Construct Validity: TIMECS Quantity and Quality Scale Scores Two scale scores were created from TIMECS items: Quantity and Quality. Scale scores were generated by averaging items from each scale for each observation. To assess the score validity of the TIMECS Quantity and Quality scales, we assessed the magnitude of the correlations between scores on the TIMECS scales (Quantity, Quality), scores on an observational teacher–child interactions measure (OTCIS), and a teacher-report relationship measure

(STRS Closeness and Conflict scales). Analyses with the STRS are based on 59 children as the STRS was not collected during the first year of the project. We hypothesized that the strongest correlation would be observed between scores on the TIMECS Quantity and Quality scales (i.e., convergent validity) and that correlations with the teacher–child relationship scales would evidence smaller correlations (OTCIS, STRS Closeness and Conflict scales; i.e., discriminant validity). Follow-up contrasts were used to evaluate the hypotheses.

Discriminative Validity: TIMECS Quantity and Quality Scale Scores We conducted a series of variance components analyses on the BEST in CLASS Adherence and BEST in CLASS Competence scales that consist of the Quantity and Quality items that correspond to the practices used in the BEST in CLASS intervention (Rules, Precorrection, Opportunities to Respond, Praise, Error Correction [i.e., Corrective Feedback], and Instructive Feedback). Scores on the BEST in CLASS Adherence scale were created by producing a mean of the six TIMECS Quantity items for each observation, and scores on the BEST in CLASS Competence scale were created by producing a mean of the six TIMECS Quality items scored for each observation. A variance component analysis was performed on each TIMECS scale using the VCA package (Schuetzenmeister & Dufey, 2019) in *R* (R Core Team, 2019) to gauge whether targets of measurement that might impact quantity and quality influenced scores on the TIMECS scales. Our goal was to compare the proportion of variance accounted for by each factor to determine if the pattern differed across the scales. Variance component analysis partitions the total variance among scores into facets of measurement variance (i.e., group, teacher, child, time, coder). The nested design was accounted for in the ICC calculations using mixed-model procedures (see McLeod et al., 2018). A separate analysis was run for each TIMECS scale. Variance components were calculated using a mixed model with restricted maximum likelihood estimation for (a) group (BEST in CLASS, business as usual), (b) teacher (nested in group), (c) child (nested in group, teacher), (d) time (nested in child, teacher, group), and (e) coder. Each factor represents a possible source of variation in integrity scores. *Group* reflects the influence of the groups (BEST in CLASS, business as usual); *teacher* represents variability across teachers; *child* reflects variability across children; *time* reflects change over time (measured in weeks since first observation); *coder* reflects variability in coder ratings (tendency to score high or low). Variance estimates were transformed into proportions of variance based on estimates of the total variance. Our last set of analyses evaluated whether there were differences between BEST in CLASS and business as usual on the two TIMECS scales. Using the same linear mixed model, we estimated a coefficient representing the mean difference in scores between teachers in the two groups using the *lme4* package in *R* (Bates et al., 2019). We hypothesized that scores on the TIMECS BEST

in CLASS Adherence and Competence scales would evidence discriminative validity and be significantly higher in the BEST in CLASS group.

Results

Interrater Reliability: TIMECS Quantity and Quality Items

Table 2 reflects the descriptive statistics and ICC results for all items. Three Quantity items had a range below 3 points (Choices, Premack principle, Tangible reinforcement); no Quality items had a range below 3 points. For the Quantity items, interrater reliability for the items had a mean of .81 ($SD = .07$; range from .68 to .95.). All 21 items were in the “good” or “excellent” range using the standards described by Cicchetti (1994), with 18 of the 21 in the “excellent” range. For the Quality items, the mean interrater reliability was 0.69 ($SD = .08$; range from .52 to .80). Out of the 21 items, 17 of the 21 ICC values were in the “good” or better range, and none were below .50. Overall, these findings suggest that the full range was used for most items and that the interrater reliability for the items were “good” to “excellent”.

Construct Score Validity: TIMECS Quantity and Quality Item Scores

As seen in Table 3, the mean of the absolute value of the correlations between the corresponding Quantity and Quality items was $r = 0.42$ ($SD = 0.17$; range .06 to .73). Only Time Out was correlated above .70 ($r = .73$, $p < .01$). The mean of the absolute value of the correlations between the Quantity items and the OTCIS was 0.12 ($SD = .12$; range from .01 to .44), whereas the mean of the absolute value of the correlations between the Quality items and the OTCIS was .28 ($SD = .12$; range from .01 to .52). The mean correlation between Quantity and Quality items was significantly higher than the mean correlation between the Quantity items and the OTCIS ($z = 2.52$, $p < .001$), but was not significantly different from the mean correlation between the Quality items and the OTCIS ($z = 1.10$, $p = .27$). Overall, these findings provide mixed support for the construct validity of TIMECS item scores. The Quantity and Quality item scores appear to assess distinct integrity components. However, while the Quantity items do not overlap with an observational measure of the teacher–child interactions, the Quality items do.

Construct Score Validity: TIMECS Quantity and Quality Scale Scores

As seen in Table 4, the correlations among the TIMECS Quantity and Quality scales ranged from small to large. The strongest correlation was between the TIMECS Quantity and Quality scales ($r = .46$, $p < .001$), followed by the correlation

between the TIMECS Quality and OTCIS scales ($r = .44$, $p < .001$). The correlations between the TIMECS Quantity and Quality scales and the OTCIS were large, whereas the correlations between the TIMECS Quantity and Quality scales and the STRS Closeness and Conflict scales were small. The correlation between the TIMECS Quantity and Quality scales was significantly higher than the correlation between the TIMECS Quantity scale and the OTCIS ($z = 2.53$, $p < .01$), but the correlation between the TIMECS Quantity and Quality scales was not significantly different from the correlation between scores on the TIMECS Quality scale and the OTCIS ($z = .18$, $p = .86$). The correlation between the TIMECS Quantity and Quality scales was significantly higher than the correlation between the TIMECS Quantity and the STRS Closeness ($z = 2.37$, $p < .001$) and STRS Conflict ($z = 2.42$, $p < .001$) scales. Similarly, the correlation between the TIMECS Quantity and Quality scales was significantly higher than the correlation between the TIMECS Quality scale and the STRS Closeness ($z = 1.88$, $p < .05$) and STRS Conflict ($z = 1.92$, $p < .05$) scales. These findings indicate that the Quantity and Quality scales may assess different integrity components, but the discriminant validity of the Quality scale is not fully supported.

Discriminant Validity of TIMECS Quantity and Quality Scale Scores

Separate variance component analyses were run for the BEST in CLASS Adherence and Competence scales. All models accounted for variation within-child (residual), between-child, between-teacher, between-group, between-coder, and across time. As seen in Table 5, teacher accounted for a high proportion of variance in the BEST in CLASS Adherence and Competence scales (.19 and .48), with teacher accounting for more than twice the variance in the BEST in CLASS Competence scale. Group accounted for variation in the BEST in CLASS Adherence scale (.06), but group did not account for any meaningful variation in the BEST in CLASS Competence scale ($< .01$). Time (.08 and .06), child (.08 and .002), and coder (.07 and .04) accounted for a small amount of variation in each TIMECS BEST in CLASS Adherence and Competence scales. The proportion of residual variance for the BEST in CLASS Adherence and Competence scales within child was .52 and .41. Overall, these analyses suggest scores on the BEST in CLASS Adherence scale varied across groups and that teacher may account for a different proportion of variance in the BEST in CLASS Adherence and BEST in CLASS Competence scales. Using a linear mixed model with random effects for time, coder, teacher, and child, we tested whether scale scores differed according to group (BEST in CLASS vs. business as usual). As hypothesized, scores on the BEST in CLASS Adherence scale were significantly higher than business as usual

Table 3 Construct validity of TIMECS Quantity and Quality item scores

Item name	Quan-Qual	Quan-OTCIS	Qual-OTCIS
1. Social skills	.39***	.05	.32**
2. Emotion regulation	.41***	.01	.31**
3. Problem solving	.60***	.11**	.01
4. Promoting behavioral competence	.43***	.01	.37**
5. Teacher–child relationships	.56***	.40**	.47**
6. Rules	.50***	−.02	.10
7. Narrating	.41***	.04	.24**
8. Supporting listening	.48***	.44**	.52**
9. Choices	.43***	.08	.23**
10. Monitoring	.68***	.22*	.40**
11. Modeling	.43***	.04	.23**
12. Rehearsal	.09	.06	.25
13. Precorrection	.36**	.07	.23**
14. Opportunities to respond	.56***	.17*	.34**
15. Visual cueing	.38***	−.02	.18**
16. Premack principle	.50***	−.15	.17
17. Tangible reinforcement	.16	−.05	.27
18. Time-out	.73**	−.15	.42
19. Praise	.44***	.16	.21
20. Error correction	.06	−.08	.40**
21. Instructive feedback	.29***	.13	.23**
<i>M</i>	.42	.12	.28
<i>SD</i>	.17	.12	.12

TIMECS treatment integrity measure for early childhood classrooms, OTCIS observational teacher–child interaction scale. *Quan-Qual* correlation between quantity and quality item scores, *Quan-OTCIS* correlation between quantity items and OTCIS, *Qual-OTCIS* correlation between quality item scores and OTCIS

* $p < .05$; ** $p < .01$; *** $p < .001$

($\beta = .25$, $t = 2.51$, $p < .05$), but no significant difference was found for the BEST in CLASS Competence scale ($p > .05$). These findings support the discriminant validity of the BEST in CLASS Adherence scale.

Discussion

The goal of this study was to describe the development of the TIMECS and report initial score reliability and validity evidence for the items and scales. Analyses provide

Table 4 Construct score validity of TIMECS Quantity and Quality scale scores

Scale	2	3	4	5
1. Quantity	.46***	.20**	−.08	−.07
2. Quality		.44**	.15	−.14
3. OTCIS			.28*	−.24
4. STRS closeness				−.41**
5. STRS conflict				

TIMECS treatment integrity measure for early childhood classrooms, STRS student teacher relationship scale, OTCIS observational teacher child interaction scale

* $p < .05$; ** $p < .01$

initial support for the reliability of the TIMECS Quantity and Quality items. The findings generally support the score validity of the TIMECS items and scales: the pattern of correlations among the items and scales were in the expected direction, Quantity and Quality items and scales were distinct from each other, and the TIMECS Quantity and Quality scales were distinct from a teacher-report measure of the teacher–child relationship. However, scores on the TIMECS Quality and Quality items and scales were not consistently distinct from scores on an observational measure of teacher–child interactions. Our analyses also revealed that different factors accounted for systematic variation in the Quantity and Quality scale scores, and the Quantity scale differentiated between teachers trained in an EBI. Overall, the TIMECS appears to be a promising integrity measure that may contribute to efforts to evaluate and implement EBIs.

The TIMECS Quantity and Quality items were coded reliably by independent coders conducting live observations in early childhood settings. The interrater reliability for the Quantity and Quality items reported here compares favorably to a study in the early childhood literature that used an observational integrity measure (see Sutherland et al., 2014)

Table 5 Variance components for TIMECS scales

TIMECS scale	Variance components					
	Group	Teacher	Child	Time	Coder	Residual
BiC adherence	.06	.19	.08	.08	.07	.52
BiC competence	< .01	.48	.002	.06	.04	.41

TIMECS treatment integrity measure for early childhood settings, BiC BEST in CLASS

Variance component estimates represent the proportion of variance that is attributed to each source of variance

as well as studies from the mental health field that employed a similar observational integrity measurement approach (e.g., Hogue et al., 2008; McLeod et al., 2018). As with previous efforts to assess adherence and competence (e.g., Hogue et al., 2008; Sutherland et al., 2014), interrater reliability was lower for the Quality items than for the Quantity items. This may be due to the fact that quality ratings require coders to consider teacher and child behavior (i.e., responsiveness to child behavior), which may be more challenging to rate than just teacher behavior. If rating quality items requires more inference relative to quantity items this could have implications for training coders to use the TIMECS. That said, the mean ICCs for the Quality items were still in the good to excellent range. Thus, using the TIMECS the trained observers were able to reliably code Quantity and Quality items during live observations in early childhood settings. Importantly, the interrater reliability indicates that each TIMECS item can be used in isolation (e.g., to estimate quantity and quality of praise), or combined together to form scales, which supports using the TIMECS as a generic tool to estimate treatment integrity for practices or programs.

Psychometric Findings

In developing the TIMECS, our aim was to assess quantity (i.e., adherence) and quality (i.e., competence). To demonstrate that the TIMECS can assess these components, it is important to establish that the items designed to assess quantity and quality are distinct. Correlations range in strength from 0.38 (Sutherland et al., 2014) to .96 (Barber et al., 2003) in previous research that has used the same coder to rate quantity and quality, raising the question of whether the same coder can distinguish between these components. Our analyses indicated that the correlations between scores on the corresponding TIMECS Quantity and Quality items were large. However, only Time Out had a correlation above .70 (i.e., $r = .73$), suggesting that the Quantity and Quality items are not redundant ($r < .70$; Kline, 1979). Together, these findings suggest that each set of items assess different integrity components and that the same coder can be used to assess both treatment integrity components.

Our findings provide mixed support for the discriminant validity of the TIMECS item scores. At the item level, scores

on the TIMECS Quantity items evidenced discriminant score validity with an observational measure, but the Quality items did not evidence discriminant validity. The magnitude of the correlations between the TIMECS Quantity and Quality items and the observational measure of teacher–child interactions was consistent with previous research in the mental health field (e.g., Hogue et al., 2008; McLeod et al., 2018), though the correlations between the Quantity and Quality items are lower than typically seen in previous research (see e.g., Hogue et al., 2008; McLeod et al., 2018). Thus, findings support the discriminant validity of the Quantity items, but only provide mixed support for the Quality items.

The findings at the scale level were mixed as well. The pattern supported the discriminant score validity of the TIMECS Quantity scale, but the same pattern was not observed for scores on the Quality scale. The correlation between the TIMECS Quality scale and the observational teacher–child interaction scale was the same magnitude as the correlations between the TIMECS Quantity and Quality scales. Yet, the correlation between the TIMECS Quality scale and the self-report teacher–student relationship scales was significantly lower than the correlation between the TIMECS Quality scale and the observational teacher–child relationship scale. This suggests that shared method variance—the same coders rating the relationship and competence—may explain the large correlations between the Quality scale and the observational teacher–child interaction scale. Previous studies have found small to moderate correlations between scores on measures designed to assess quantity, quality, and the relationship (e.g., Hogue et al., 2008; McLeod et al., 2018). It will be important for future research to further investigate the discriminant validity of scores on the TIMECS Quality scale.

The TIMECS was developed as a generic measure. On the one hand, the content validity of the TIMECS items has been established (see McLeod et al., 2017), suggesting that the items map onto the content of existing EBIs. However, only the TIMECS Adherence scale discriminated between teachers who had previously received training in the BEST in CLASS intervention from teachers who did not. Sustainment of EBIs beyond implementation has been difficult to achieve in school settings (Locke et al., 2015), so it is plausible that the teachers who had been trained in the BEST in CLASS intervention may no longer be delivering the core

practices with competence. That said, this study provides some evidence that the TIMECS items can be mapped onto the content of a social and behavioral EBI and used to assess adherence. An important direction for future research is to evaluate if both the Quantity and Quality items can be used in this manner with specific practices and with different EBIs.

Our findings suggest that there was teacher to teacher variation in scores on the TIMECS BEST in CLASS Adherence and Competence scales. Compared to the Adherence scale, the teacher accounted for a higher proportion of variance in the Competence scale suggesting that the competence scores evidenced greater variation across teachers. That variation observed across teachers is consistent with integrity research that indicates providers vary in quality (e.g., McLeod et al., 2018). If teachers do consistently vary in their overall competence, then this has important implications for professional development in early childhood settings. The child also accounted for variation in the BEST in CLASS Adherence scale, suggesting that the quantity of practices delivered varies from child to child. It is possible that child behavior is influencing what practices teachers deliver. Time and coders accounted for small amounts of variation, suggesting scores may not vary across these factors. Our findings are consistent with research that has found various factors account for variation across quantity and quality scales (e.g., Hogue et al., 2008; McLeod et al., 2018) and indicate that these sources of variation may need to be considered in future TIMECS applications.

Implications, Limitations, and Conclusions

Given the potential applications of the TIMECS to intervention evaluation, professional development, and implementation research, it is important to consider how the observational procedures may need to change across different stakeholders. In considering how procedures may change, it is important to note that our coding team was composed of post-BA and master's level personnel, suggesting that individuals with a range of training can use the TIMECS. At the same time, early childhood programs may not have access to post-BA and masters' level personnel to collect these data, and the amount of time and resources allocated to training coders may not be feasible for professional development application or even large-scale implementation research. Future feasibility research that examines how best to train and support early childhood professionals to collect reliable TIMECS data that can be used for professional development and program evaluation would help advance evidence-based practice around social, emotional and behavioral learning in early childhood programs.

A few limitations bear mentioning. First, the TIMECS was designed to assess the delivery of practices to children at risk for EBDs and thus may not generalize to other populations. Second, the TIMECS was not able to discriminate

between quality of BEST in CLASS practices delivered by teachers trained in BEST in CLASS and teachers in business as usual. It is unclear whether this is due to a sustainment issue or a measurement issue and thus warrants further research. Finally, our ability to estimate variance components for teachers was somewhat limited given the number of children in each class.

The TIMECS was designed to characterize the implementation of practices found in EBIs that target social and behavioral skills in early childhood classrooms. The findings support the potential of the TIMECS to assess integrity of teacher-delivered practices. The development of a generic measure that can be used across existing EBIs for children at risk for EBDs that focus on child social and behavioral outcomes increases the utility of the measure, and holds promise for assessing not only the teacher delivery of EBIs in evaluation and implementation research but also characterizing business as usual classrooms in these same studies.

Funding This research was supported by a grant (R305A140487) from the US Department of Education, Institute for Education Sciences. Preparation of this article was supported in part by a grant from the Institute of Education Science (R305A140487; McLeod & Sutherland).

Declarations

Ethics Approval All study procedures involving human participants were in accordance with the ethical standards of the researchers' Institutional Review Boards and with the 1964 Helsinki declaration and its later amendments of comparable ethical standards. The institutional review board of Virginia Commonwealth University granted approval of all the work reported in this paper.

Informed Consent Informed consent and assent was obtained from all teachers, parents, from parents for their child's participants, and verbal assent from children.

Disclaimer The opinions expressed by the authors are not necessarily reflective of the position of or endorsed by the US Department of Education.

Conflict of Interest The authors declare that they have no conflict of interest.

References

- Abry, T., Rimm-Kaufman, S. E., Larsen, R. A., & Brewer, A. J. (2013). The influence of fidelity of implementation on teacher–student interaction quality in the context of a randomized controlled trial of the Responsive Classroom approach. *Journal of School Psychology, 51*, 437–453.
- Barber, J. P., Liese, B. S., & Abrams, M. J. (2003). Development of the cognitive therapy adherence and competence scale. *Psychotherapy Research, 13*, 205–221.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendick, G., Green, P., &

- Fox, J. (2019). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version, 1*, 1–21.
- Berry, V., Axford, N., Blower, S., Taylor, R. S., Edwards, R. T., Tobin, K., & Bywater, T. (2016). The effectiveness and micro-costing analysis of a universal, school-based, social-emotional learning programme in the UK: A cluster-randomised controlled trial. *School Mental Health, 8*, 238–256.
- Caron, E., Muggeo, M. A., Souer, H. R., Pella, J. E., & Ginsburg, G. S. (2019). Concordance between clinician, supervisor, and observer ratings of therapeutic competence in CBT and treatment as usual: Does clinician competence or supervisor session observation improve agreement? *Behavioural and Cognitive Psychotherapy, 48*, 350–363.
- Chapman, J. E., McCart, M. R., Letourneau, E. J., & Sheidow, A. J. (2013). Comparison of youth, caregiver, therapist, trained, and treatment expert raters of therapist adherence to a substance abuse treatment protocol. *Journal of Consulting and Clinical Psychology, 81*, 674–680.
- Chorpita, B. F., & Daleiden, E. L. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*, 566–579.
- Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.
- Collyer, H., Eisler, I., & Woolgar, M. (2019). Systematic literature review and meta-analysis of the relationship between adherence, competence and outcome in psychotherapy for children and adolescents. *European Child & Adolescent Psychiatry, 29*, 417–431.
- Conroy, M. A., Sutherland, K. S., Algina, J., Ladwig, C., Werch, B., & Gyure, M. (2019). Outcomes of the BEST in CLASS intervention on teachers' use of effective practices, self-efficacy, and classroom quality. *School Psychology Review, 48*, 31–45.
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*, 405–432.
- Fleiss, J. (1981). Balanced incomplete block designs for interrater reliability studies. *Applied Psychological Measurement, 5*, 105–112.
- Garland, A. F., Brookman-Frazee, L., Hurlburt, M. S., Accurso, E. C., Zoffness, R. J., Haine-Schlagel, R., & Ganger, W. (2010). Mental health care for children with disruptive behavior problems: A view inside therapists' offices. *Psychiatric Services, 61*, 788–795.
- Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationship and the trajectory of children's school outcome through eighth grade. *Child Development, 72*, 625–638.
- Hedges, L. V. (1994). Fixed effect models. In H. E. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 285–299). Russell Sage Foundation.
- Hemmeter, M. L., Snyder, P., & Fox, L. (2018). Using the Teaching pyramid observation tool (TPOT) to support implementation of social-emotional teaching practices. *School Mental Health, 10*, 202–213.
- Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy: Theory, Research, Practice, Training, 33*, 332–345.
- Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inlan, J., Reiner, R. H., & Liddle, H. A. S. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment, 35*, 137–147.
- Howes, C., & Ritchie, S. (1999). Attachment organizations in children with difficult life circumstances. *Development and Psychopathology, 11*, 251–268.
- Kline, P. (1979). *Psychometrics and psychology*. Academic Press.
- Locke, J., Olsen, A., Wideman, R., Downey, M. M., Kretzmann, M., Kasari, C., & Mandell, D. S. (2015). A tangled web: The challenges of implementing an evidence-based social engagement intervention for children with autism in urban public school settings. *Behavior Therapy, 46*, 54–67.
- McLeod, B. D., Southam-Gerow, M. A., Rodriguez, A., Quinoy, A., Arnold, C., Kendall, P. C., & Weisz, J. R. (2018). Development and initial psychometrics for a therapist competence instrument for CBT for youth anxiety. *Journal of Clinical Child and Adolescent Psychology, 47*, 47–60.
- McLeod, B. D., Sutherland, K. S., Broda, M., Granger, K., & Markowicz, K. (2021). Initial development of the Observational Teacher-Child Interaction Scale of early childhood settings. Manuscript in preparation.
- McLeod, B. D., Sutherland, K. S., Martinez, R. G., Conroy, M. A., Snyder, P. A., & Southam-Gerow, M. A. (2017). Identifying common practice elements to improve social, emotional, and behavioral outcomes of young children in early childhood classrooms. *Prevention Science, 18*(2), 204–213.
- Newborg, J. (2005). *Battelle developmental inventory, 2nd edition, Examiner's manual*. Riverside Publishing.
- Pianta, R. C., & Hamre, B. (2001). *Students, teachers, and relationship support (STARS)*. Lutz, FL: Psychological Assessment Resources.
- Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research, 38*, 65–76.
- R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. McGraw-Hill.
- Sanetti, L. M., Charbonneau, S., Knight, A., Cochrane, W. S., Kulcyk, M. C. M., & Kraus, K. E. (2020). Treatment fidelity reporting in intervention outcome studies in the school psychology literature from 2009 to 2016. *Psychology in the Schools, 57*, 901–922.
- Schuetzenmeister, A., & Dufey, F. (2019). R package 'VCA'. Downloaded from <https://CRAN.R-project.org/package=VCA>
- Schweinhart, L. J., & Weikart, D. P. (1997). The high/scope preschool curriculum comparison study through age 23. *Early Childhood Research Quarterly, 12*, 117–143.
- Sheridan, S. M., Rispoli, K. M., & Holmes, S. R. (2014). Treatment integrity in conjoint behavioral consultation: Active ingredients and potential pathways of influence. *Faculty Publications from CYFS, 96*.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420.
- Sutherland, K. S., Conroy, M., McLeod, B. D., Algina, J., & Wu, E. (2018). Teacher competence of delivery of BEST in CLASS as a mediator of treatment effects. *School Mental Health, 10*, 214–225.
- Sutherland, K. S., McLeod, B. D., Conroy, M., Abrams, L., & Smith, M. M. (2014). Preliminary psychometric properties of the Best in Class Adherence and Competence Scale. *Journal of Emotional and Behavioral Disorders, 22*(4), 249–259.
- Sutherland, K. S., McLeod, B. D., Conroy, M. A., & Cox, J. R. (2013). Measuring implementation of evidence-based programs targeting young children at risk for emotional/behavioral disorders conceptual issues and recommendations. *Journal of Early Intervention, 35*, 129–149.
- Sutherland, K. S., & Snyder, A. (2007). The effects of peer tutoring and self-graphing on the reading fluency and classroom behavior of middle school students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 15*, 103–118.
- Walker, H., Severson, H., & Feil, E. (1995). *Early Screening Project: A proven child find process, Examiner's manual*. Sopris West Publishing.

- Wanless, S. B., & Domitrovich, C. E. (2015). Readiness to implement school-based social-emotional learning interventions: Using research on factors related to implementation to maximize quality. *Prevention Science, 16*, 1037–1043.
- Webb, C. A., Derubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*, 200–211.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.