School Mental Health
https://doi.org/10.1007/s12310-021-09437-7

ORIGINAL PAPER

Examining the Correspondence Between Teacher- and Observer-Report Treatment Integrity Measures

Bryce D. McLeod[1] · Kevin S. Sutherland[2] · Michael Broda[2] · Kristen L. Granger[2] · Jennifer Cecilione[1] · Clayton R. Cook[3] · Maureen A. Conroy[4] · Patricia A. Snyder[4] · Michael A. Southam-Gerow[1]

Accepted: 26 February 2021

Abstract

Teacher-reported measures of treatment integrity (the extent to which prescribed practices are delivered as intended by teachers) have the potential to support efforts to evaluate and implement evidence-based interventions in early childhood settings. However, self-report treatment integrity measures have shown poor correspondence with observer-report treatment integrity measures, raising questions about score validity. This paper reports on the development and initial evaluation of the score reliability and validity of the Treatment Integrity Measure for Early Childhood Settings Teacher Report (TIMECS-TR), which is designed to address limitations of previous self-report treatment integrity measures that may have contributed to low correspondence with observer-rated measures. The TIMECS-TR includes 24 items designed to represent practices found in evidence-based interventions delivered in early childhood settings that target child social, emotional, and behavioral skills, rather than adherence to practices found in a specific evidence-based intervention. Fifty-four teachers (92.6% female, 7.4% male; 61.1% White) completed the TIMECS-TR weekly for a total of 618 times ($M = 6.79$ per child; SD = 2.16; range 2 to 11) about the practices they delivered with 91 children (45.1% female, 54.9% male; $M = 4.53$ years old; SD = 45.1% Black) who were at risk for emotional and behavioral disorders. Analyses indicated that the TIMECS-TR items evidenced mild to moderate test–retest score reliability over one week. However, analyses did not support the convergent score validity of the TIMECS-TR items or scale with observational ratings of the same practices. Teachers reported higher levels of practice delivery on the TIMECS-TR items relative to observer report. Overall, our findings raise concerns about the accuracy of teacher-report adherence measures. Lessons from this research can be used to identify possible reasons for the low correspondence between teacher- and observer-report treatment integrity measures so that future research can strive to dependably capture teacher delivery of the practices found in evidence-based interventions.

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

Introduction

In education research, treatment integrity measures are designed to gauge the extent to which prescribed or selected practices are delivered as intended by teachers. Though definitions vary, four components of treatment integrity are often emphasized in the education literature: adherence, dosage, responsiveness, and competence (Sanetti et al., 2020; Sutherland et al., 2013). Adherence refers to the extent to which a practice, such as behavior-specific praise (Sutherland et al., 2000), is delivered according to an established protocol. Dosage focuses on how much exposure to the practices a child gets (Sanetti et al., 2020). Responsiveness refers to the extent to which children understand and engage in intervention activities (Bellg et al., 2004). Competence

Preparation of this article was supported in part by a grant from the Institute of Education Sciences (R305A140487; McLeod & Sutherland).

✉ 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, USA

[3] Department of Educational Psychology and Organizational Leadership and Policy Development, University of Minnesota, Minneapolis, USA

[4] Anita Zucker Center for Excellence in Early Childhood Studies, University of Florida, Gainesville, USA

Published online: 13 March 2021

refers to the responsiveness and skill demonstrated by a teacher when delivering a practice specified in a treatment protocol. Measurement of treatment integrity is critical for efforts to evaluate and implement individual practices and evidence-based interventions (EBIs; Sanetti et al., 2020) not only in K-12 education settings, but also in early care and education settings.

Until relatively recently, few studies reported on or adequately assessed treatment integrity in the mental health or school psychology fields (e.g., Perepletchikova et al., 2007; Sanetti et al., 2011). It appears, however, that this gap is beginning to be addressed. A recent review of the school psychology literature concluded that more studies include an assessment of treatment integrity (Sanetti et al., 2020). Representing early childhood to high school samples, Sanetti et al., (2020) noted that their review of the literature indicated approximately 75% of studies reported on treatment integrity, with most (65%) of the treatment integrity measurement focusing on adherence.

Despite the importance of assessing treatment integrity, barriers exist to utilizing treatment integrity measures in early care and education settings outside of funded research projects. Observational measures are the most common data collection method for treatment integrity data (Sanetti et al., 2020) and are generally considered the "gold standard" in treatment integrity research (Sanetti et al., 2020; Sutherland et al., 2013). However, there are several features of observer-rated measures that limit their feasibility (i.e., extent to which a measure can be successfully used in a setting; Proctor et al., 2011) and usability (i.e., extent to which a measure can be used to achieve specific goals; Lyon & Koerner, 2016). First, observer-rated measures are costly and time intensive to use (Hogue et al., 2014; Schoenwald et al., 2011). Second, observer-rated measures do not provide easy access to practices that are delivered infrequently, as most observations last roughly 30–60 min in duration (McLeod et al., 2009). Finally, stakeholders (e.g., teachers, administrators) do not always support observational assessment due in part to their perceived intrusiveness (Yoder et al., 2018). For these reasons, pragmatic treatment integrity measures (i.e., brief, easy to use, and valid; Stanick et al., 2019) are needed for use by stakeholders and researchers in early care and education settings.

Development of teacher-report integrity measures might address some of these concerns (Hogue et al., 2014). Compared to other commonly used data collection methods—i.e., observational and permanent product review (see Sanetti et al., 2020)—teacher self-report measures have the potential to be more cost-effective and easier to use. Self-report treatment integrity measures can also be used to support the continuous improvement in the delivery of individual practices and EBIs by facilitating ongoing collection and monitoring of delivery of practices (Connors et al., 2020; Hogue et al., 2013). In mental health research, the potential of this approach is exemplified by research conducted with Multi-Systemic Therapy for youth offenders. To illustrate, scores generated on self-report adherence measures have been linked to improved clinical outcomes and used as a feedback tool in the training and supervision of mental health providers (e.g., Henggeler et al., 1999; Schoenwald et al., 2000). Moreover, emerging evidence suggests that self-report tools might provide feasible, consistent, and meaningful ways of gathering data about teacher-delivered instructional and behavioral practices. For example, the Classroom Strategies Scale—Teacher Form, a teacher-report measure of classroom strategies used in elementary schools, has shown promising score reliability and validity (Reddy et al., 2015, 2016).

Teacher-report measures that assess adherence have particular benefits for intervention evaluation and implementation research (Proctor et al., 2011; Sutherland et al., 2013). Each treatment integrity component provides valuable information about different facets of intervention delivery. That said, adherence data provides critical information relevant to the evaluation and implementation of EBIs. When an intervention outperforms a control group, the measurement of adherence allows researchers to determine if the effect can be attributed to the intervention (i.e., if adherence is high then the effect is likely due to the intervention). In contrast, when an intervention does not produce significant effects, adherence measurement can help researchers interpret the null findings (i.e., if adherence is low then the lack of an effect may be due to poor delivery, but if adherence is high then the intervention may not work; Schoenwald et al., 2011). Measurement of adherence also informs implementation research, which focuses on gauging the success of training and coaching supports on teacher behavior change (Proctor et al., 2011). Thus, teacher-report measures that assess adherence has the potential to benefit efforts to evaluate and implement interventions in early care and education settings.

Though teacher-report measures have the potential to overcome existing barriers and support intervention evaluation and implementation, several advances are needed to produce technically sound self-report adherence measures. First, concerns about the dependability of self-report adherence measures need to be addressed. Self-report adherence measures have traditionally evidenced weak correspondence with observer-rated measures (see e.g., Caron et al., 2019; Chapman et al., 2013; Hurlburt et al., 2010). Given the questions about convergent score validity, it is important to evaluate the degree of overlap between scores on self- and observer-rated adherence measures. Observer-rated adherence measures with score reliability and validity evidence are needed to develop teacher-report measures (McLeod et al., 2009), so that research can examine whether discrepancies between observer- and self-report ratings of

adherence represent meaningful differences and not just measurement error.

Though conventional wisdom may view that the discrepancies represent measurement error in self-report measures, recent research pushes back against the notion that scores on self-report adherence measures are incapable of being reliable and valid (Hogue et al., 2017; Sanetti & Collier-Meek, 2019). For example, research suggests that it may be possible to improve the accuracy of self-report integrity measures by training teachers to self-report on their delivery of practices (e.g., Dart et al., 2020; Fallon et al., 2018). Training teachers to self-report on the practices used to address the social and behavioral skills of children may thus help address problems with correspondence.

Another potential source of variance in self-report adherence measures involves item design (e.g., scaling differences) and wording. To date, most self-report integrity tools have not been designed specifically for (or by) end-users. For example, most items are written by researchers so there may be a mismatch between how items are written and how the items are interpreted by end-users (Haynes et al., 1995; Ware et al., 2003). Using mixed method approaches to ensure that the item design and wording is acceptable to end-users may therefore increase data accuracy of self-report adherence measures.

A final issue is related to the utility (i.e., usefulness of a measure to stakeholders) of self-report adherence measures. To date, most adherence measures are tied to a specific purveyor-based EBI model. Schools sometimes implement more than one EBI, in which case they would need to deploy multiple adherence measures for adequate coverage. By developing items for adherence measures that represent practice elements (i.e., "discrete clinical technique used as part of a larger intervention plan"; Chorpita & Daleiden, 2009, p. 560), as opposed to practices associated with a specific EBI protocol, a more flexible adherence measure can be produced for use with multiple EBIs as well as develop an understanding of typical or usual care to identify specific practice elements that are missing (Hogue et al., 2019; McLeod et al., 2013).

The purpose of this study is to describe the development and initial psychometric evaluation of the Treatment Integrity Measure for Early Childhood Settings Teacher Report (TIMECS-TR; Sutherland & McLeod). The TIMECS-TR was designed to assess the quantity (i.e., extensiveness or adherence) of teacher-delivered practices in early childhood classrooms. In developing the TIMECS-TR, our goal was to produce a teacher-report measure that could be used to support the evaluation and implementation of practices and EBIs delivered in early childhood settings. To achieve this goal, we developed items that (a) capture practices found across EBIs designed for young children with social, emotional, and behavioral challenges (McLeod et al., 2017) and (b) assess the quantity (i.e., adherence) with which the practices were delivered. To our knowledge, the TIMECS-TR is the first "generic" teacher-report measure in the early care or education literature designed to assess practice elements, as opposed to the practices that are found in a particular EBI protocol.

To evaluate the initial psychometric properties of the TIMECS-TR, teachers in early childhood classrooms serving children ages 3–5 years old were asked to fill out the TIMECS-TR based on their delivery of practices for children within their classrooms deemed to be at risk for social–emotional and behavioral challenges. A multicomponent measurement model (Carroll et al., 2000; Hogue, 2002; McLeod et al., 2013) used in previous treatment integrity research that included multiple constructs (quantity [adherence], quality [competence], teacher–child relationship) and methods (observer report and self-report) was used to assess the score reliability and validity of the TIMECS-TR. We evaluated the score reliability (test, retest) and validity (construct validity) of the TIMECS-TR items and scale. The main focus of our analyses was the correspondence between the TIMECS-TR and an observer-rated treatment integrity measure called the TIMECS that has demonstrated initial score reliability and validity (McLeod et al., 2020). The TIMECS-TR items parallel the content of TIMECS, which assesses both quantity (adherence) and quality (competence) of practice delivery. In light of the steps we took to address potential limitations of existing self-report adherence measures (i.e., improving item design, training teachers to self-report), we hypothesized that scores on the TIMECS-TR items and scale would demonstrate (a) evidence of convergent validity with the TIMECS Quantity scale and (b) evidence of discriminant validity with distinct domains (quality and teacher–child relationships). We also hypothesized that this would be consistent with patterns of associations seen in previous studies (e.g., Carroll et al., 2000; Hogue et al., 2008; Sutherland et al., 2014).

Method

Participants and Settings

Children and their teachers in early childhood classrooms located within urban and suburban communities in a Southeastern state participated in the study. Each classroom had an average of 17.26 (SD = 3.54) children and 2.09 (SD = 0.29) adults. Our goal was to assess teacher delivery of practices targeted at children at risk for EBD during the school year (i.e., *focal* children). To identify children at risk for EBD, multiple measures were used to screen child participants for inclusion. 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 were approached in October and asked to identify up to five 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* for those children with caregiver consent to confirm risk for EBD. Last, the *Battelle Developmental Inventory, Second Edition Screener* (Newborg, 2005) was used for each selected child; if a child demonstrated average or above average scores on the cognitive domain, they were retained. Children with the two most extreme scores on the *ESP* were retained in the sample as the focal children.

The screening process resulted in the inclusion of 54 teacher and 91 child participants. Twenty-one teachers were not included because caregiver consents for the focal children were not returned, children in their classroom did not qualify for the study, or because they withdrew from the study. Withdrawal reasons included expressing that they were no longer interested in the study, going on medical leave, and being too overwhelmed by other responsibilities to participate. Some teachers withdrew without providing a reason. Seven children withdrew during the study. Reasons for withdrawal included teacher withdrawal from the study, caregiver no longer being interested, unable to obtain pre-test measures from the child, chronic absences, and children moving out of the classroom.

Teacher Participants

The 54 teacher participants had the following demographic characteristics: 94.4% 26–55+ years old, 7.4% male, 92.6% female, 61.1% White, 35.2% Black, and 3.7% multiracial. The teachers had the following educational backgrounds: 38.9% Bachelors degree, 48.1% Master's degree, and 13.0% other degree, and the teachers averaged 7.69 (SD = 7.98) years teaching in early childhood classrooms. Seventeen of the 54 teachers had previously received training and practice-based coaching (Snyder et al., 2015) in BEST in CLASS (Sutherland et al., 2018), a manualized Tier-2 program that targets the reduction of problem behaviors demonstrated by young children at risk for EBD (see Sutherland et al., 2020). See Table 1 for information about the teacher participants.

Child Participants

Teachers self-reported on the practices used with 91 children (54.9% male, 25.3% female, 19.8% not reported) who were on average 4.31 years old (SD = 0.67) and identified as: 45.1% Black, 8.8% White, 1.1% Native American/American Indian, 1.1% Asian/Pacific Islander, 5.5% multiracial/other, and 7.7% Latino. 38.4% did not report on race/ethnicity.

Table 1 Teacher sample demographic and education data

Variable	*M* (SD) or %
Sex	
Male	7.4%
Female	92.6%
Age	
18–25	1.9%
26–35	40.7%
36–45	16.7%
46–55	22.2%
> 55	14.8%
Prefer not to answer	3.7%
Race	
Black	35.2%
White	61.1%
Multiracial/Other	3.7%
Highest level education	
High School Diploma	3.7%
Bachelor's Degree	38.9%
Education specialist	0
Associates Degree	3.7%
Master's Degree	48.1%
Doctoral Degree	1.9%
Other	3.7%
Years teaching	12.99 (9.50)
Years teaching preschool	7.69 (7.98)

Development of the Treatment Integrity Measure for Early Childhood Classrooms Teacher Report

The TIMECS-TR is a 24-item measure designed to assess the quantity of teacher-delivered practices that foster social, emotional, and behavioral skills for children at risk for EBD in early childhood settings. The TIMECS-TR was developed via a three-step process.

Step 1: Item Development

Our main aim was to produce a measure to assess the quantity (adherence and dosage) of practices found across EBIs designed for early childhood classrooms that target child social, emotional, and behavioral skills, rather than adherence to the practices found in a particular EBI. We thus performed a systematic review of practices, interventions, and EBIs designed to foster positive social–emotional learning outcomes that had been evaluated in single-case experimental designs, quasi-experimental designs, and randomized group designs. In all, 50 published articles were identified and an iterative process was used to distill the practice elements (i.e., "discrete clinical technique used as part of a larger intervention plan"; Chorpita & Daleiden,

2009, p. 560) from the practices and interventions evaluated with each of the articles. Experts in the early childhood field reviewed the practice elements. In all, 24 practice elements were identified. The complete process of defining the items is detailed in McLeod et al., (2017). Once the 24 practice elements were identified, definitions were written for each practice (see Table 2 for item definitions).

To ensure the TIMECS-TR item design and wording were acceptable to end-users, a target population review of each item was obtained from eight teachers designed to maximize the match between how items were designed and how they were interpreted. The teachers were recruited from federally funded or state-funded early childhood programs to participate in a cognitive interview and complete a draft version

Table 2 TIMECS-TR item names, definitions, scores, and distribution

Item name	Definition	*M* (SD)	Range	Skewness	Kurtosis
Social skills	Teacher provides instruction on strategies that can facilitate positive social interactions with peers or adults	3.81 (0.91)	4	−0.63	0.25
Emotion regulation	Teacher provides instruction focused on helping to identify, label, or regulate emotions	3.57 (1.03)	4	−0.51	−0.21
Problem solving	Teacher provides instruction designed to generate solutions to social, behavioral, or preacademic problems	3.63 (0.94)	4	−0.47	−0.16
Promoting behavioral competence	Instruction that focuses on promoting positive behavior (e.g., engagement) during instructional activities	4.44 (0.80)	4	−1.63	2.80
Teacher–child relationship	Teacher behavior that conveys warmth, closeness, and interest when listening to and interacting	4.47 (0.64)	3	−0.88	0.05
Rules	Teacher uses prescribed guidelines to teach the rules and behavioral expectations of the classroom	4.42 (0.81)	4	−1.51	2.10
Narrating	Teacher provides a verbal description of behavior	3.89 (0.90)	4	−0.40	−0.54
Supportive listening	Teacher actively demonstrates understanding of the topic	4.22 (0.78)	3	−0.77	0.12
Choices	Teacher provides an opportunity to select between two or more options related to instructional activities	3.69 (1.02)	4	−0.51	−0.36
Monitoring	Teacher actively monitors	4.35 (0.79)	3	−0.94	0.01
Modeling	Teacher demonstrates, or has a peer demonstrate, a behavioral or preacademic skill to promote learning	3.56 (1.02)	4	−0.44	−0.29
Rehearsal	Teacher encourages practice of a behavioral skill (e.g., during interactions with peers)	3.53 (0.99)	4	−0.32	−0.43
Precorrection	Teacher uses prompts prior to the occurrence of a behavior to remind of appropriate behavior and correct responding	4.07 (0.90)	4	−0.74	−0.04
Opportunities to respond	Teacher uses prompts (i.e., gestural, verbal, visual) that seek an active, observable, and specific response	4.02 (0.95)	4	−0.72	−0.16
Visual cueing	Teacher uses visual cues to prompt for appropriate behavioral responses or consequence	4.11 (0.94)	4	−0.88	0.02
Premack principle	Teacher uses a more reinforcing behavior (e.g., playtime) to reinforce less probable behaviors (e.g., lesson time)	3.48 (1.09)	4	−0.38	−0.47
Tangible reward	Teacher gives a tangible reward in response to an appropriate social, emotional, or behavioral response	3.28 (1.33)	4	−0.30	−1.06
Time-out	Teacher removes a child from a preferred activity for a specified period of time following a problem behavior	2.86 (1.27)	4	0.07	−1.06
Praise	Teacher provides positive verbal statements in response to an appropriate social, emotional, or behavioral response	4.26 (0.73)	4	−0.90	1.35
Differential reinforcement	Providing attention or praise to other children in order to remind a child of the behavioral expectation	4.18 (0.82)	4	−0.10	1.10
Ignoring	Teacher ignores undesirable behaviors	2.97 (1.22)	4	−0.09	−0.91
Error correction	Teacher provides corrective feedback following an incorrect response or undesirable behavior	3.78 (1.08)	4	−0.65	−0.29
Instructive feedback	Teacher provides extra instructional information while responding to correct response or appropriate behavior	3.82 (1.01)	4	−0.69	0.02
Scaffolding	Using an instructional method designed to facilitate learning that is appropriate for the child's developmental level	3.09 (1.09)	4	−0.21	−0.64

TIMECS-TR Treatment Integrity Measure for Early Childhood Classrooms Teacher Report

of the TIMECS-TR. Teachers averaged 4 years of experience as early childhood teachers (SD = 4.00). The interviews solicited information on how teachers understood each item, recalled relevant information, and formulated answers to improve the quality of data collected from the TIMECS-TR. The first and second authors conducted the interviews. A close-ended, "respondent debriefing" approach was employed that has been used to develop measures for mental health services (see Haynes et al., 1995; Ware et al., 2003). This approach is designed to identify problematic areas of questionnaires by determining whether the questions, definitions, and instructions convey the intended message. Items and response options that were not clear or failed to convey the intended message were revised to ensure the items had sufficient specificity and appropriate grammatical structure. This process was done iteratively, taking each respondent's suggestions into account prior to the next interview.

Step 2: Scoring Strategy

The TIMECS-TR items were incorporated into a rating form designed to provide information on the extent to which each practice was used in the past week. Since the current study focused on teacher report of practices delivered with focal children, the teachers were asked to report on the practices used with a specific focal child over the past week. The items were scored on a 5-point scale with the following anchors: 1 = *not at all,* 3 = *some*, and 5 = *a lot*. This scoring strategy has been used in previous self-report treatment integrity measures (see Henggeler & Borduin, 1992; Hogue et al., 2014).

Step 3: Training Manual Development

We developed a training manual to help improve teachers' accuracy in reporting on their own behavior. The training manual includes item descriptions and scoring guidelines in a concise, user-friendly format (i.e., 8 pages). The training manual was distributed to four teachers who were asked to provide feedback on the scoring instructions and item descriptions. The feedback was used to produce a final version of the training manual.

Measures for Validity Analyses

Treatment Integrity Measure for Early Childhood Classrooms (TIMECS; Sutherland & McLeod, 2015a, 2015b) is an observer-rated measure designed to assess the quantity and quality of teacher-delivered practices used with focal children in early childhood classrooms. The TIMECS-TR was designed to parallel the content of the TIMECS. The TIMECS consists of 21 items that could be used in an observer-rated system. Each TIMECS item is rated for quantity (adherence) and quality (competence). The quantity ratings involve extensiveness ratings (see Hogue et al., 1996), which requires coders to estimate the extent to which teachers engage in each practice during an observation using a 7-point Likert-type extensiveness scale with the following anchors: 1 = *not at all,* 3 = *somewhat,* 5 = *considerably,* and 7 = *extensively*. Extensiveness ratings are comprised of two components: thoroughness and frequency. Thoroughness refers to the depth, complexity, or persistence with which the teacher engages in each practice element. Frequency refers to the number of times throughout an observation that a given practice is delivered regardless of the thoroughness of the practice. Both thoroughness and frequency are considered in making an extensiveness rating on each item; thus, extensiveness ratings provide an estimate of dosage for each practice. For competence ratings, 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*. For each item, coders are asked to consider the extent to which a teacher demonstrated the following dimensions: (a) expertise, commitment, motivation; (b) clarity of language; (c) appropriate timing of a practice (responsiveness); and (d) ability to read and respond to where the child appears to be (responsiveness). TIMECS items are combined to create Quantity and Quality scales that are generated by averaging together the quantity or quality items. Scores on the TIMECS Quantity and Quality items and scales showed evidence of score validity, with the magnitude of the correlations suggesting that the quantity and quality items assess distinct components of treatment integrity (i.e., quantity and quality; McLeod et al., 2020). In the current study, inter-rater reliability (ICC[2,2]) for the quantity item scores had a mean of 0.81 (SD = 0.07; range from 0.68 to 0.95), whereas the quality item scores had a mean of 0.69 (SD = 0.08; range from 0.52 to 0.80).

Student Teacher Relationship Scale (STRS; Pianta & Hamre, 2001). The STRS assesses teacher perceptions of relationships with children and contains 15 items measured on a 5-point Likert-type scale: 1 = *definitely does not apply* and 5 = *definitely applies*. The STRS, which has two subscales, Closeness and Conflict, has demonstrated score validity in preschool through the elementary grades by predicting academic and social functioning (Hamre & Pianta, 2001; Pianta et al., 2002). The STRS has been used widely in studies of preschool and elementary-age children (e.g., Birch & Ladd, 1997, 1998; Howes & Richie 1999) and has been found to demonstrate score validity with children who are under-represented (Hamre & Pianta, 2001). The internal consistency for both subscales was acceptable (Cronbach's $\alpha = 0.78$ and 0.89 for closeness and conflict, respectively) for the current sample.

Observational Teacher–Child Interaction Scale (OTCIS; McLeod & Sutherland, 2015). The six items on the OTCIS

assess affective and interactional aspects of teacher–child exchanges. Trained observers assess teacher–child interactions and rate each item on a 6-point scale (0 = *not at all* to 5 = *a great deal*). The OTCIS has evidenced promising score reliability and validity (see McLeod et al., 2021). The inter-rater score reliability for the current study was ICC(2,2) = 0.81, and internal consistency (Cronbach's alpha) score reliability was $\alpha = 0.82$.

Procedures for the Treatment Integrity Measures

Data collection for the treatment integrity measures began in November. Our goal was to sample 12 self-report (TIMECS-TR) and 12 observer-rated (TIMECS) treatment integrity measurements for each focal child between November and April, as this sampling plan had produced reliable integrity data in a previous study (Sutherland et al., 2014).

TIMECS-TR Procedures

At the beginning of the study, each teacher was provided the TIMECS-TR training manual. They were asked to review the training manual and given the opportunity to ask questions about specific items or rating procedures. Once data collection began, each teacher was sent the TIMECS-TR via e-mail to fill out for each focal child each week on Friday based on the practices they delivered with a specific focal child over the past week. The TIMECS-TR was administered through REDCap; e-mails were sent to the teachers that contained a unique link for each child that led to an electronic copy of the TIMECS-TR that could be filled out on computer, tablet, or phone. Each e-mail had a list of instructions for filling out the TIMECS-TR for each focal child as well as a link to the TIMECS-TR training manual. If the TIMECS-TR was not filled out by the following Monday, a reminder was sent to the teacher. Teachers were given until the following Thursday to fill out the TIMECS-TR. Teachers were sent the TIMECS-TR weekly up to 12 times. A total of 654 TIMECS-TRs were sent and 618 were returned by the 54 teachers. On average, teachers filled out 6.79 TIMECS-TR per child (SD = 2.16; range 2 to 11).

TIMECS and OTCIS Procedures

The following procedures were used to generate scores on the TIMECS and OTCIS.

Coders. Four doctoral students in clinical psychology, three doctoral students in education, and five data staff were trained to use the TIMECS and OTCIS. The coders were on average 25.92 years old (SD = 3.90) and identified as the following: 17.0% male, 83% female; 84.0% White, 8.0% Asian; 8.0% multiracial, and 25% Latinx.

Coder Training. Training progressed through four steps over a 2-month period. Coders first received didactic instruction and discussion of the scoring manual, reviewed recordings of early childhood classrooms with the trainers, and engaged in 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. Third, coders independently conducted live coding in early childhood classrooms. Finally, coders scored 40 recordings, and reliability for each coder was assessed against master codes produced by the first and second authors. Once coders met "good" reliability on each item (i.e., ICC[2,2] > 0.59; Cicchetti, 1994), independent coding commenced. Regular reliability assessments were performed, and the results were discussed in weekly meetings to prevent coder drift (Margolin et al., 1998).

Coding Assignment Plan. Coders were assigned to observations using a balanced incomplete block design (Fleiss, 1981), stratifying for classroom and time. Study staff aimed to schedule up to 12 observations for each focal child. Observations were scheduled during the week at a convenient time for the teacher. Two observers were sent to each class and were instructed to conduct separate observations for each focal child in the classroom. If multiple focal children were in a class, the order of the observations was determined at random. The observers were instructed to sit in a discrete location in the class and not interact with the teacher, teacher assistant, children, or each other. Observations occurred during a teacher-led instructional (e.g., circle time, small group, story time) activity, a child-led activity (e.g., center time), or transitions. The observations could be comprised of more than one instructional context; however, most of the observation (e.g., at least 30 min) had to be comprised of teacher-directed instructional time where the teacher was engaged with the focal child or a group that the focal child was part of (i.e., a group that the focal child part of and is the target of an item). Only observations that were at least 30 min were retained. Observations less than 30 min were discarded. For example, if the child left the room for 30 min or more. Observations were an average of 40.56 (SD = 11.08) minutes. Coders took notes throughout each observation. At the end of each observation, the coders scored the TIMECS and OTCIS.

Results

Our analyses progressed through five steps designed to investigate the performance of the TIMECS-TR items (i.e., descriptive statistics) and the possible impact of missing data, reliability of the TIMECS items (i.e., test–retest), construct validity of the item scores, inter-informant agreement

of the teacher - and observer-report items, and the construct validity of the TIMECS-TR scale.

Item Performance

We first investigated patterns of missingness in the data, comparing the sample of 654 TIMECS-TRs that were sent to teachers to the 618 TIMECS-TRs that were returned. Using Little's MCAR Test (Little, 1988), we tested whether the missing data could be considered missing completely at random (MCAR) and therefore ignorable. Results of this test failed to reject the null hypothesis that the data can be considered MCAR ($\chi^2[1] = 0.42$, $p = 0.52$). This suggests that the missing data mechanism can be treated as ignorable. Descriptives for each item were evaluated to determine if the items displayed the full range of scores (i.e., 4 points or 1 to 5) and were normally distributed. Table 2 displays the mean, standard deviation, kurtosis, skewness, and range of each item. All items displayed a 3- or 4-point range and acceptable skewness and kurtosis (−2 to +2; Trochim & Donnelly, 2006), except for *Promoting Behavioral Competence* ($M = 4.44$; SD = 0.80; Skewness = −1.63; Kurtosis = 2.80) and *Rules* ($M = 4.42$; SD = 0.81; Skewness = −1.51; Kurtosis = 2.10).

Test–Retest Score Reliability

To estimate the test–retest score reliability for the TIMECS-TR items, two approaches were used. First, we calculated a multilevel standardized coefficient (equivalent to a correlation, adjusting for repeated observations within teacher) using a two-level multilevel model, with observations nested within children and teachers. For this model, the TIMECS-TR item score in the current time point was the outcome, and the lagged (1-week prior) score was the predictor. This resulted in observations from 90 children and 53 teachers. These raw correlations were then divided by the observed standard deviation to produce the equivalent of a test–retest correlation. The mean correlation across all items was $r = 0.16$ (SD = 0.08), which means that on average, items shared about 3.0% of variance across time points. Items with the highest correlations included *Rules* ($r = 0.25$, $r^2 = 0.06$), *Modeling* ($r = 0.26$, $r^2 = 0.07$), *Tangible Rewards* ($r = 0.35$, $r^2 = 0.12$), and *Ignoring* ($r = 0.27$, $r^2 = 0.07$). Table 3 presents correlations and shared variance estimates for the items.

The second approach to estimating test–retest score reliability was to use intraclass correlations (ICCs). These were again based on a multilevel model, with observations nested within teacher. The ICC represents the extent to which observations vary within teacher versus between teacher, with higher values between 0 and 1 representing higher within-person reliability. Table 3 presents ICC estimates by item. The mean ICC was 0.38 (SD = 0.08) and ranged from a low of 0.26 for *Problem Solving* to a high of 0.57 for *Tangible Reward*. Overall, these analyses suggest that individual items evidence small to moderate test–retest score reliability.

Construct Validity: TIMECS-TR Item Scores

We created a sample of "matched" observations to facilitate comparisons between the TIMECS-TR and TIMECS. A TIMECS-TR and TIMECS were considered "matched" if the TIMECS observation occurred up to five business days prior to the TIMECS-TR, since teachers were asked to report on their practices over a five-day time period. This resulted in a total of 396 matched observations. We investigated potential patterns of missingness in the data, comparing the original sample of 618 teacher reports to the reduced (matched) sample of 396 observations that were conducted on average 3.59 days apart (SD = 2.08). Using Little's MCAR Test (Little 1986), we tested whether the missing data could be considered missing completely at random (MCAR) and therefore ignorable. Results of this test failed to reject the null hypothesis that the data can be considered MCAR ($\chi^2(1) = 0.25$, $p = 0.62$). This suggests that the missing data mechanism can be treated as ignorable and facilitates valid comparisons between the full and matched sample.

To estimate the convergence between TIMECS-TR and TIMECS items, we used a multilevel model, with observations nested within child. A separate model was run for each item, with the TIMECS-TR item as the outcome, and the corresponding TIMECS Quantity and Quality item scores as predictors. The full results are in Table 3. Three TIMECS items were found to significantly predict their TIMECS-TR counterpart. The score on the TIMECS Quality *Social Skills* item negatively predicted the score on the TIMECS-TR *Social Skills* item (−0.28, $p = 0.003$), and the score on the TIMECS Quantity *Precorrection* item positively predicted the score on the TIMECS-TR *Precorrection* item (0.37, $p = 0.01$). Finally, the score on the TIMECS Quality *Tangible Reward* item positively predicted the score on the TIMECS-TR *Tangible Reward* item (0.71, $p < 0.001$). Overall, these findings do not support the construct validity of the TIMECS-TR item scores as few TIMECS-TR items were significantly related to the corresponding TIMECS item scores.

Inter-Informant Agreement

The TIMECS-TR item analyses showed little correspondence with the corresponding TIMECS Quantity items so we calculated two indices of inter-informant agreement to identify (a) discrepant and non-discrepant TIMECS-TR items, and (b) the direction of the disagreement for discrepant items (see De Los Reyes et al., 2019). We first calculated two estimates of single-item inter-informant agreement: r_{wg}

Table 3 TIMECS-TR item reliability and validity

Item name	Test–retest reliability			Convergent validity		Inter-rater agreement		
	Corr	Pseudo R squared	ICC	TIMECS Quantity	TIMECS Quality	r_{wg}	a_{wg}	SMD
Social skills	.07	< .01	.29	− .02	− .28*	.19	− .13	1.80
Emotion regulation	.08	.01	.34	.00	− .15	.20	− .45	1.95
Problem solving	.18*	.03	.26	− .51	.05	.12	− .58	2.56
Promoting behavioral competence	.08*	.01	.34	.02	.01	.63	.85*	.27
Rules	.25*	.06	.33	.08	− .14	.09	− .89	2.59
Teacher–child relationship	.13*	.02	.41	− .02	.02	.31	− .27	1.34
Narrating	.04	< .01	.43	− .10	.02	.18	− .62	2.18
Supportive listening	.13*	.02	.37	.09	− .07	.18	− .43	1.97
Choices	.11*	.01	.46	.24	.11	.12	− .70	1.71
Monitoring	.14*	.02	.41	− .03	− .02	.64	.98*	− .42
Modeling	.26*	.07	.34	− .12	.00	.40	.19	.86
Rehearsal	.08	.01	.37	.08	.04	.18	− .57	2.13
Precorrection	.24*	.06	.36	.37*	− .16	.15	− .50	2.14
Opportunities to respond	.21*	.04	.42	− .01	− .03	.58	.92*	− .49
Visual cueing	.15*	.02	.36	− .04	− .07	.19	− .54	1.78
Premack principle	.14*	.02	.37	.41	− .03	.18	− .44	1.96
Tangible reward	.35*	.12	.57	.22	.71*	.26	− .23	1.58
Time-out	.27*	.07	.41	.03	− .32	.36	.22	1.41
Praise	.06	< .01	.28	.03	.03	.47	.26	1.06
Differential reinforcement	.16*	.03	.31	–	–	–	–	
Ignoring	.27*	.07	.50	–	–	–	–	
Error correction	.15*	.02	.49	.11	− .04	.48	.41	.79
Instructive feedback	.09*	.01	.41	.04	− .12	.20	− .41	1.82
Scaffolding	.15*	.02	.29	–	–	–	–	–

TIMECS-TR Treatment Integrity Measure for Early Childhood Classrooms Teacher Report, *Corr*. Correlation, *ICC* Intraclass correlation, *TIMECS Quantity* TIMECS Quantity scale, *TIMECS Quality* TIMECS Quality scale, *SMD* standardized mean difference

*$p < .05$

(James et al., 1984), and a_{wg} (Brown & Hauenstein, 2005). These statistics estimate the extent to which scores generated by teachers on a TIMECS-TR item are discrepant or non-discrepant from scores generated by observers on the corresponding TIMECS item. For both indices, values above 0.70 suggest that an item has acceptable levels of inter-informant agreement (LeBreton et al., 2003). Only three items had values above 0.70 on a_{wg}: *Promoting Behavioral Competence* ($r_{wg} = 0.63$, $a_{wg} = 0.85$), *Monitoring* ($r_{wg} = 0.64$, $a_{wg} = 0.98$), and *Opportunities to Respond* ($r_{wg} = 0.58$, $a_{wg} = 0.92$). Overall, these analyses indicated that most TIMECS-TR items can be considered to be discrepant from the corresponding TIMECS item.

The estimates of single-item inter-informant agreement determined that the TIMECS-TR items were discrepant, but the estimates did not indicate the direction of the discrepancies. We thus calculated the standardized mean difference (SMD; De Los Reyes et al., 2019) to provide information about the direction of the discrepancies. The SMD produces an effect size estimate that can be interpreted using Cohen's (1988) criteria as a guideline for the magnitude of a discrepancy between two informants on a specific item: small SMD = 0.20 to 0.49; moderate SMD = 0.50 to 0.79; large SMD > 0.80. Items are considered to be non-discrepant if values fall between − 0.20 and − 0.20, whereas items are considered discrepant when they are above 0.80 or below − 0.80. SMD values were calculated such that positive scores indicate teacher report was higher than observer report. As can be seen in Table 3, all but four TIMECS-TR items (*Promoting Behavioral Competence*, *Monitoring*, *Opportunities to Respond*, *Error Correction*) were discrepant (*M* SMD = 1.48; SD = 0.86). Moreover, all but two items (*Monitoring*, *Opportunities to Respond*) were rated higher by teachers. These findings suggest there is little correspondence between teacher report and observer report and that teachers tend to report higher scores.

Construct Validity: TIMECS-TR Scale Scores

Our last set of analyses evaluated the construct validity for scores on the TIMECS-TR Quantity scale created by producing a mean score for all the TIMECS-TR items. First, we used a multilevel model, with observations nested within children. The TIMECS-TR Quantity scale score was used as the outcome, and the TIMECS Quantity scale, TIMECS Quality scale, and OTCIS were used as predictors in three separate multilevel models. The TIMECS-TR was not significantly associated with TIMECS Quantity scale (B = − 0.01, *SE* = 0.05, *p* = 0.840), TIMECS Quality scale (B = − 0.03, *SE* = 0.04, *p* = 0.439), or the OTCIS (B = − 0.03, *SE* = 0.04, *p* = 0.347). Second, we calculated correlations between the TIMECS-TR Quantity scale, TIMECS Quantity scale, TIMECS Quality scale, OTCIS, STRS Conflict scale, and STRS Closeness scale to facilitate comparisons with previous treatment integrity studies. 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. As can be seen in Table 4, the correlations between the TIMECS-TR Quantity scale and the remaining measures were all small in magnitude. The hypothesized correlations between the quantity, quality, and teacher–child relationship measures were not observed. Thus, these analyses do not support the construct validity of the TIMECS-TR Quantity scale score for the three measures used in the present study.

Discussion

Self-report measures are used to assess treatment integrity with regard to the delivery of EBIs within early care and education, though not nearly as often as observer-rated measures (Sanetti et al., 2020). Pragmatic self-report adherence measures could help support the evaluation and implementation of practices and EBIs focused on developing child social, emotional, and behavioral skills in early care settings. However, low correspondence between self- and observer-rated adherence measures have raised concerns about the score validity of self-report measures (e.g., Caron et al., 2019; Chapman et al., 2013; Hurlburt et al., 2010). The purpose of this study was to report on the development and initial evaluation of the psychometric properties of the TIMECS-TR, a teacher-report adherence measure designed to address limitations that may have influenced correspondence between self- and observer-report adherence measures. The TIMECS-TR parallels the content of the observer-rated TIMECS and is designed to be a cost-effective and efficient teacher report measure. Findings from the present study showed the majority of the teacher ratings of TIMECS-TR items were normally distributed and utilized the full range of scores. Test–retest score reliability of the items was small to moderate. Scores on the TIMECS-TR items and scale did not show evidence of convergent or discriminant validity with the TIMECS Quantity scale, the TIMECS Quality Scale, or the OTCIS. Follow-up analyses indicated that teachers may overreport on practice delivery relative to observer-reported delivery.

Overall, the results support the performance of the TIMECS-TR items. The majority of the TIMECS-TR items utilized the full range of scores and demonstrated a normal distribution. Only two TIMECS-TR items (*Promoting Behavioral Competence*, *Rules*) did not demonstrate a normal distribution. Of the 24 items, three did not exhibit the full range of scores, though all items showed a range of at least 3 out of 4 possible points. The items that did not exhibit a full range (*Supportive Listening*, *Monitoring*, *Teacher–Child Relationship*) were all negatively skewed. Thus, most item scores were normally distributed and utilized the full range of scores.

The test–retest reliability of scores on the TIMECS-TR items evidence small to moderate stability. This suggests that some variability in scores existed when completed week to week. Though some stability in the item scores is expected, high stability is not. Classrooms are dynamic contexts (Sutherland et al., 2008), and the social, emotional, and behavioral needs of children vary over time and activities, leading to changes in teachers' practice delivery. Future research should focus on the variability between teacher reports of their practice delivery and its association with direct observations of child behavior to better determine if the measure is sensitive to teacher responsiveness to child social, emotional, and behavioral needs. Moreover, it is important to determine whether the variability from one rating to the next is favorable or unfavorable with regard to practice elements due to alignment or misalignment with child need.

Table 4 Construct validity of TIMECS-TR scale scores

Scale	2	3	4	5	6
1. TIMECS-TR quantity	.05	− .17**	− .03	.08	.10
2. TIMECS quantity		.47**	.52**	− .08	− .07
3. TIMECS quality			.55**	.15	− .14
4. OTCIS				.28*	− .24
5. STRS closeness					− .41**
6. STRS conflict					

TIMECS-TR Treatment integrity measure for early childhood classrooms teacher report, *TIMECS Quantity* TIMECS quantity scale, *TIMECS Quality* TIMECS quality scale, *OTCIS* Observational teacher child relationship scale, *STRS* Student teacher relationship scale

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

Scores on the TIMECS-TR items and scale did not support construct validity. Neither the TIMECS-TR items nor the scale evidenced the hypothesized pattern of correlations with the corresponding observer-rated TIMECS Quantity items and scale. Moreover, the correlations between the TIMECS-TR items and scale scores demonstrated little to no association with scores on measures of teacher competence and the teacher–child relationship. This pattern is consistent with what has been seen in the mental health field (e.g., Caron et al., 2019; Chapman et al., 2013; Hurlburt et al., 2010), but we do not know of any relevant comparisons in the school-based literature. Advances in the self-report of instructional and behavioral practices have been made (e.g., Reddy et al., 2015, 2016), but the extent to which these teacher-report measures correspond with observer-report measures is unknown.

Our analyses indicated that the majority of the TIMECS-TR items were discrepant from the corresponding TIMECS items (De Los Reyes et al., 2019). Follow-up analyses indicated that teachers reported higher scores on all but two items. Previous research has indicated that self-report treatment integrity measures tend to overestimate integrity relative to observer-report measures (Breitenstein et al., 2010; Hogue et al., 2015). Our findings suggest that this may be the case with the TIMECS-TR. If so, then rater biases may account for the low correspondence between the TIMECS-TR and the TIMECS. Compared to the pattern of scores for the TIMECS-TR, scores on the TIMECS items and scales provided support for construct validity. It thus may be possible that teachers are not able to dependably report on their use of practices.

Our findings raise serious concerns about the accuracy of teacher report of adherence. To our knowledge, this was the first study to report on the correspondence between teacher- and observer-report integrity measures designed to assess parallel content. In this study, we attempted to address certain limitations of previous self-report adherence measures in an effort to improve correspondence with observer-rated measures. However, the innovative design of the TIMECS-TR failed to produce adequate correspondence with observer reports of the adherence. These findings are consistent with studies from the mental health literature that have found little agreement between self-report and observer-report measures (see e.g., Chapman et al., 2013; Hurlburt et al., 2010). Considered together, the accumulating evidence suggests that self-report measures may not produce an accurate estimate of adherence.

Despite concerns about the accuracy of teacher-report adherence measures, self-report measures do have the potential to address certain limitations of observer-rated adherence measures. For example, self-report adherence measures have the potential to be more cost-effective and feasible to use (Schoenwald et al., 2011). Moreover, self-report adherence measures have been used to predict clinical outcomes and support quality assurance in mental health research (e.g., Henggeler et al., 1999; Schoenwald et al., 2000). Thus, it is important to determine if lessons from this research can be applied to future efforts to design teacher-report adherence measures.

First, it is possible that the discrepancies between teacher and observer report were due to training. The TIMECS raters received didactic training, engaged in independent ratings, and received feedback on their scoring. In contrast, the teachers were only provided a training manual, asked if they understood the item definitions, and provided opportunities to ask questions about using the TIMECS-TR. It is likely that the steps taken to train the teachers was insufficient. A few studies indicate that the provision of training and feedback can improve the accuracy of ratings (e.g., Fallon et al., 2018). It is likely that training needs to be more intensive and include feedback. For example, Reddy et al., (2015) used didactic training and video examples to train teachers to rate their own instructional behavior. Thus, more intensive training that is more comparable to that provided to the coders using the TIMECS may be necessary to generate scores that demonstrate adequate correspondence with observational ratings.

Second, the TIMECS-TR scoring strategy may have contributed to the discrepant findings. We asked teachers to report on the extent to which each practice was used in the past five days. There are at least two potential problems with this scoring strategy. First, early childhood classrooms differ across a number of dimensions that may influence the delivery of practices with focal children both during and across the days in a week. For example, early childhood programs vary across a variety of structural dimensions (e.g., teachers' years of education, number of program hours, teacher–child ratio), but they also differ across process dimensions (e.g., interactions and teaching, provisions for learning, and emotional climate; Dotterer et al., 2013). Second, the reporting period may be too long and, consequently, teachers may not accurately remember the practices used over the course of a week. Most self-report treatment integrity measures in mental health ask clinicians to report on the practices used within a 45–60 min treatment session, as opposed to a five-day period. Considered together, asking teachers to report on a smaller time period (e.g., an hour or day) may provide a more accurate estimate of the practices used with a focal child.

A third source of the discrepant ratings may be the timing of assessments. We took steps to ensure that teacher- and observer-reported data were collected within the same 5-day period, which aligns with the time period teachers were asked to report on when filling out the TIMECS-TR. Despite this attempt to align the timing of assessment, our approach still had a discrepancy. Observers generated ratings based

on a 45-min observation, and it is unclear whether these ratings generalize to the practices used across a full week. In contrast, teachers were asked to report on the practices used with a focal child over the course of a 5-day period, which may not represent what happened with a child on any single day or occasion. Given the extent that the timing between observer and self-report ratings differ, our findings indicate that it is likely observer and teacher reports of integrity will be discrepant (Collier-Meek et al., 2019). Based on this, by aligning the timing between observer and teacher ratings of integrity or reducing the time between assessment, congruence between the two may increase.

A final potential source is item design. We took steps to ensure that items were written in accessible language. However, there is a need to explore how best to write items to facilitate accurate and consistent assessment. Items on existing integrity measures vary in their level of specificity (Schoenwald et al., 2011). A specificity approach includes more information to capture precise and specific aspects of the assessment target, while a breadth approach involves more holistic representativeness of the assessment target. Our items vary in how specific or broad each is, and this may interact with training and timing in important ways. Despite arguments for either approach, no research has tested whether broad or specific items best capture teacher behavior (Schoenwald et al., 2011).

Limitations and Implications for Future Research

An important goal for the field is to develop psychometrically sound teacher-report treatment integrity measures to advance implementation research and intervention science. Our findings highlight important areas for future research. First, researchers should investigate the role professional development, training, and implementation support and feedback plays in teachers' ability to rate their own practice delivery. A limitation of the current study was that teachers were only provided a brief manual describing the measure, items, and procedures for completing. Moreover, treatment integrity and comprehension of the training was not monitored. Future work should examine correspondence between teacher-reported and observer-reported measures of treatment integrity under different conditions of training (see Reddy et al., 2015), measure fidelity to the training, and gauge teacher comprehension to establish standards for the field to follow in order to maximize the utility of teacher-report measures.

Second, future research should examine the role of different scoring strategies for teacher-report measures (e.g., one hour, a day, a week) along with the role of timing in the completion of teacher-report integrity measures in comparison with observer reports. In the current study, teachers were asked to self-report on their use of practices with a focal child in the past week, which may have provided too much opportunity for recency bias (e.g., to influence scores). Future work should empirically manipulate the scoring strategy and the timing of teacher report in comparison to observer report to identify the optimal timing necessary for accurate self-reporting of practice delivery.

Finally, each of these areas for future research should be conducted with an end-goal in mind (e.g., to develop psychometrically sound, pragmatic teacher self-reports of adherence). This is critically important to advancing implementation and intervention science in the area of school-based delivery of social, emotional, and behavioral interventions (Sutherland et al., 2013). To illustrate, the field needs to determine the minimal amount of training necessary to support teachers in reliably assessing their own delivery of practices in order to increase the usability of these measures in applied settings. Further, we need to identify the optimal timing for teacher reports to correspond with reliable observer-reported measures and include this information in training for teachers and other service providers in schools. By optimizing the training and timing of teacher self-report measures, we can increase the scale of usage of these measures in early childhood settings, potentially enhancing our ability to monitor and support teachers' delivery of social, emotional and behavioral interventions.

Conclusion

A parallel set of self-report and observer-report treatment integrity measures represents an innovative and practically important next step to assist in addressing the science-to-practice gap. Multiple end-users could benefit from a suite of measures, including administrators, consultants, purveyors, intermediaries who support EBI implementation, internal teams operating in early childhood programs, and teachers. However, as with previous research, our self-report and observer-report measures failed to converge. As a result, the current evidence does not support the use of teacher-report measures to assess adherence. More research is needed to help identify possible reasons for the discrepancies and develop a self-report measure that can dependably capture teacher delivery of practices found in EBIs.

References

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Orwig, D., & Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH behavior change consortium. *Health Psychology, 23*(5), 443–451. https://doi.org/10.1037/0278-6133.23.5.443

Birch, S. H., & Ladd, G. W. (1997). The teacher–child relationship and children's early school adjustment. *Journal of School Psychology, 35*(1), 61–79. https://doi.org/10.1016/S0022-4405(96)00029-5

Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher–child relationship. *Developmental Psychology, 34*(5), 934–946. https://doi.org/10.1037/0012-1649.34.5.934

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing & Health, 33*(2), 164–173. https://doi.org/10.1002/nur.20373

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the rwg indices. *Organizational Research Methods, 8,* 165–184. https://doi.org/10.1177/1094428105275376

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*(3), 350–363. https://doi.org/10.1017/S1352465819000699

Carroll, K. M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., Fenton, L., & Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence, 57*(3), 225–238. https://doi.org/10.1016/s0376-8716(99)00049-6

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. https://doi.org/10.1037/a0033021

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. https://doi.org/10.1037/a0014565

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Collier-Meek, M. A., Sanetti, L. M., Fallon, L., & Chafouleas, S. (2019). Exploring the influences of assessment method, intervention steps, intervention sessions, and observation timing on treatment fidelity estimates. *Assessment for Effective Intervention, 46*(1), 3–13. https://doi.org/10.1177/1534508419857228

Connors, E., Lawson, G., Wheatley-Rowe, D., & Hoover, S. (2020). Exploration, preparation, and implementation of standardized assessment in a multi-agency school behavioral health network. *Administration and Policy in Mental Health and Mental Health Services Research*. https://doi.org/10.1007/s10488-020-01082-7

Dart, E. H., Collier-Meek, M. A., Chambers, C., & Murphy, A. (2020). Multi-informant assessment of treatment integrity in the classroom. *Psychology in the Schools, 57,* 805–822. https://doi.org/10.1002/pits.22351

De Los Reyes, A., Cook, C. R., Gresham, M., Bridget, A., Makol, A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology, 74,* 74–89. https://doi.org/10.1016/j.jsp.2019.05.005

Dotterer, A. M., Burchinal, M., Cryant, D., Early, D., & Pianta, R. (2013). Universal and targeted pre-kindergarten programmes: A comparison of classroom characteristics and child outcomes. *Early Child Development and Care, 183,* 931–950. https://doi.org/10.1080/03004430.2012.698388

Fallon, L. M., Sanetti, L. M. H., Chafouleas, S. M., Faggella-Luby, M. N., & Briesch, A. M. (2018). Direct training to increase agreement between teachers' and observers' treatment integrity ratings. *Assessment for Effective Interventions, 43*(4), 196–211. https://doi.org/10.1177/1534508417738721

Fleiss, J. (1981). Balanced incomplete block designs for interrater reliability studies. *Applied Psychological Measurement, 5,* 105–112. https://doi.org/10.1177/014662168100500115

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. https://doi.org/10.1111/1467-8624.00301

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Henggeler, S. W., & Borduin, C. M. (1992). *Multisystemic therapy adherence scales*. Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina.

Henggeler, S. W., Pickrel, S. G., & Brondino, M. J. (1999). Multisystemic treatment of substance-abusing and dependent delinquents: Outcomes, treatment fidelity, and transportability. *Mental Health Services Research, 1,* 171–184. https://doi.org/10.1023/a:1022373813261

Hogue, A. (2002). Adherence process research on developmental interventions: Filling in the middle. In A. Higgins-D'Alessandro & K. R. B. Jankowski (Eds.), *New directions for child and adolescent development, Vol. 98: Science for society: Informing policy and practice through research in developmental psychology* (pp. 67–74). Jossey Bass.

Hogue, A., Liddle, H. A., & Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy, 33,* 332–345. https://doi.org/10.1037/0033-3204.33.2.332

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inclan, J., Reiner, R. H., & Liddle, H. A. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment, 35*(2), 137–147. https://doi.org/10.1016/j.jsat.2007.09.002

Hogue, A., Ozechowski, T. J., Robbins, M. S., & Waldron, H. B. (2013). Making fidelity an intramural game: Localizing quality assurance procedures to promote sustainability of evidence-based practices in usual care. *Clinical Psychology: Science and Practice, 20,* 60–77. https://doi.org/10.1111/cpsp.12023

Hogue, A., Dauber, S., Henderson, C. E., & Liddle, H. A. (2014). Reliability of therapist self-report on treatment targets and focus in family-based intervention. *Administration and Policy in Mental Health and Mental Health Services Research, 41,* 697–705. https://doi.org/10.1007/s10488-013-0520-6

Hogue, A., Dauber, S., Lichvar, E., Bobek, M., & Henderson, C. E. (2015). Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: Correspondence between therapists and observers. *Administration and Policy in Mental Health and Mental Health Services Research, 42,* 229–243. https://doi.org/10.1007/s10488-014-0548-2

Hogue, A., Bobek, M., Dauber, S., Henderson, C. E., McLeod, B. D., & Southam-Gerow, M. A. (2017). Distilling the core elements of family therapy for adolescent substance use: Conceptual and empirical solutions. *Journal of Child and Adolescent Substance Abuse, 26*(6), 437–453. https://doi.org/10.1080/1067828X.2017.1322020

Hogue, A., Bobek, M., Dauber, S., Henderson, C. E., McLeod, B. D., & Southam-Gerow, M. A. (2019). Core elements of family therapy for adolescent substance use: Empirical distillation of three manualized treatments. *Journal of Clinical Child and Adolescent*

Psychology, 48(1), 29–41. https://doi.org/10.1080/15374416.2018.1555762

Howes, C., & Ritchie, S. (1999). Attachment organizations in children with difficult life circumstances. *Development and Psychopathology, 11*(2), 251–268. https://doi.org/10.1017/S0954579499002047

Hurlburt, M. S., Garland, A. F., Nguyen, K., & Brookman-Frazee, L. (2010). Child and family therapy process: Concordance of therapist and observational perspectives. *Administration and Policy in Mental Health and Mental Health Services Research, 37,* 230–244. https://doi.org/10.1007/s10488-009-0251-x

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85–98. https://doi.org/10.1037/0021-9010.69.1.85

Lebreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80–128. https://doi.org/10.1177/1094428102239427

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Lyon, A. R., & Koerner, K. (2016). User-centered design for psychosocial intervention development and implementation. *Clinical Psychology: Science and Practice, 23*(2), 180–200. https://doi.org/10.1111/cpsp.12154

Margolin, G., Oliver, P., Gordis, E., O'Hearn, H., Medina, A., Ghosh, C., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review, 1*(4), 195–213. https://doi.org/10.1023/a:1022608117322

McLeod, B. D., & Sutherland, K. S. (2015). *Scoring manual for the observational teacher–child relationship measure*. Unpublished scoring manual prepared at Virginia Commonwealth University.

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review, 38,* 541–546.

McLeod, B. D., Southam-Gerow, M. A., Bair, C. E., Rodriguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychological treatment quality indicator. *Clinical Psychology: Science and Practice, 20*(1), 14–32. https://doi.org/10.1111/cpsp.12020

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. https://doi.org/10.1007/s11121-016-0703-y

McLeod, B. D., Sutherland, K. S., Broda, M., Granger, K. L., Martinez, R. G., Conroy, M. A., Snyder, P. A., & Southam-Gerow, M. A. (2020). Development and initial psychometrics of a generic treatment integrity measure designed to assess practice elements of evidence-based interventions for early childhood settings. Manuscript submitted for publication.

McLeod, B. D., Sutherland, K. S., Broda, M., Granger, K. L., Frey, A., & Markowicz, K. (2021). Development and initial psychometrics of the observational teacher-child interactions scale for early childhood settings. Manuscript in preparation.

Newborg, J. (2005). *Battelle developmental inventory, 2nd edition, examiner's manual*. Riverside Publishing.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*(6), 829–841. https://doi.org/10.1037/0022-006X.75.6.829

Pianta, R. C., & Hamre, B. (2001). *Students, teachers, and relationship support (STARS)*. Psychological Assessment Resources.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal, 102,* 225–238. https://doi.org/10.1086/499701

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffery, R., & 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*(2), 65–76. https://doi.org/10.1007/s10488-010-0319-7

Reddy, L. A., Dudek, C. M., Fabiano, G. A., & Peters, S. (2015). Measuring teacher self-report on classroom practices: Construct validity and reliability of the classroom strategies scale—Teacher form. *School Psychology Quarterly, 30*(4), 513–533. https://doi.org/10.1037/spq0000110

Reddy, L. A., Dudek, C. M., Rualo, A. J., & Fabiano, G. A. (2016). Concurrent validity of the classroom strategies scale—teacher form: A preliminary investigation. *Educational Assessment, 21*(4), 267–277. https://doi.org/10.1080/10627197.2016.1236675

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

Sanetti, L. M., & Collier-Meek, M. (2019). Increasing implementation science literacy to address the research-to-practice gap in school psychology. *Journal of School Psychology, 76,* 33–47. https://doi.org/10.1016/j.jsp.2019.07.008

Sanetti, L. M., Gritter, K. L., & Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review, 40*(1), 72–84. https://doi.org/10.1177/01430343134763999

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*(6), 901–922. https://doi.org/10.1002/pits.22364

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. *Family Process, 39,* 83–103. https://doi.org/10.1111/j.1545-5300.2000.39109.x

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38,* 32–43. https://doi.org/10.1007/s10488-010-0321-0

Snyder, P. A., Hemmeter, M. L., & Fox, L. (2015). Supporting implementation of evidence-based practices through practice-based coaching. *Topics in Early Childhood Special Education, 35*(3), 133–143. https://doi.org/10.1177/0271121415594925

Stanick, C. F., Halko, H. M., Nolen, E. A., Powell, B. J., Dorsey, C. N., Mettert, K. D., Weiner, B. J., Barwick, M., Wolfenden, L., Damschroder, L. J., & Lewis, C. C. (2019). Pragmatic measures for implementation research: development of the psychometric and pragmatic evidence rating scale (PAPERS). *Translational Behavioral Medicine. Advance Online Publication.* https://doi.org/10.1093/tbm/ibz164

Sutherland, K. S., & McLeod, B. D. (2015a). *Scoring manual for the treatment integrity measure for early childhood settings: the adherence and competence scale*. Unpublished scoring manual prepared at Virginia Commonwealth University.

Sutherland, K. S., & McLeod, B. D. (2015b). *Scoring manual for the treatment integrity measure for early childhood settings: the*

teacher report scale. Unpublished scoring manual prepared at Virginia Commonwealth University.

Sutherland, K. S., Wehby, J. H., & Copeland, S. R. (2000). Effect of varying rates of behavior-specific praise on the on-task behavior of students with EBD. *Journal of Emotional and Behavioral Disorders, 8*(1), 2–8. https://doi.org/10.1177/106342660000800101

Sutherland, K. S., Lewis-Palmer, T., Stichter, J., & Morgan, P. (2008). Examining the influence of teacher behavior and classroom context on the behavioral and academic outcomes for students with emotional or behavioral disorders. *Journal of Special Education, 41,* 223–233. https://doi.org/10.1177/0022466907310372

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. https://doi.org/10.1177/1053815113515025

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. https://doi.org/10.1177/1063426613497258

Sutherland, K. S., Conroy, M. A., Algina, J., Ladwig, C., Jessee, G., & Gyure, M. (2018). Reducing child problem behaviors and improving teacher-child interactions and relationships: A randomized controlled trial of BEST in CLASS. *Early Childhood Research Quarterly, 42,* 31–43. https://doi.org/10.1016/j.ecresq.2017.08.001

Sutherland, K. S., Conroy, M. A., & Granger, K. (2020). BEST in CLASS: A Tier-2 program for children with and at-risk for emotional/behavioral disorders. In T. Farmer, M. Conroy, E. Farmer, & K. Sutherland (Eds.), *Handbook of research on emotional and behavioral disorders: interdisciplinary developmental perspectives on children and youth* (pp. 214–226). Routledge/Taylor & Francis.

Trochim, W. M., & Donnelly, J. P. (2006). *The research methods knowledge base* (3rd ed.). Atomic Dog.

Walker, H., Severson, H., & Feil, E. (1995). *Early screening project: A proven child find process*. Sopris West Publishing.

Ware, N. C., Dickey, B., Tugenberg, T., & McHorney, C. A. (2003). CONNECT: A measure of continuity of care in mental health services. *Administration and Policy in Mental Health and Mental Health Services Research, 5*(4), 209–221. https://doi.org/10.1023/A:1026276918081

Yoder, P. J., Symons, F. J., & Lloyd, B. (2018). *Observational measurement of behavior* (2nd ed.). Brookes Publishing.