School Mental Health
https://doi.org/10.1007/s12310-021-09423-z

ORIGINAL PAPER

Developing Treatment Integrity Measures for Teacher-Delivered Interventions: Progress, Recommendations and Future Directions

Kevin S. Sutherland[1] · Bryce D. McLeod[2] · Maureen A. Conroy[3] · Nicholas Mccormick[1]

Accepted: 6 February 2021

Abstract

While the measurement of treatment integrity is important to determine how much, and how well, interventions are delivered in schools, the science of treatment integrity is not well developed in education research. The purpose of this paper is to describe a program of research that has developed treatment integrity measures over the past 10 years to assess teacher delivery of an indicated program targeting reductions in problem behavior in early childhood and elementary school classrooms. Specifically, this paper will highlight the importance of active use of conceptual models to guide treatment integrity measure development, multidimensional assessment of treatment integrity and training procedures for observers, using several studies to illustrate the evolution and refinement of our measurement approach. Recommendations for researchers developing and evaluating interventions in schools are provided, as are recommendations to help the field move toward a more rigorous science of treatment integrity.

The ability to measure treatment integrity, the quantity and quality of how teachers deliver practices and intervention programs designed to promote social and emotional learning in schools, is important for intervention development, evaluation and implementation. First, when evidence-based practices and programs are delivered with integrity, students are more likely to learn social, emotional and behavioral skills that promote their well-being and maximize development and learning opportunities (Durlak 2010). Second, if practices and programs are delivered with integrity, learning contexts improve (Conroy et al. 2019). Third, understanding how much and how well teachers delivered the practices found in a treatment protocol (i.e., treatment adherence) can help researchers interpret study findings and identify the key ingredients of the intervention (Sutherland et al. 2013b). Finally, by understanding how, and how well, teachers deliver practices and programs researchers can identify factors that influence the delivery of the program; thus, factors that influence program implementation can be identified and addressed to maximize the effectiveness of practices and programs in various school contexts (McLeod et al. 2020). To better understand the quantity and quality of how teachers deliver practices and programs to promote social and emotional learning, researchers need psychometrically sound measurement tools.

Research has shown that while the number of school-based studies reporting treatment integrity (also referred to as treatment fidelity, intervention integrity) has increased (Sanetti et al. 2020), many studies only minimally address treatment integrity (e.g., Sanetti et al. 2012, 2011). Treatment integrity is conceptualized as a multidimensional construct (see below); yet, most studies that do report on treatment integrity focus only on adherence (Sanetti et al. 2012). Further, Sanetti and Reed found that researchers reported that the time required to assess treatment integrity and lack of agreement about how best to assess treatment integrity

✉ Kevin S. Sutherland
kssuther@vcu.edu

Bryce D. McLeod
bmcleod@vcu.edu

Maureen A. Conroy
mconroy@coe.ufl.edu

Nicholas Mccormick
mccormicknp@mymail.vcu.edu

[1] Department of Counseling and Special Education, Virginia Commonwealth University, 1015 W. Main St, PO Box 842020, Richmond, VA 23284, USA

[2] Department of Psychology, Virginia Commonwealth University, 806 W. Franklin St, PO Box 842018, Richmond, VA 23284, USA

[3] Department of Special Education, School Psychology and Early Childood Studies, University of Florida, 1345 Norman Hall, PO Box 117050, Gainesville, FL 32611, USA

Published online: 20 February 2021

were two primary barriers to the measurement of treatment integrity in school-based research. Clearly, there is a need in school-based research to provide further information on how best to measure treatment integrity in a comprehensive manner. This article is designed to address this need by describing a program of research that has developed measurement tools to assess multiple dimensions of treatment integrity of teacher-delivered practices in early childhood and elementary classrooms to support the social and emotional learning of young children and students who demonstrate chronic problem behavior. An overarching purpose of the current article is to provide a framework for treatment integrity measure development that can assist school-based researchers in developing treatment integrity measures to support intervention development, evaluation and implementation efforts.

After defining key terms that appear throughout the article, we will discuss the importance of assessing multiple dimensions of treatment integrity as well as provide a conceptual model that has guided our work. Next, we will use the development and evaluation of BEST in CLASS (Conroy et al. 2019; Sutherland et al. 2020), a Tier 2 program designed to support children and young students with chronic problem behavior, as a context for a description of how treatment integrity measures, and training and monitoring of integrity measurement, have contributed to our understanding of how, and how well, teachers deliver the core elements of BEST in CLASS. We will finish with recommendations for measuring treatment integrity of teacher-delivered practices and programs as well as future directions to advance intervention science in the delivery of social, emotional and behavioral support programming.

Definitions and Conceptual Model

It is important to place our discussion of treatment integrity within the model of translational research that starts with basic research and progresses to implementation research. The translational research model stipulates that the evaluation of interventions begins with basic research (what is the importance of the teacher–student relationship?), progresses to efficacy trials (does an intervention work under controlled conditions when implemented by the researchers?), moves to effectiveness trials (does an intervention work when implemented in authentic settings by authentic providers?) and then moves to implementation trials (what activities and strategies or adaptations are required to integrate and sustain an intervention into a specific context?). The needs and focus of treatment integrity measurement differ as an intervention progresses along the translational pipeline. Early in the development of an intervention treatment integrity measurement focuses on determining whether the core components of the treatment protocol were delivered in order to inform intervention refinement. During efficacy trials, treatment integrity measures are often used for manipulation checks (i.e., a test to ascertain whether a variable was successfully manipulated) intended to determine whether an intervention under study was delivered as intended (Perepletchikova and Kazdin 2005; Waltz et al. 1993). As an intervention arrives at effectiveness and implementation research, treatment integrity measures often are used as dependent variables (were training and coaching successful? Proctor et al. 2011). Also, the design of treatment integrity measures often changes across the translational pipeline (i.e., more detailed and specific measures used early in the process, whereas more generic and pragmatic measures are used in implementation research). Thus, treatment integrity instruments designed for one phase may not be a good fit for all research questions along the translational pipeline (McLeod et al. 2013; Schoenwald 2011).

In general, treatment integrity is defined as the degree to which practices or programs are delivered as intended (McLeod et al. 2009; Sanetti and Kratochwill 2009; Sutherland et al. 2013a, b). We conceptualize treatment integrity of teacher-delivered practices as being comprised of four components (Fig. 1): adherence, competence, differentiation and child responsiveness (Sutherland et al. 2013b). Situated within multi-tiered systems of support (e.g., Positive Behavior Interventions and Supports, PBIS), our integrity measurement is focused on children and students with more indicated support needs (e.g., Tier 2 and Tier 3). Therefore, when assessing treatment integrity our coders are taught to focus on teacher practices targeted toward a particular student (i.e., focal student); treatment integrity measurement of Tier 1 (universal practices) would target teacher-delivered practices to all children or students in a classroom.

Adherence is defined as the extent to which a teacher delivers the core components of an intervention (e.g., the teacher provides multiple opportunities, with scaffolding, for a student to demonstrate a behavior). Competence is defined as how well those core components are delivered (e.g., when providing opportunities for a student to demonstrate a skill the teacher is responsive to the student's needs, is encouraging and uses developmentally appropriate language). Treatment differentiation is defined as the extent to which a teacher delivers proscribed practices (i.e., not representing core components). Finally, child responsiveness represents how the recipients of an intervention respond to a teacher's attempts to deliver the core components of the intervention; this dimension of treatment integrity may be represented by behaviors such as child engagement, responsiveness to teacher attempts to deliver core components of the intervention, or contra-indicated behavior, such as disruption. Each of these dimensions of treatment integrity has been

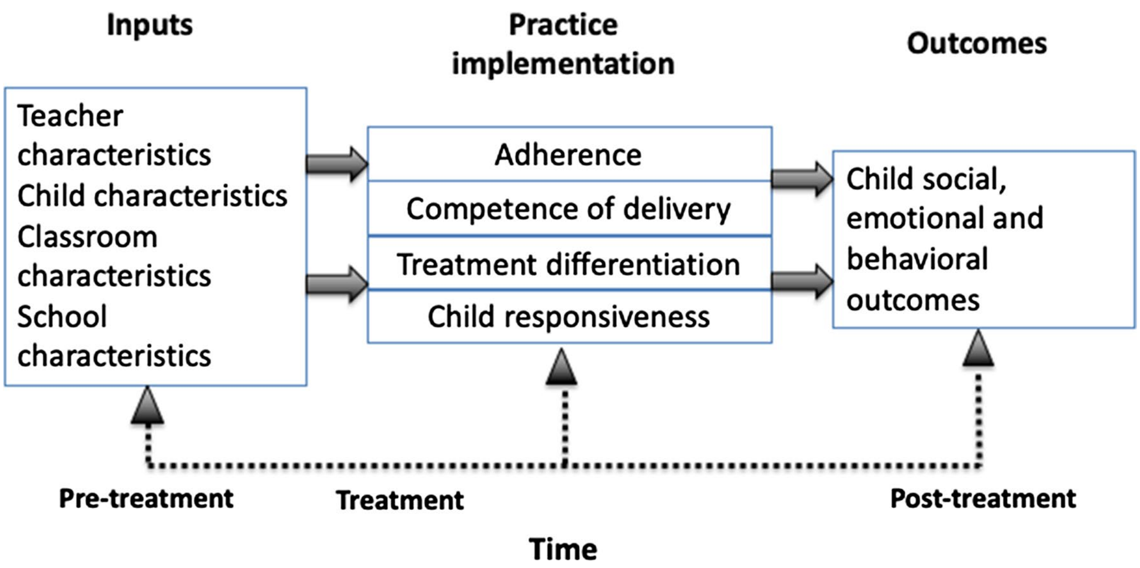

Fig. 1 Conceptual model of treatment integrity

associated with treatment outcomes (e.g., Durlak 2010; Sutherland et al. 2018b; Vroom et al. 2020).

We have included these dimensions of treatment integrity in a conceptual model that guides our work in intervention science (Fig. 1). Specifically, we suggest that teacher, child, classroom and school characteristics influence how, and how well, teacher-delivered interventions are implemented in classrooms. Researchers have used social–ecological models (Bronfenbrenner 1979) to describe a number of influences on teacher delivery of interventions at the child, teacher, classroom and school level (see Domitrovich et al. 2008; Durlak 2015; Han and Weiss 2005), and research has supported the influence of these factors on teacher treatment integrity (e.g., Sutherland et al. 2018b; Williford et al. 2015).

The middle part of our model represents the dimensions of treatment integrity that are influenced by these factors and in turn influence outcomes on the right side of our model (i.e., child social, emotional and behavioral outcomes). Research indicates that these dimensions of integrity are associated with intervention effects across a number of studies. For example, Sutherland, Conroy, McLeod, Algina and Wu (2018c) found that teacher competence of delivery mediated the effects of BEST in CLASS on reductions in child externalizing problem behavior. Similarly, Vroom et al. (2020) found that student responsiveness was associated with students' social–emotional learning skills at posttest in a study of the Life Skills Training program, an evidence-based social–emotional learning program (Botvin and Griffin 2004). In sum, we propose that a variety of factors influence treatment integrity dimensions, which in turn influence child outcomes. Thus, the measurement of these treatment integrity dimensions is critical to understanding how teacher-delivered interventions affect, or do not affect, child social, emotional and behavioral outcomes. In the next section, we briefly describe BEST in CLASS, a Tier 2 program designed to support children and young students with chronic problem behavior, as a context for a description of how we have developed a suite of treatment integrity measures.

BEST in CLASS

BEST in CLASS is a Tier 2 program delivered by classroom teachers, with support from trained coaches, that targets improvements in teacher–child interactions and relationships in order to reduce the chronic problem behavior of young children and students with or at risk of emotional/behavioral disorders (EBD). BEST in CLASS is comprised of a number of evidence-informed practices (McLeod et al. 2017; Sutherland et al. 2019) that teachers deliver to focal children (i.e., children identified as having chronic problem behavior) during authentic learning activities in the classroom throughout the day. BEST in CLASS has demonstrated reductions in child problem behavior, improvements in teacher behavior and improvements in teacher–child interactions and relationships across a number of studies (Conroy et al. 2018, 2019; Sutherland et al. 2018a, c; Sutherland et al. 2020).

The measurement of treatment integrity has been critical to the development and testing of BEST in CLASS. Initially, the dimensions of adherence and competence were assessed (e.g., Conroy et al. 2018), and in later iterations of measure development, the dimensions of child responsiveness and differentiation were added (e.g., McLeod et al. 2020; Sutherland et al. 2020). Our ability to use valid and reliable measures of treatment integrity (see Sutherland et al.

2014) allowed our research team to learn more about factors associated with teacher delivery of BEST in CLASS (Sutherland et al. 2018b) and to examine the relationship between treatment integrity and child social, emotional and behavioral outcomes (Sutherland et al. 2018c). In the next section, we will first describe the measurement approach that guides assessment of teacher integrity of delivery of BEST in CLASS. We will then describe the process we have used to guide treatment integrity measure development, starting with the BEST in CLASS Adherence and Competence Scale (BiCACS; Sutherland et al. 2014). Within this description, we will emphasize steps we have taken to improve our operational definitions of codes, as well as training, supervision and monitoring of data collectors. Throughout the following section, we will provide data to highlight the effect of these improvements on the reliability of our treatment integrity measurement.

Development of Treatment Integrity Measure

To provide an objective estimate of treatment integrity, we developed an observer-rated treatment integrity instrument for use by trained coders. In order to estimate the treatment integrity of teacher delivery of the core components of BEST in CLASS specified in the treatment protocol, we used a four-step approach to measure development (see Hogue et al. 1996; McLeod and Weisz 2010): scale development, item development, selection of scoring strategies and pilot coding.

Scale Development

The first step in measure development was to determine what treatment integrity dimensions are important to capture. The main purpose of our treatment integrity measures was to provide a means of documenting how extensively (i.e., adherence) and how well (i.e., competence) practices found in BEST in CLASS were delivered by teachers. We therefore determined that we would develop separate Adherence and Competence scales. In addition, in later iterations of our measure development work we sought to document child responsiveness to the BEST in CLASS practices so we also included a Child Responsiveness scale. Last, given the value-added nature of BEST in CLASS (i.e., teachers may be using prescribed practices, just not extensively, with high quality or with students identified with Tier 2 needs) as well as the complexities of early childhood and elementary school classrooms, we sought to also characterize teacher delivery of proscribed practices (i.e., those not represented in BEST in CLASS) via a Differentiation scale.

Item Development

It was important that we be able to measure teacher delivery of specific practices; therefore, when creating items, we focused on operational definitions of discrete practices. Initially, these were the practices that comprised the BEST in CLASS model (e.g., rules, precorrection, opportunities to respond, behavior-specific praise, instructive feedback, corrective feedback), and in later measure development work, we used a practice elements approach (McLeod et al. 2017; Sutherland et al. 2019; see below) to identify items.

Scoring Strategies

Because it was expected that teachers would vary in the extent to which they delivered different practices, it was important that the scoring strategies used for the scales capture the breadth, depth and quality of practice delivery. To achieve this goal, we used scoring strategies used in exemplar coding systems from mental health treatment research (see Carroll et al. 2000; Hogue et al. 1996; McLeod and Weisz 2010).

For items on the Adherence scale, the scoring strategy is designed to yield quantitative data that are non-subjective and specific with regard to how teachers deliver the core practices found in BEST in CLASS. Existing treatment integrity measures differ greatly in their scoring strategies and range from microanalytic strategies (e.g., frequency counts) to macroanalytic scoring of an entire observation (i.e., generating a single score based on a longer observation). Because we expected that teachers would vary in the extent to which they employed different practices, it was important that the scoring strategy capture both the breadth and depth of practice delivery. Microanalytic scoring strategies were thus ruled out (e.g., scoring of frequency counts) because these scoring strategies fail to capture the important contextual variables (e.g., the depth or complexity of a practice) that can influence the effectiveness of a practice (Greenberg 1986). For example, the exclusive use of frequency counts can misrepresent treatment integrity by giving a higher weight to practices that are used more often, but not in a more thorough manner (Greenberg 1986). With microanalytic strategies ruled out, we turned to macroanalytic scoring strategies.

The scoring strategy involves macroanalytic extensiveness ratings of practices designed to measure the degree to which teachers use a specific practice during an observation. This extensiveness rating strategy was based directly upon the scoring strategy used in exemplar treatment integrity measures (e.g., Carroll et al. 2000; Evans et al. 1984;

Hogue et al. 1996). In making extensiveness ratings, coders are asked to estimate the extent to which teachers engage in each practice during the entire observation using a seven-point Likert-type scale with the following anchors: 1 = *not at all*, 3 = *somewhat*, 5 = *considerable* and 7 = *extensively*. In other words, if a practice is observed, then coders determine the extensiveness of delivery ranging from 2 to 7. Two components are considered when making extensiveness ratings of observed practices: thoroughness and frequency. Thoroughness is defined as the persistence and depth with which a teacher attempts to deliver a practice. Frequency refers to the amount of times a practice occurs during the observation, and coders are trained to use both of these components in making extensiveness ratings. Both thoroughness and frequency are considered in making a rating on each item; for example, persistence in delivering three opportunities to respond in quick succession to a student in order to solicit a specific correct response would be considered more thorough than three opportunities to respond delivered independently during an observation. Therefore, extensiveness ratings provide quantity, or dosage, information about each practice. In other words, these ratings determine how much of each practice the child is exposed to in a given observation (e.g., how strong a dose of behavior-specific praise the teacher provided to the child).

We adopted a scoring strategy for the competence items that involves macroanalytic competence ratings that consider the quality of delivery (skillfulness) and the timing and appropriateness of delivery for a given child and context (responsiveness). For each item, coders consider the extent to which a teacher demonstrated the following skillfulness and responsiveness dimensions in an observation (Carroll et al. 2000): (a) expertise, (b) clarity of communication, (c) appropriate timing of delivery and (d) read and respond to the child. In making 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*. This scoring strategy was adapted slightly from exemplar competence coding systems developed for youth (Hogue et al. 2008) and adult (Barber et al. 1996; Carroll et al. 2000) mental health treatment. Coders are instructed to consider ratings of "4" as average competence, ratings above "4" as above average and ratings below "4" as below average.

Pilot Coding

Once the previous steps were completed, a preliminary coding manual was developed and pilot coding was used to refine the manual. The coding manual was designed to provide coders with a comprehensive guide for coding observations. Coders across all development phases were graduate students (i.e., doctoral students in clinical psychology, educational psychology or special education) and post-baccalaureate research assistants. The manual serves as a companion document for training new coders as well as a reference document for trained coders to use while coding. As such, the manual contains a thorough description of each item and provides additional information to help the coder make coding decisions in an informed and reliable manner. Our coding manuals were modeled after exemplar systems in the mental health field (see Evans et al. 1984; Hogue et al. 1996; Hollon et al. 1988) and are organized into two sections. The first section, *General Instructions*, provides an overview of procedural guidelines, scoring strategies and coder caveats to help coders acquire and maintain coding reliability (e.g., how to avoid "haloed" ratings). The second section, *Item Descriptions*, provides detailed descriptions and examples for each item. Each of the items that comprise the measure is presented in the following format: (a) item as it appears on the extensiveness or competence scoring sheet (Table 1); (b) brief description of the item and its purpose within the scale; (c) supplemental coding information including specific examples of different levels of extensiveness or competence; (d) exemplar teacher statements; and (e) guidelines for differentiating the item from other items. Published manuals are available upon request from the authors.

BEST in CLASS Adherence and Competence Scale (BiCACS)

The BiCACS (see Sutherland et al. 2014) was developed as part of an Institute of Education Sciences (IES)-funded project that supported the initial development of BEST in CLASS. Our goals in developing this initial integrity measure were threefold. First, we wanted to measure teacher delivery of each of the practices that comprised the BEST in CLASS model in a way that allowed for item variability and to measure the extensiveness of practice delivery (i.e., not a dichotomous checklist). Second, since BEST in CLASS is a value-added intervention (i.e., teachers are likely already using some of the practices to some degree in their classrooms, just not with the extensiveness or quality focal children may need), we wanted to be able to assess teacher delivery of practices at pretest as well as in business-as-usual (BAU) classrooms. Third, we wanted to be able to assess both the extensiveness (i.e., adherence) of delivery and the quality (i.e., competence) of delivery. Within the translational research model, we were early in the development of BEST in CLASS treatment integrity measurement and therefore focused on determining how extensively and how well the core components of the treatment protocol were delivered.

Once the core components of BEST in CLASS were identified during the intervention development process,

Table 1 Intraclass correlation coefficients (ICCs) across measures

	BiCACS		BiCACS—Web		TIMECS		TIES	
	N	*ICC*	*N*	*ICC*	*N*	*ICC*	*N*	*ICC*
Adherence Items								
Emotion Regulation	–	–	–	–	650	0.890	132	0.785
Self-Management	–	–	–	–	–	–	132	0.839
Instructional Feedback	628	0.690	24	0.778	650	0.760	132	0.802
Peer Tutoring	–	–	–	–	–	–	132	0.793
Problem-Solving	–	–	–	–	650	0.830	132	0.915
Punishment	–	–	–	–	–	–	132	0.796
Reinforcement	–	–	–	–	–	–	132	0.774
Routines	–	–	–	–	–	–	132	0.799
Social Skills	–	–	–	–	650	0.880	132	0.671
Teacher–Student Relationships	–	–	–	–	650	0.870	132	0.725
Active Supervision	–	–	–	–	–	–	132	0.657
Behavioral Momentum	–	–	–	–	–	–	132	1.000
Choice	–	–	–	–	650	0.680	132	0.561
Error Collection	–	–	–	–	650	0.790	132	0.671
Opportunities to Respond	417	0.815	24	0.809	650	0.720	132	0.525
Praise					650	0.820	132	0.812
Precorrection	629	0.707	24	0.904	650	0.770	132	0.721
Rules	626	0.813	24	0.938	650	0.900	132	0.827
Behavior-Specific Praise	629	0.798	24	0.651	–	–	–	–
Corrective Feedback	628	0.651	24	0.278	–	–	–	–
Promoting Behavioral Competence	–	–	–	–	650	0.800	–	–
Narrating	–	–	–	–	650	0.800	–	–
Supportive Listening	–	–	–	–	650	0.830	–	–
Monitoring	–	–	–	–	650	0.690	–	–
Modeling	–	–	–	–	650	0.810	–	–
Rehearsal	–	–	–	–	650	0.800	–	–
Visual Cueing	–	–	–	–	650	0.800	–	–
Premack Principle	–	–	–	–	650	0.800	–	–
Tangible Reward	–	–	–	–	650	0.890	–	–
Time-out	–	–	–	–	650	0.950	–	–
Competence Items								
Emotion Regulation	–	–	–	–	125	0.740	7	0.909
Self-Management	–	–	–	–	–	–	14	0.615
Instructional Feedback	223	0.424	5	0.150	179	0.520	57	0.673
Peer Tutoring	–	–	–	–	–	–	11	0.667
Problem-Solving	–	–	–	–	47	0.740	58	0.729
Punishment	–	–	–	–	–	–	13	0.752
Reinforcement	–	–	–	–	–	–	15	0.559
Routines	–	–	–	–	–	–	55	0.723
Social Skills	–	–	–	–	248	0.770	23	0.393
Teacher–Student Relationships	–	–	–	–	422	0.770	114	0.828
Active Supervision	–	–	–	–	–	–	121	0.754
Behavioral Momentum	–	–	–	–	–	–	0	–
Choice	–	–	–	–	–	–	1	–
Error Collection	–	–	–	–	449	0.700	108	0.495
Opportunities to Respond	392	0.533	24	0.746	644	0.780	132	0.480
Praise	–	–	–	–	533	0.730	119	0.703
Precorrection	168	0.413	5	0.732	202	0.560	54	0.746

Table 1 (continued)

	BiCACS		BiCACS—Web		TIMECS		TIES	
	N	*ICC*	*N*	*ICC*	*N*	*ICC*	*N*	*ICC*
Rules	306	0.497	14	0.453	149	0.710	18	0.328
Behavior-Specific Praise	264	0.284	6	0.541	–	–	–	–
Corrective Feedback	230	0.446	2	0.800	–	–	–	–
Promoting Behavioral Competence	–	–	–	–	635	0.800	–	–
Narrating	–	–	–	–	170	0.640	–	–
Supportive Listening	–	–	–	–	239	0.750	–	–
Monitoring	–	–	–	–	643	0.690	–	–
Modeling	–	–	–	–	331	0.590	–	–
Rehearsal	–	–	–	–	56	0.580	–	–
Visual Cueing	–	–	–	–	248	0.580	–	–
Premack Principle	–	–	–	–	58	0.660	–	–
Tangible Reward	–	–	–	–	44	0.760	–	–
Time-out	–	–	–	–	17	0.680	–	–
Student Responsiveness	414	0.608	24	0.687	–	–	132	0.754
Disruptive Behavior	413	0.549	24	0.557	–	–	132	0.647

Note. BEST in CLASS Adherence and Competence Scale (BiCACS); BEST in CLASS Adherence and Competence Scale—Web (BiCACS-Web); Treatment Integrity Measure for Early Childhood Settings (TIMECS); Treatment Integrity Instrument for Elementary School Classrooms (TIES); development of these measures occurred sequentially, beginning with the BiCACS

we began to operationally define each of the practices in order to produce a scoring manual. Using examples from the literature, operational definitions of each practice were created; within the manual, these definitions were preceded by clear scoring procedures (for both adherence and competence). Within each item section of the scoring manual, a list of exemplar examples of the practice was listed, as well as examples and non-examples. Coders then received a brief didactic training (approximately two hours) on the BiCACS and received the scoring manual for reference. Coders used this manual to practice code a number of video-recorded sessions of teachers delivering BEST in CLASS in early childhood classrooms and provided feedback on definitions, exemplars and examples and non-examples resulting in a revised scoring manual. This manual was used in both the initial BEST in CLASS efficacy study (e.g., Sutherland, Conroy, Algina, Ladwig, et al. 2018a; Conroy et al. 2019) and a study examining the efficacy of a Web-based version of BEST in CLASS (Conroy et al. 2020). In addition, following the initial training on using the BiCACS, coders in both studies received a one-hour booster session training at the midpoint of intervention delivery to reduce possible observer drift, answer any questions that arose during coding, and remind coders of observational procedures.

In order to assess reliability, we have coders score the same observation (live or, in the case of training, video-recorded) independently and compare item-level scores. We use intraclass correlation coefficients (ICCs) to assess reliability, which provides an estimate of the ratio of true score variance to total variance, following the guidelines of Cicchetti (1994). Using these guidelines, ICCs greater than 0.75 reflect "excellent" agreement, ICCs between 0.60 and 0.74 reflect "good" agreement, ICCs between 0.40 and 0.59 reflect "fair" agreement and ICCs less than 0.40 reflect "poor" agreement. As given in Table 1 (columns 1 and 2), item-level ICCs for the BiCACS adherence items (Sutherland et al. 2018a) in the initial efficacy study ranged from 0.65 to 0.82, representing "good" to "excellent" agreement, and item-level ICCs for the responsiveness items were "fair" to "good" (0.55 and 0.61; these items were only included in years 3 and 4 of the initial BEST in CLASS efficacy study). Item-level ICCs for the competence scale were lower, ranging from 0.28 to 0.53, representing "poor" to "fair" agreement. All data in Table 1 are from the corresponding intervention trials.

We also used the BiCACS to assess treatment integrity in the BEST in CLASS-Web study (Conroy et al. 2020). This study adapted BEST in CLASS for Web-based delivery and assessed the efficacy of the model in a small, randomized controlled trial. Intraclass correlations using the BiCACS in this study ranged from 0.28 to 0.94 for the adherence items, representing "poor" to "excellent" agreement (Table 1, columns 3 and 4); ICCs for the child responsiveness items ranged from 0.56 to 0.69, representing "fair" to "good" agreement. Competence items ICCs in this study ranged from 0.15 to 0.80, representing "poor" to "excellent" agreement. It is important to note that researchers have found the scoring of competence to be more difficult than the scoring of adherence, with consistently lower ICCs for competence (e.g., Hogue et al. 2008). In our case, the lower competence scores can partially be explained by the scoring method; that is, when an adherence item does not occur, the coders may score it a "1" (*not at all*); however, when adherence is scored a "1," no competence rating can be made. As an example, given the small sample size in the Conroy et al. (2020) study, there were only five instances of precorrection opportunities to rate competence, which may have influenced the low ICC (0.15) noted for this item. In general, there are fewer instances of opportunities to rate competence than there are opportunities to rate adherence. Lower reliability estimates for competence in comparison with adherence and child responsiveness are a consistent finding across our studies.

Treatment Integrity Measure for Early Childhood Settings

While we were pleased about coders' ability to score adherence, competence and child responsiveness on the BiCACS, this measure did not allow for the measurement of the fourth integrity dimension in our conceptual model, treatment differentiation. Moreover, as we became more interested in later stages of the translational research model (i.e., effectiveness and implementation research), it became important for us to be able to assess other practices not prescribed by BEST in CLASS in order to better understand the contexts in which the intervention was being implemented and tested. A measure development grant from IES supported our research team in addressing this measurement limitation via the development of the Treatment Integrity Measure for Early Childhood Settings (TIMECS; McLeod et al. 2020). While the measurement approach (i.e., macroanalytic ratings) remained the same for this measure, a broader number of items were identified to allow for the assessment of treatment differentiation of teacher-delivered practices targeting social, emotional or behavioral outcomes in early childhood classrooms. Our team used a practice elements (Chorpita and Daleiden 2009) approach to develop items for the TIMECS (see McLeod et al. 2017). As the number of items on this measure increased threefold over the BiCACS, we also intensified both our training and coder supervision to support acceptable reliability estimates given the increased load on coders, and below we will describe these procedures.

A goal of the development of the TIMECS was to develop a psychometrically sound tool that could assess teacher delivery of evidence-based practice elements that target social, emotional and behavioral outcomes of young children served in early childhood classroom settings. To do this, we conducted a systematic review of the early childhood literature and distilled practice elements from the practices that comprise evidence-based practices and interventions (McLeod et al. 2017). Five experts in early childhood education rated the practice elements, with all 24 identified practice elements rated as useful or essential (see McLeod et al. 2017 for more detail). Next, practice elements were defined to allow for the measurement of adherence and competence. The same macroanalytic scoring strategy used for the BiCACS was used for the TIMECS, and a scoring manual was produced.

As mentioned earlier, training, checkout and supervision procedures for coders became more intensive given the larger number of items coders needed to reliably score and with the goal of increasing the reliability of individual items. Training for the TIMECS occurred across several steps over a 2-month period of time, and coders were required to achieve item-level ICC reliability of greater than 0.60 before coding could proceed. First coders were trained in the coding procedures and definitions and were provided with the scoring manual. During this training, exemplar items were identified in video examples and practice coding was used to generate questions and discussion. Next, coders began independently coding videos and weekly meetings were held to address coder questions, which were documented using a running record of both questions and decision rules. In the third step, coders began practice coding in early childhood classrooms in order to orient themselves to live coding; last, coders independently coded 40 master-coded videos and were required to achieve greater than "good" reliability (ICC > 0.60; Cicchetti 1994) on each item before they could begin independently coding in early childhood classrooms. Once live coding began, coders met weekly with trainers to answer questions and review ICC data to prevent coder drift.

These training procedures resulted in coders being able to reliably code items on the TIMECS (see McLeod et al. 2020 for more detail). All of the adherence items scored "good" or better, with item-level ICCs ranging from 0.68 to 0.95. Overall, the competence item ICCs were lower than the adherence item ICCs, ranging from 0.52 to 0.80, with 17 of the 21 items scoring "good" or better.

Treatment Integrity Instrument for Elementary School Classrooms

As we completed work on developing the TIMECS, our research team received funding from IES to adapt BEST in CLASS for use in early elementary school classrooms. As a result of this adaptation, we needed to develop integrity measures to assess the core components of BEST in CLASS—Elementary and also wanted to use what we had learned in developing the TIMECS to be able to also assess treatment differentiation, in addition to student responsiveness. The measure development of the Treatment Integrity

Instrument for Elementary School Classrooms (TIES) largely mirrored the work done on the TIMECS and is described below.

First, we used the practice elements approach to identify common teacher-delivered practices in evidence-based programs and interventions in early elementary school (see Sutherland et al. 2019). After these practice elements were identified and reviewed by experts, we created a scoring manual using the same macroanalytic approach used in the previously described measures. The final TIES measure includes 18 items, each of which is scored on two dimensions: adherence and competence. Of these 18 items, 6 are prescribed by BEST in CLASS—Elementary (supportive relationships, emotion regulation, rules, precorrection, opportunities to respond and praise), while 12 are not part of the training and coaching of BEST in CLASS—Elementary and are used to assess treatment differentiation (e.g., self-management, problem-solving). In addition, two items are used to assess student responsiveness: responsiveness and disruptions.

Training, checkout and supervision of coders were similar to training for the TIMECS. Training for the TIES occurs across several steps over approximately a 2-month period of time, and coders are required to achieve item-level ICC reliability of greater than 0.60 before live coding can proceed. First coders are trained in the coding procedures and definitions and are provided with the scoring manual during an initial two-hour meeting. Exemplar items are identified using video examples, and practice coding is used to generate questions and discussion. Next, coders begin coding videos in pairs to generate questions for weekly meetings, where a running record of questions and decisions is maintained. This phase lasts approximately two weeks and is followed by coders independently coding videos and weekly group meetings with trainers to address coder questions, which are also documented using a running record of both questions and decision rules. Last, coders independently code nine master-coded videos and are required to achieve greater than "good" reliability (ICC > 0.60; Cicchetti 1994) on each item before they can begin independently coding in elementary classrooms. Once live coding begins, ICC data are reviewed in order to identify any drift or coding problems. For this project, TIES data are collected at pretest, midpoint of intervention, posttest and maintenance. Prior to the midpoint data collection, a booster training is held and coders are required to check out on three videos, with greater than "good" reliability across all items.

One difference between the integrity observations in this project and the previous BEST in CLASS trial (Sutherland et al. 2018a) is that observers in this study are blind to condition. Initial reliability data from the first two years of the BEST in CLASS—Elementary study are promising; 16 of the 18 adherence items scored "good" or better, with item-level ICCs ranging from 0.53 to 1.00. (One of the items, Behavioral Momentum, was never observed, which resulted in perfect agreement.) Overall, the competence item ICCs were lower than the adherence item ICCs, ranging from 0.39 to 0.91 (not counting Choice, which was only observed once), with 11 of the 18 items scoring "good" or better. The student responsiveness item ICCs were 0.75 and 0.65, for responsiveness and disruptions, respectively.

Discussion

The purpose of this article was to describe a framework for developing treatment integrity measures and the processes used to develop a suite of measurement tools to assess multiple dimensions of teacher integrity of delivery of practices in early childhood and elementary classrooms to support the social and emotional learning of young children and students who demonstrate chronic problem behavior. This research led to the development of integrity tools to reliably assess adherence, competence, differentiation and child responsiveness. In the following sections, we will provide recommendations for measuring treatment integrity of teacher-delivered practices and programs as well as future directions to advance intervention research in the delivery of social, emotional and behavioral support programming.

Recommendations

We have several recommendations for the development and use of treatment integrity measures, particularly for teacher-delivered interventions. First, researchers should use conceptual models to help guide their measure development. While developing common measures to assess child and teacher outcomes is a priority for the field, integrity measures are often inextricably linked to the intervention being delivered. While a common elements approach (McLeod et al. 2017; Sutherland et al. 2019) is promising for identifying common practice elements delivered by teachers, researchers often create their own integrity measures linked to core components or active ingredients of their particular intervention. In this case, having clear conceptual models to guide this work is critical; however, there is a lack of conceptual models of treatment integrity in school-based intervention research to guide investigators (see Sanetti and Kratochwill 2009). That said, using examples such as the model we provided earlier or models from other fields (e.g., Berkel et al. 2011; McLeod et al. 2013) may help researchers use conceptual models to guide their treatment integrity measurement approach.

One reason that conceptual models are critical to teacher-delivered interventions is that there are so many potential factors associated with intervention delivery in the complexity of classroom environments (see Durlak 2010; Sanetti and Kratochwill 2009). Thus, it is important that researchers identify both factors hypothesized to influence teacher integrity of delivery, but also different dimensions of integrity that are to be assessed. School-based intervention research has lagged behind other fields in assessing multiple dimensions of treatment integrity with psychometrically sound measurement tools (Sanetti and Kratochwill 2008, 2009). For example, given the value-added nature of BEST in CLASS, it became clear to our research team that assessing adherence and competence of delivery in both treatment and BAU classrooms was important, as was the measurement of treatment differentiation and child responsiveness.

Another underreported dimension of treatment integrity that we have only recently begun to assess is child responsiveness. Indeed, interventions that target social, emotional and behavioral outcomes of young children and students who have chronic problem behavior, such as BEST in CLASS, must assess child responsiveness given that a presenting problem of these children are behaviors consistent with non-responsiveness (i.e., poor engagement, disruptive behaviors) to interventions. Thus, while teachers may deliver an intervention with high adherence and competence, if the focal child does not “receive” the intervention it will not be effective. To further illustrate, in a recent small trial of BEST in CLASS—Elementary we found that student responsiveness to teacher attempts to use BEST in CLASS practices increased from pretest to posttest for students in the treatment group, but decreased from pretest to posttest for students in the BAU condition, providing a potential explanation for treatment effects (Sutherland et al. 2020). This example highlights the importance of not only assessing child responsiveness but also assessing the dimensions of treatment integrity that make up a conceptual framework in both treatment and BAU classrooms. This illustration also highlights how treatment integrity measures may differ across different phases of the translational pipeline. To illustrate, early in our intervention work it was important for us to assess how much, and how well, teachers were delivering the BEST in CLASS practices (e.g., Conroy et al. 2015); in later work, it became important to assess other dimensions of treatment integrity, such as child responsiveness, in order to assist us in interpreting findings from efficacy trials (e.g., Sutherland et al. 2020).

Another recommendation involves the training of observers to collect integrity data. We have learned a great deal over the past decade or so in our own work, and our training approach has changed over time to reflect these lessons. While we have always relied primarily on direct observation of teacher delivery of practices to assess treatment integrity, the depth and amount of time we have spent training observers have improved and increased and these improvements are reflected in our reliability estimates (Table 1) even when we have increased the number of items on our adherence and competence scales. The importance of several aspects of our training has become apparent and is likely to be useful for other researchers assessing treatment integrity in classrooms and schools.

First, it is critical to have clearly written manuals that outline procedures for both collecting data in classrooms and detailed operational definitions of codes including examples and non-examples. These manuals become integral in helping make coding decisions to address questions that arise during training, and serve as working documents during training via the additions and clarifications of examples and non-examples of codes. In addition, we have found it helpful to have observers read through the manual at the end of data collection to provide suggestions on improving clarity and detail in both procedures and definitions of codes for future use.

Second, it is important to have quality didactic training of coders initially to orient them to the coding system, and having video exemplars to share during this time is particularly important. Over time we have been fortunate to collect a library of tapes of teachers delivering BEST in CLASS (as well as some videos of BAU classrooms) which has allowed us to use these tapes in training and checkout. Further, having master-coded tapes allows us to ensure that coders are reliable at the item level before they begin collecting data in classrooms. Relatedly, we enter observational data into our database as soon as possible and this allows for ongoing reliability checks to identify any potential trouble spots that need attention. Last, recalibration training at the midpoint of data collection, after observers have not collected data in classrooms for a period of time, serves to reorient coders to item definitions and procedures and requires that they are reliable on codes before reentering classrooms. All of these training and supervision strategies have played a significant role in our ability to reliably code teacher delivery of practices in early childhood and elementary classrooms.

Limitations

There are several limitations to keep in mind regarding the development of the treatment integrity measures included in this article. First, as with other treatment integrity measures (e.g., Hogue et al. 2008), the reliability of the competence items is lower than those of the adherence items and in some cases is poor (i.e., ICC < 0.40). That said, as we intensified the training and checkout procedures across time we were able to increase the competence item ICCs. However, this process does raise another limitation of the approach taken in our measure development process—resources. It is costly and time-consuming to train to reliability and conduct direct observations in classrooms, and the field needs common measurement tools with acceptable reliability to increase our ability to collect treatment integrity data in classrooms. Finally, while we used the broader literature (McLeod et al. 2017; Sutherland et al. 2019) to identify practice elements for items on the integrity measures and anticipate that they are representative of the broader field, data reported in this paper were collected in two regions of the USA and thus may not be representative of other geographic regions.

Future Directions

In addition to the recommendations for measuring integrity of teacher-delivered interventions targeting social, emotional and behavioral outcomes, there are ways that research can advance intervention science in the delivery of social, emotional and behavioral support programming via treatment integrity measurement. First, while the common elements approach has much promise for measure development, it also has significant implications for better understanding relations between treatments and outcomes. If researchers are able to independently assess the core components of interventions across a variety of dimensions of treatment integrity (e.g., adherence, competence), then we may be able to determine which components are more, or less, likely to be associated with treatment outcomes. Thus, we may be able to make multi-component interventions more efficient by focusing training and coaching efforts on those components that have the greatest association with positive effects.

While direct observations are considered the gold standard in treatment integrity measurement (Sanetti and Kratochwill 2009; Sutherland et al. 2013a, b), they are time-consuming and expensive (Hogue et al. 2014; Schoenwald et al. 2011), only assess practices that are observable and frequent (McLeod et al. 2009) and often are intrusive to learning contexts (Yoder and Symons 2010). Thus, one goal for the field is to develop psychometrically sound teacher reports of their use of practices within intervention models to provide a measure of treatment integrity (see McLeod et al., this issue). Teacher report measures are cost and time-effective and may allow for the assessment of teacher use of practices (e.g., scaffolding) that are not observable. Teacher report measures of treatment integrity would also enhance implementation research. To illustrate, having psychometrically sound teacher report measures would allow for more frequent integrity checks during both implementation and sustainment phases, particularly in large-scale studies where direct observations are not feasible. That said, developing psychometrically sound observational tools of treatment integrity is a critical step toward developing teacher report measures, providing a gold standard metric for comparison during the development process.

Conclusion

Demonstrating that evidence-based programs and practices that target social, emotional and behavioral outcomes are delivered as intended in early childhood and elementary classrooms is critical to scaling up programs and practices (Schoenwald et al. 2011). In addition, demonstrating that teachers are implementing practices with adherence and competence is also critical to efforts to scale implementation supports such as coaching models (Schoenwald et al. 2012), and evaluating treatment integrity thus represents an important outcome in implementation research (Proctor et al. 2011). However, in order to assess teacher integrity of implementation it is necessary to have valid and reliable measures of multiple dimensions of treatment integrity grounded in conceptual models. This article described the iterative development of a suite of treatment integrity measures that assess multiple dimensions of treatment integrity of teacher delivery of evidence-based practices and hopefully will be useful for other school-based researchers interested in improving the delivery and access to programs and practices that improve the social, emotional and behavioral outcomes of children and students.

Funding This research was supported by Grants from the U.S. Department of Education's Institute for Education Sciences (R324A11017, R305A140487, R305A150246, R324A160158 and R305A180182). The opinions expressed by the authors are not necessarily reflective of the position of or endorsed by the U.S. Department of Education.

Compliance with Ethical Standards

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

Ethical 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.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Barber, J. P., Mercer, D., & Krakauer, I. (1996). Development of an adherence/competence rating scale for individual drug counseling. *Drug and Alcohol Dependence*, *43*, 125–132.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science, 12*(1), 23–33. https://doi.org/10.1007/s11121-010-0186-1.

Botvin, G. J., & Griffin, K. W. (2004). Life skills training: Empirical findings and future directions. *Journal of Primary Prevention, 25*(4), 211–232.

Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge: Harvard University Press.

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.

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*(3), 566. https://doi.org/10.1037/a0014565.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284. https://doi.org/10.1037/1040-3590.6.4.284.

Conroy, M. A., Sutherland, K. S., Algina, J., Ladwig, C., Werch, B., Martinez, J., et al. (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. https://doi.org/10.17105/SPR-2018-0003.V48-1.

Conroy, M. A., Sutherland, K. S., Algina, J., Werch, B., & Ladwig, C. (2018). Prevention and treatment of externalizing problem behaviors in young children: Clinical implications from a randomized controlled trial of BEST in CLASS. *AERA Open, 4*(1), 1–16. https://doi.org/10.1177/2332858417750376.

Conroy, M. A., Sutherland, K. S., Granger, K. L., Marcolouides, K. M., Feil, E., Huang, K., & Montession, A. (2020). *Preliminary study of the effects of BEST in CLASS-Web on young children's social-emotional and behavioral outcomes.* Manuscript submitted for publication.

Conroy, M. A., Sutherland, K. S., Wilson, R. E., Martinez, J., Whalon, K. J., & Algina, J. (2015). Measuring teacher implementation of the BEST in CLASS intervention program and corollary child outcomes. *Journal of Emotional and Behavioral Disorders, 23,* 144–155. https://doi.org/10.1177/1063426614532949.

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., et al. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework Advances in School Mental. *Health Promotion, 1*(3), 6–28. https://doi.org/10.1080/1754730X.2008.9715730.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "Implementation research in early childhood education." *Early Childhood Research Quarterly, 25*(3), 348–357. https://doi.org/10.1016/j.ecresq.2010.03.003.

Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prevention Science, 16*(8), 1123–1127. https://doi.org/10.1007/s11121-015-0606-3.

Evans, M., Piasecki, J., Kriss, M., & Hollon, S. (1984). *Rater's manual for the collaborative study psychotherapy rating scale- form 6 (CSPRS-6)*. University of Minnesota and St. Paul-Tamsey Medical Center. Unpublished manuscript.

Han, S. S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology, 33*(6), 665–679. https://doi.org/10.1007/s10802-005-7646-2.

Hogue, A., Dauber, S., Lichvar, E., Bobek, M., & Henderson, C. E. (2014). 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*(2), 229–243. https://doi.org/10.1007/s10488-014-0548-2.

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*(2), 332. https://doi.org/10.1037/0033-3204.33.2.332.

Hollon, S., Evans, M., Auerbach, A., DeRubeus, R., Elkin, I., Lowerry, A., et al. (1988). *Development of a system for rating therapies for depression: Differentiating cognitive therapy, interpersonal psychotherapy, and clinical management pharmacotherapy*. Vanderbilt University, Nashville, TN. Unpublished manuscript.

Greenberg, L. S. (1986). Change process research. *Journal of Consulting and Clinical Psychology, 54,* 4–9.

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.

McLeod, B. D., Kunemund, R., Nemer, S. L., & Lyon, A. R. (2020). Leveraging Implementation Science and Practice to Support the Delivery of Evidence-Based Practices in Services for Youth with Emotional and 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. 417–432). New York: Routledge/Taylor & Francis.

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

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review, 38*(4), 541.

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., & Weisz, J. R. (2010). The therapy process observational coding system for child psychotherapy strategies scale. *Journal of Clinical Child & Adolescent Psychology, 39*(3), 436–443. https://doi.org/10.1080/15374411003691750.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*(4), 365–383. https://doi.org/10.1093/clipsy.bpi045.

Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., et al. (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.

Sanetti, L. M. H., 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. https://doi.org/10.1002/pits.22364.

Sanetti, L. M. H., Dobey, L. M., & Gritter, K. L. (2012). Treatment integrity of interventions with children in the Journal of Positive Behavior Interventions from 1999–2009. *Journal of Positive Behaviour Interventions, 14,* 29–46. https://doi.org/10.1177/1098300711405853.

Sanetti, L. M. H., Gritter, K. L., & Dobey, L. M. (2011). Treatment fidelity of interventions with children in the school psychology literature from 1995–2008. *School Psychology Review, 40,* 72–84.

Sanetti, L. M. H., & Kratochwill, T. R. (2008). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy, 4*(1), 95–114. https://doi.org/10.1037/h0100835.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review*, *38*(4), 445–459.

Schoenwald, S. K. (2011). It's a bird, it's a plane, it's… fidelity measurement in the real world. *Clinical Psychology: Science and Practice, 18*(2), 142–147. https://doi.org/10.1111/j.1468-2850.2011.01245.x.

Schoenwald, S. K., Mchugh, R. K., & Barlow, D. H. (2012). The science of dissemination and implementation. In R. K. McHugh & D. H. Barlow (Eds.), *Dissemination and implementation of evidence-based psychological interventions* (pp. 16–42). Oxford: Oxford University Press.

Sutherland, K. S., Conroy, M. A., Abrams, L., Vo, A., & Ogston, P. (2013a). An initial evaluation of the Teacher-Child Interaction Direct Observation System: Measuring teacher-child interaction behaviors in classroom settings. *Assessment for Effective Intervention, 39,* 12–23. https://doi.org/10.1177/1534508412463814.

Sutherland, K. S., Conroy, M. A., Algina, J., Ladwig, C., Jessee, G., & Gyure, M. (2018a). 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. L. (2020). BEST in CLASS A Tier-2 program for children with and at risk for emotional/behavioral disorders. In T. W. Farmer, M. A. Conroy, E. M. Z. Farmer, & K. S. Sutherland (Eds.), *Handbook of Research on Emotional and Behavioral Disorders* (pp. 214–226). Abingdon: Routledge.

Sutherland, K. S., Conroy, M. A., McLeod, B. D., Algina, J., & Kunemund, R. L. (2018b). Factors associated with teacher delivery of a classroom-based Tier 2 prevention program. *Prevention Science, 19*(2), 186–196. https://doi.org/10.1007/s11121-017-0832-y.

Sutherland, K. S., Conroy, M. A., McLeod, B. D., Algina, J., & Wu, E. (2018c). Teacher competence of delivery of BEST in CLASS as a mediator of treatment effects. *School Mental Health, 10*(3), 214–225. https://doi.org/10.1007/s12310-017-9224-5.

Sutherland, K. S., Conroy, M. A., McLeod, B. D., Kunemund, R., & McKnight, K. (2019). Common practice elements for improving social, emotional, and behavioral outcomes of young elementary school students. *Journal of Emotional and Behavioral Disorders, 27*(2), 76–85. https://doi.org/10.1177/1063426618784009.

Sutherland, K. S., McLeod, B. D., Conroy, M. A., Abrams, L. M., & 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., McLeod, B. D., Conroy, M. A., & Cox, J. R. (2013b). Measuring implementation of evidence-based programs targeting young children at risk for emotional/behavioral disorders: Conceptual issues and recommendations. *Journal of Early Intervention, 35*(2), 129–149. https://doi.org/10.1177/1053815113515025.

Vroom, E. B., Massey, O. T., Yampolskaya, S., & Levin, B. L. (2020). The Impact of Implementation Fidelity on Student Outcomes in the Life Skills Training Program. *School Mental Health, 12*(1), 113–123. https://doi.org/10.1007/s12310-019-09333-1.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of consulting and clinical psychology, 61*(4), 620. https://doi.org/10.1037/0022-006X.61.4.620.

Williford, A. P., Wolcott, C. S., Whittaker, J. V., & Locasale-Crouch, J. (2015). Program and teacher characteristics predicting the implementation of Banking Time with preschoolers who display disruptive behaviors. *Prevention Science, 16*(8), 1054–1063. https://doi.org/10.1007/s11121-015-0544-0.

Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. Berlin: Springer Publishing Company.

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