2021

A Psychometric Evaluation of the BEST in CLASS Adherence and Competence Scale

Katrina A. Markowicz
Virginia Commonwealth University

A Psychometric Evaluation of the BEST in CLASS Adherence and Competence Scale

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science
at Virginia Commonwealth University.

by

Katrina Ashley Markowicz
Bachelor of Science, Wayne State University, 2014

Director: Bryce D. McLeod, Ph.D.
Professor, Department of Psychology

Virginia Commonwealth University
Richmond, Virginia
May 2021

Acknowledgments

I would like to express my gratitude to several people who have been key supports to me in accomplishing this thesis milestone and during my graduate education thus far. I could not have completed this project without the mentorship of my research advisor, Dr. Bryce McLeod. Thank you, Bryce, for your wisdom, guidance, and feedback in crafting this project, but also for your holistic support, patience, and encouragement. I am thankful for an advisor who expands my research knowledge, challenges my thought process, continues to help me grow as a writer, and helps me hone in my research interests. Thank you to my thesis committee, Drs. Kevin Sutherland and Joshua Langberg, for kindly agreeing to serve on my committee, as well as your support, feedback, and direction on this project. Of course, thank you to the BEST in CLASS team. Drs. Kevin Sutherland and Maureen Conroy, thank you for the opportunity to join the team. I am thankful for the use of this data to pursue a thesis project I am passionate about. Thank you, Dr. Kristen Granger, for providing me with statistical guidance on the MPlus software. To the coders and staff who have helped code and compile the data of which this project has used, I thank you for your contributions. To my lab, thank you for your feedback, suggestions, and support during brainstorming meetings and presentations of my thesis project. Lastly, I would like to extend a thank you to my partner, family and friends in their unwavering love, support and accountability. I appreciate each one of you who have uniquely supported my dreams and passions. I am grateful for everything you all have done to support me, including words of encouragement, long days of writing together, mailing me my favorite Midwestern snacks, and making sure I always have a warm cup of tea or coffee nearby.

Table of Contents

List of Tables

List of Figures

Abstract

A PSYCHOMETRIC EVALUATION OF THE BEST IN CLASS ADHERENCE AND COMPETENCE SCALE

By Katrina A. Markowicz, B.S.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2021

Major Director: Bryce D. McLeod
Professor, Department of Psychology

It is critical to utilize treatment integrity instruments to support the evaluation of evidence-based programs in early classroom contexts. However, in the early childhood field, guidelines for collecting treatment integrity data are underdeveloped. Consequently, most treatment integrity instruments employed in the field solely assess adherence, vary in design features and have little psychometric evidence supporting their use. As such, this represents a gap in the field that might slow efforts to implement evidence-based programs. The current study examines the score reliability and validity of an observational treatment integrity instrument (The BEST in CLASS Adherence and Competence Scale [BiCACS]; Sutherland et al., 2014). The BiCACS is designed to assess adherence and competence of the practices found in the BEST in CLASS program, a teacher-delivered evidence-based program for children at-risk for emotional and behavioral disorders. Data were drawn from observations of 179 teachers who were randomized to BEST in CLASS (n = 89) or business-as-usual (n = 90) and 416 children (n = 211 in the BEST in CLASS condition; n = 205 in the business-as-usual condition) at risk for emotional and behavioral disorders. Based on double-coded observations (25% of sample) the mean single-measure intraclass correlation (ICC[2,1]) was .74 (*SD* = 0.06) for the Adherence items and .46 (*SD* = 0.14) for the Competence items. The ICC(2,1) for the Adherence and Competence subscales were .81 and .43, respectively. Findings also suggested initial evidence of convergent and discriminant validity at the BiCACS item and subscale-levels. The magnitude of correlations among the BiCACS items suggests that the adherence and Competence items overlap the most with items within the same subscale, but also measure distinct BEST in CLASS practices. At the subscale level, the correlation among the Adherence and Competence items are more related to each other than their correlations with scores on measures of child responsiveness, child engagement, closeness, and conflict of student-teacher relationships. Validity evidence at the subscale level suggests that the BiCACS can distinguish between intervention groups and detect change over time. The reliability and validity findings support the use of the BiCACS as a program evaluation instrument. Although, future research is still needed to replicate these findings and test the construct validity of the BiCACS with other instruments that assess adherence and competence. Still, results provide valuable information about the psychometric properties of a treatment integrity instrument used in early classroom contexts and inform the growing knowledge of this area in the field.

Statement of Purpose

Many young children in early childhood classroom contexts exhibit high levels of behavioral problems that put them at risk for more impairing mental health issues later in life (Barbarin, 2007; Brauner & Stephens, 2006; Campbell et al., 2000; Mesman et al., 2001). Fortunately, teacher-delivered evidence-based programs exist for young children displaying behavioral difficulties in early childhood classrooms. Examples of these programs include the *Incredible Years* (Webster-Stratton et al., 2004) and *Preschool PATHS* (Domitrovich et al., 2007). However, despite the potential promise of these programs, there are various factors (e.g., the intensity of child problem behavior, level of teacher training; Pas et al., 2015; Pianta & Rimm-Kaufman, 2006) that make it difficult for teachers to deliver these programs in their classrooms. Notably, these factors may cause variation in how evidence-based programs are delivered, which may undermine program effects.

In part to help standardize program delivery, Carroll and Nuro (2002) proposed guidelines to support the development and evaluation of evidence-based programs that include four elements: treatment manuals, well-defined treatment population, standardized training, coaching procedures, and treatment integrity instruments. Each of these four elements intends to ensure programs are delivered consistently across contexts. In the early childhood literature, most randomized-controlled trials have utilized three of the four elements, but they lack standardized treatment integrity instruments (Sanetti et al., 2011; Sanetti et al., 2020).

Treatment integrity is defined as the degree to which an evidence-based program is delivered as intended (McLeod et al., 2009; Sanetti & Kratochwill, 2009). Two components of treatment integrity are beneficial for assessing evidence-based programs delivered in early classroom contexts: *adherence* and *competence*. Adherence refers to how frequent and thorough a teacher delivers the practices in a treatment protocol (Sutherland et al., 2013). Competence describes the level of skill and degree of responsiveness with which the teacher delivers the core components specified in the protocol (Schulte et al., 2009). These two components of treatment integrity are critical for interpreting the outcomes of an intervention, evaluating the efficacy of evidence-based programs in randomized controlled trials, and assessing whether providers of evidence-based programs can establish and maintain treatment integrity over time (Cross & West, 2011; Fiske, 2008; Noell et al., 2005).

Assessing adherence and competence in randomized-controlled trials is critical as it aids in the interpretation of research findings. For example, when evidence-based programs are evaluated using treatment integrity instruments, they help ascertain that the program was delivered as designed, permitting researchers to better attribute promising intervention outcomes to the program's effects (Schoenwald et al., 2011). However, when unfavorable results are achieved and adherence and competence are assessed, researchers can then decipher whether the poor intervention outcomes were because the program did not work or whether the program was delivered a with variation (Perepletchikova et al., 2007; Sanetti & Kratochwill, 2009).

Although assessing treatment integrity provides valuable information about program delivery, measurement and design guidelines of treatment integrity instruments within early childhood contexts in the early educational field have not been fully established (Sanetti et al., 2020, Sutherland et al., 2013; Wolery, 2011). Early childhood literature suggests that there is variation among the response format of the extant treatment integrity instruments, but observer report instruments instead of self-report seem to be most common (e.g., Bierman et al., 2008; Hamre et al., 2012; Sanetti et al., 2020; Sutherland et al., 2018). For program evaluation, observational treatment integrity instruments that assess both adherence and competence using

Likert-type scales have certain benefits. For example, observational instruments typically exhibit more psychometric evidence (Hansen et al., 1991; Harachi et al., 1999; Perepletchikova, 2007). Also, observational instruments are often preferred over self-report instruments to provide a specific and objective assessment of behavior because self-report instruments may suffer from bias of one's own behavior (Hogue et al., 1996; Martinez et al., 2014; Schoenwald et al., 2011). Additionally, Likert-type scales are the ideal response-format over checklist or frequency count response formats as they capture variation in delivery that is critical to tracking changes over time (Wolery, 2011).

Although these treatment integrity features are ideal, the scores produced by the instrument must demonstrate reliability and validity. Currently, treatment integrity instruments that assess adherence and competence that possess evidence of score reliability and validity are wanting in the early educational field (Sutherland et al., 2013, Sanetti et al., 2020). To my knowledge, only five randomized-controlled trials evaluating an evidence-based program delivered in early classroom settings for children at risk for emotional and behavioral disorders utilize treatment integrity instruments with a report of the reliability or the validity of the instrument (e.g., Conroy et al., 2018; Barnett et al., 2008; Feil et al., 2014; Hemmeter et al., 2016; Sutherland et al., 2018). Additionally, only three of these studies utilize instruments that are observational and use a Likert-type response format (i.e., Conroy et al., 2018; Feil et al., 2014; Sutherland et al., 2018). Therefore, there is a need for treatment integrity instruments in the early educational field that demonstrate score reliability and validity.

Treatment integrity instruments must exhibit specific properties of score reliability and validity to be used for program evaluation (Gresham, 2009; Mowbray et al., 2003; Sanetti & Kratochwill, 2009; Sheridan et al., 2009). These instruments must possess evidence of interrater reliability, meaning that scores produced by two independent raters should be consistent with each other (Kazdin, 2016). These instruments should also possess evidence of score validity, including evidence of representative (i.e., whether scores represent the constructs) and elaborative validity (i.e., whether scores have utility in assessing the constructs; Foster & Cone, 1995). Evidence for construct validity occurs when scores indicate that an instrument assesses the intended construct it was designed to assess (Foster & Cone, 1995). This type of validity is achieved through evidence of convergent and discriminant validity. Evidence of convergent validity is established by demonstrating that scores of one instrument are highly correlated with another instrument designed to assess the same construct (DeVellis, 2017; Furr & Bacharach, 2014; McLeod et al., 2013). In contrast, discriminant validity is evidenced through small to moderate correlations between the instrument and another instrument that assesses an unrelated construct (i.e., discriminant validity; DeVellis, 2017; Furr & Bacharach, 2014; McLeod et al., 2013). Without evidence of construct validity, it is possible that the instrument may be assessing a different construct than the one intended (Hill & Lambert, 2004).

One way to establish elaborative validity is to produce evidence of discriminative validity. Discriminative validity is established when an instrument can distinguish between groups expected to differ (McLeod et al., 2015, 2018). For example, a treatment integrity instrument would be expected to differentiate between a group of teachers who were trained and coached on the delivery of an evidence-based program and a group of teachers who did not receive this training or coaching (Carroll et al., 2000). Evidence of discriminative validity suggests that the scores produced by the instrument can detect differences in treatment conditions. This is important because if the instrument cannot detect these effects, it is difficult to determine whether an intervention was delivered as intended.

Measurement sensitivity represents another dimension of elaborative validity. Measurement sensitivity provides evidence that the scores on an instrument can detect program effects such that the scores produced by the instrument can detect changes over the course of the program (Kazdin, 2016). Evidence of measurement sensitivity is essential for a treatment integrity instrument because it affords researchers the ability to determine whether adherence and competence scores changed during training and coaching. Additionally, treatment integrity instruments must be able to detect changes over a short time (e.g., two weeks), thus suggesting that the instrument can capture changes in the delivery of a program.

Given the lack of treatment integrity instruments that demonstrate score reliability and validity in the early childhood literature, the current study aims to evaluate the psychometric properties of a treatment integrity instrument designed to support program evaluation. The treatment integrity instrument under evaluation is the BEST in CLASS Adherence and Competence Scale (BiCACS; Sutherland & McLeod, 2010). The BiCACS is an observational treatment integrity instrument with a Likert-type scale that assesses adherence and competence of the core components of BEST in CLASS. BEST in CLASS is a teacher-delivered evidence-based program delivered in early classroom contexts and targets children at risk for emotional and behavioral disorders. Investigations of BEST in CLASS have produced promising results for its use in early classroom settings in randomized-controlled trials (see Conroy et al., 2018; Sutherland et al., 2018).

To achieve the current study’s aims, interrater reliability, construct validity, discriminative validity, and measurement sensitivity of the BiCACS will be assessed. These psychometric domains will be examined by evaluating BiCACS ratings collected by observational coders. Additionally, these ratings were collected from teachers who were trained and then coached on the BEST in CLASS program in addition to teachers who did not receive training nor coaching (i.e., business as usual). More specifically, the aims of the current study attempt to evaluate whether the BiCACS: (a) produces reliable scores of adherence and competence at the item and subscale-level; (b) produces a pattern of correlations consistent with past research that provide evidence at the item and subscale-levels; (c) identifies group differences in levels of adherence and competence; and (d) demonstrates a positive change over time utilizing adherence and Competence subscale scores.

Literature Review

Tier-2 teacher-delivered evidence-based programs exist for young children who exhibit problem behaviors in early classroom contexts, and these programs have demonstrated favorable results in randomized-controlled trials (e.g., Bierman et al., 2008; Domitrovich et al., 2007; Webster-Stratton et al., 2004). Tier-2 interventions are aimed at ameliorating problem behavior for children at-risk for emotional and behavioral disorders. It is fortunate that these evidence-based programs exist as many young children attending early childhood programs exhibit high levels of problem behaviors that place them at increased risk for emotional and behavioral disorders (Basten et al., 2016). Children who exhibit problem behaviors at an early age are at increased risk for more severe problem behaviors and clinical psychiatric disorders that manifest in their later years of childhood and late adolescence (Brennan et al., 2012; Campbell et al., 2000; Finsaas et al., 2018; Keller et al., 2005; Mesman et al., 2001).

Though tier-2 teacher-delivered evidence-based programs help mitigate problem behavior for children at-risk for emotional and behavioral disorders, barriers exist that hinder the delivery of these programs. The complex nature of these programs and the context in which the programs are delivered can make it difficult for teachers to accurately and adequately deliver the programs (Durlak, 2010). For example, various factors may influence how well the program is delivered, such as the level and type of teacher training (Pianta & Rimm-Kaufman, 2006), quality of teacher-child relationships (Driscoll & Pianta, 2010), and teachers' instructional ability (Domitrovich et al., 2010; Hamre et al., 2010). Further, contextual factors may play a role in hindering the delivery of these programs, including the intensity of child problem behavior (Pas et al., 2015), teacher education level, and experience (Domitrovich et al., 2008, Durlak, 2010), principal leadership (Kam et al., 2003), and school size (Pas et al., 2015). As such, there may be variation in the way the program is delivered across different children, teachers, classrooms, and schools.

In the early childhood field, guidelines are underdeveloped that are aimed at reducing the variation in the delivery of evidence-based practices. To support the development and evaluation of evidence-based programs delivered in early classroom settings, four elements are needed: treatment manuals, a well-defined treatment population, standardized training and coaching procedures, and treatment integrity instruments (see Carroll & Nuro, 2002). Many studies that evaluate evidence-based programs meet three of these criteria, but they lack standardized treatment integrity instruments (Sanetti et al., 2011; Sanetti et al., 2020). In order to support the development and evaluation of evidence-based programs, treatment integrity instruments are needed.

Treatment integrity (also referred to as treatment fidelity, treatment adherence, intervention integrity, and procedural reliability; Noell & Gansle, 2014; Sanetti & Fallon, 2011) is a broad term used to refer to the degree to which an evidence-based program was delivered as intended (McLeod et al., 2009; Sanetti & Kratochwill, 2009). Two components of treatment integrity can be particularly helpful to assess within the context of randomized-controlled trials in order to determine if program outcomes were detected due to the program itself or because of variation in the delivery of the program. These components include adherence and competence. *Adherence* is defined as the extent to which a teacher delivers the core practices found in the treatment protocol (Sutherland et al., 2013). *Competence* is defined as the level of skill and Responsiveness through which the teacher delivers the core practices found in a treatment protocol (Schulte et al., 2009; Sutherland et al., 2013). Treatment integrity instruments that capture both components of adherence and competence are particularly useful when evaluating programs during efficacy trials (Schoenwald et al., 2011).

To support the evaluation and delivery of evidence-based programs in early classroom contexts, it is important to utilize treatment integrity instruments that assess how the program was delivered. Across fields (e.g., mental health, medical, education, school psychology) researchers often fail to report treatment integrity data (Perepletchikova et al., 2007; Sanetti et al., 2013; Sanetti et al., 2020; Schoenwald et al., 2011). However, recent trends suggest that in education, it is becoming more common to report treatment integrity data. For example, recent reviews suggest that 72.8% of published studies in in school psychology and education-related journals that evaluate interventions delivered within school settings report treatment integrity data (Sanetti et al., 2020). This is an increase of 35% over the past decade (Sanetti et al., 2011). However, the majority of the studies that report treatment integrity data tend to report on adherence (i.e., 98.7% of studies), whereas competence data is still underreported (i.e., 8.7%). The underutilization of instruments that assess multiple treatment integrity components is a disservice to these fields because, as stated above, treatment integrity instruments have important implications for the interpretation of study findings (Perepletchikova et al., 2007; Sanetti & Kratochwill, 2009) and the assessment of the outcomes of training and coaching (Hemmeter et al., 2015; Noell et al., 2005).

Treatment Integrity Measurement in Early Childhood Literature

Research in the education literature suggests that treatment integrity methods are not consistently reported. Within the educational literature, a clear and consistent conceptual understanding of treatment integrity is lacking such that researchers use different terms (e.g., integrity, adherence, fidelity) when talking about treatment integrity and include different components when assessing treatment integrity (Sanetti et al., 2020; Sanetti & Kratochwill, 2009; Schulte et al., 2009). Most commonly, a single treatment integrity component is assessed, which evaluates the degree to which an intervention or program was delivered (i.e., adherence; Noell et al., 2005). To determine if similar rates and methods of assessing of treatment integrity occurred within the early childhood literature, an abbreviated literature review was conducted to identify the current state of treatment integrity measurement specifically for programs delivered in early classroom contexts. This literature review was conducted in March of 2018 and aimed to update an abbreviated review conducted by Sutherland et al. (2013). The following search strategy was used to identify randomized controlled trials to determine whether they reported treatment integrity information: randomized AND trial AND (prekindergarten OR preschool) AND (social skills OR emotion regulation OR behavior). A total of 218 studies were identified in ERIC. Studies were included if they met the following criteria: (a) target children aged 3 to 4 at risk for emotional-behavioral disorders, (b) teachers delivered instructional practices that targeted child problem behaviors and/or pre-academic outcomes, and (c) children randomly assigned to condition. Studies were excluded if they targeted a specific population (e.g., preschool children with ADHD, preschool children with autism spectrum disorder).

When these criteria were applied, a total of 16 studies were identified evaluating 10 evidence-based programs. These programs included: *BEST in CLASS* (Conroy et al., 2018; Sutherland et al., 2018), *Chicago School Readiness Project* (CSRP; Raver et al., 2008, 2009), *Incredible Years* (Webster-Stratton et al., 2001, 2008, 2004), *Head Start Research-Based, Developmentally Informed* (REDI; Bierman et al., 2008, Nix et al., 2013), *Preschool PATHS* (Domitrovich et al., 2007, Hamre et al., 2012), *Preschool First Step to Success* (Feil et al., 2014), *Prevent-Teach-Reinforce* (Dunlap et al., 2018), *Pyramid Model* (Hemmeter et al., 2016), *Reaching Educators, Children, and Parents* (RECAP; Han et al., 2005), and *Tools of the Mind* (Barnett et al., 2008). Table 1 provides information pertaining to whether these studies reported treatment integrity data and, if so, information about what was reported. As seen, 10 out of 16 studies reported treatment integrity (62.5%). Since 2013, this indicates a rise in treatment integrity reporting (37.5%; Sutherland et al., 2013), however as the rest of this review will address, there are other factors that need to be taken into consideration when assessing treatment integrity. For example, treatment integrity instruments must be designed with certain applications in mind (Cross & West, 2011), and evidence needs to exist to support the use of these instruments for these applications (e.g., evidence of score reliability and validity; Sanetti & Kratochwill, 2009). With this in mind, out of all of these sixteen studies, two of these studies utilized a treatment integrity instrument, the BiCACS (Sutherland et al., 2014), which is an instrument with preliminary score reliability and validity that also aligns with the design recommendations provided in this review.

Treatment Integrity Components in Educational Research

When developing treatment integrity instruments for the interpretation of study findings, it is best to assess more than one component of treatment integrity (McLeod et al., 2013). That is, a one-dimensional measurement of treatment integrity does not allow for the assessment of other components of treatment integrity that may be useful for program evaluation. In contrast, a multicomponent understanding of treatment integrity includes several unique components that either together or separately contribute to the degree to which the evidence-based program was delivered as designed (Gresham, 2014).

To support this view, Sanetti and Kratochwill (2009) stated that commonly utilized treatment integrity components should address at least one or more of the following areas: (1)

content–the specific features of the intervention that were delivered, (2) quantity–how much of the intervention was delivered, (3) quality–how well the intervention was delivered, and/or (4) process–the mechanism through which the intervention was delivered. For treatment integrity instruments designed to be used for program evaluation, it is particularly important that the instrument captures content, quantity, and quality (Cross & West, 2011). The benefit to accessing process factors is they can aid in understanding how the program was received, but they do not provide information about how the program was delivered (Sanetti & Kratochwill, 2009).

In line with Sanetti and Kratochwill's (2009) review, Sutherland et al. (2013) proposed a multi-component definition of treatment integrity that includes four specific components that map onto the aforementioned areas of content, quantity, and quality. These four treatment integrity components include adherence (content and quantity), differentiation (content and quantity), competence (content and quality), and relational factors (process). Treatment adherence refers to the extent to which a teacher delivers the prescribed program's components as designed (Sutherland et al., 2013). In contrast, differentiation is the level with which a teacher deviates from the prescribed program by delivering practice elements found in proscribed programs (Perepletchikova, 2011). Whereas competence describes the level of skill and degree of responsiveness with which the teacher delivers the program's components (Schulte et al., 2009; Sutherland et al., 2013). Lastly, whereas the other components focus on treatment delivery, "relational factors" are aspects of treatment receipt (i.e., how the evidence-based program is received by a target child; Sutherland et al., 2013). This last component can include quality of the teacher-child relationship, child's response to the teacher's delivery of the program (i.e., responsiveness), and child participation with the teacher or in classroom activities (i.e., engagement).

When assessing treatment integrity within the context of program evaluation, two treatment integrity components are critical to assess (Cross et al., 2015; Cross & West, 2011). It is important to assess both adherence and competence in order to obtain information about the quantity and quality of the delivered intervention. Through assessing adherence, researchers obtain information about how much of an intervention is delivered, and an evaluation of competence provides information about the quality of the delivered intervention. This information is helpful to obtain during randomized-controlled trials to aid researchers in the interpretation of their study findings. When promising results are achieved in a randomized-controlled trial, researchers are better able to assert that the program was most likely the cause of the results when the program was delivered to a great extent and with adequate quality (Schoenwald et al., 2011). However, when unfavorable results are achieved, researchers are better able to attribute the cause of the results to the program itself or delivery of the program after determining if the program was delivered as it was intended (Schoenwald et al., 2011). Adherence is specifically important to assess in the conduct of evaluation studies since researchers can report out on the amount of program delivery needed to achieve promising results. Similarly, competence is important to assess for the purpose of program evaluation in order to determine the level of skillfulness and responsiveness in which the program components are delivered across teachers and across the intervention. However, out of 10 randomized-controlled trials conducted in early classroom contexts that assess a domain of treatment integrity, only 50% of studies assess both adherence and competence.

Design of Treatment Integrity Instruments

Guidelines have not been fully created for the measurement and design of treatment integrity instruments in the early educational field (Sutherland et al., 2013; Sanetti et al., 2020). To adequately capture treatment integrity within the context of the instrument's purpose, it is important that the design of the instrument matches the intended use of the instrument, such as program evaluation (Cross & West, 2011). Thus, this section provides suggestions for the development and measurement of treatment integrity to support program evaluation. Before these guidelines are provided, it is important to understand current trends and variations in treatment integrity measurement.

In the early childhood field, treatment integrity instruments used in randomized-controlled trials vary in design, including response format and informant. For example, out of 10 studies that assessed treatment integrity for evidence-based programs delivered in early classroom settings, five used a Likert-type scale, four used a checklist, and one response format was unknown (see Table 1). Further, observer report was most common such that all 10 studies had an observer rate the teacher's treatment integrity (e.g., Feil et al., 2014; Sutherland et al., 2018) and one of these studies also collected self-report data (i.e., Bierman et al., 2008). Given the review of the early education literature, it is clear that there is variation across response format and that observational instruments are the most frequently used format for treatment integrity instruments.

Self-report and observational instruments have been used to assess treatment integrity for teacher-delivered intervention programs for young children. Self-report instruments (e.g., Bierman et al., 2008) are instruments in which a teacher reports on the extent to which they delivered the intervention as designed. Although these instruments are cost-effective and time-efficient, they may suffer from bias given that the reporter may not be able to reflect objectively about their own behavior (Perepletchikova & Kazdin, 2005). Observational instruments of treatment integrity, considered the gold-standard, provide an objective and specific assessment of a teacher's delivery of an evidence-based program (Hintze & Matthews, 2004). It is important for treatment integrity assessment to be as accurate as possible for program evaluation as biased assessment can hinder the ability to adequately interpret study findings in randomized-controlled trials (Cross & West, 2011). Lastly, in conjunction with the reasons already provided, observational instruments are recommended for the purposes of program evaluation because they have shown more evidence of score reliability and validity than self-report instruments (Hansen et al., 1991; Harachi et al., 1999; Perepletchikova, 2007).

Lastly, the scoring strategy needs to be determined that is in line with the goal of the instrument being designed. Many treatment integrity instruments used to assess teacher delivery of intervention components use a checklist or frequency count format (e.g., Barnett et al., 2008; Bierman et al., 2008; Hamre et al., 2012). Although resource-efficient and fast to complete, these scoring strategies are limited because they do not allow for researchers to assess variation in delivery (Wolery, 2011). In addition, this lack of breadth and depth in assessment is limiting because it does not afford researchers the ability to track changes across training and coaching over time for teachers. To combat this issue, response formats that include a Likert-type scale are ideal such that they may capture change over time (e.g., Bierman et al., 2008; Feil et al., 2014).

Psychometric Properties of Treatment Integrity Instruments

Treatment integrity instruments assessing adherence and competence that demonstrate evidence of score reliability and validity are lacking in early childhood settings (Sutherland et al., 2013). Only four treatment integrity instruments used in randomized-controlled trials within the early childhood literature have documented evidence supporting score reliability (Barnett et al., 2008; Feil et al., 2014; Hemmeter et al., 2016; Sutherland et al., 2014) and two have documented evidence of score validity (Hemmeter et al., 2016; Sutherland et al., 2014). Further, only two of these instruments are observer report and use a Likert-type scale instead of a checklist response format (Feil et al., 2014; Sutherland et al., 2014). Due to the underreporting of psychometric properties of treatment integrity instruments, there is a clear need in the early education field for investigations of the score reliability and validity of treatment integrity instruments used in randomized-controlled trials.

Given the importance of adherence and competence for program evaluation, providing evidence of score reliability and validity for treatment integrity instruments is critical (Sanetti & Kratochwill, 2009). That is, treatment integrity instruments cannot be properly utilized unless they provide evidence of having score reliability and validity at the item and scale level for specific applications. Researchers have determined suggested properties of score reliability and validity of which treatment integrity instruments should possess (Gresham, 2009; Mowbray et al., 2003). However, based on the purpose or function the instrument was developed, these properties vary (Sheridan et al., 2009).

First, treatment integrity instruments must be reliable, meaning that scores produced by independent raters should be consistent (Kazdin, 2016). For observational treatment integrity instruments used for teacher-delivered programs, it is critical that they possess evidence of interrater reliability. Interrater reliability is important to assess because ideally, all raters should be consistent with each other, as in, it should not matter who observes and assesses treatment integrity because all coders should agree with each other.

Whereas reliability is concerned with whether scores are consistent, validity is concerned with whether the scores on the instrument support specific interpretations aligned with applications of the instrument (DeVellis, 2017). Foster and Cone (1995) differentiate between two types of validity: representative and elaborative. Representative validity references whether scores represent the theoretical domain of interest, in this case, adherence and competence. Elaborative validity refers to whether the scores have utility in assessing adherence and competence.

Construct validity falls within the representative validity distinction that Foster and Cone (1995) delineated. Evidence of construct validity is provided when scores indicate that an instrument accurately assesses the construct that it was designed to assess (i.e., adherence and competence) instead of an entirely different construct (Hill & Lambert, 2004). Traditionally, construct validity has been assessed by providing evidence of convergent validity and discriminant validity. Evidence of convergent validity occurs when scores on two instruments that assess the same construct (i.e., two instruments designed to measure adherence) are strongly correlated with each other (Foster & Cone, 1995). Conversely, scores on two instruments that assess unrelated constructs (i.e., adherence and treatment receipt) should demonstrate small correlations with each other (i.e., discriminant validity; DeVellis, 2017). Another way to assess construct validity is to evaluate the correlational patterns between instruments assessing various constructs. It would be expected that scores produced by items or scales of a similar construct should evidence a stronger correlation with each other than correlations between items or scales of two unrelated constructs (Furr & Bacharach, 2014). For instance, Hogue et al. (2008) found that competence ratings were more correlated with each other than when comparing correlations between competence and alliance.

It is important for program evaluation for treatment integrity instruments to demonstrate construct validity to ensure that the instruments evaluate two key integrity components: adherence and competence (Cross et al., 2015; Cross & West, 2011). Without evidence supporting construct validity, it is possible that the instrument may not assess what it was designed to assess (Hill & Lambert, 2004). If this is the case, then scores on the instrument could not be interpreted as representing the extent to which the program was delivered (i.e., adherence) or the quality of delivery (i.e., competence) in randomized-controlled trials. Instead, the scores of the instrument may be assessing a different construct or component of treatment integrity (e.g., treatment receipt, alliance).

Discriminative validity and measurement sensitivity can provide evidence of elaborative validity. Discriminative validity is established when an instrument is able to differentiate between two groups that are expected to differ (Foster & Cone, 1995). For example, if an instrument is able to distinguish between a group of teachers that were trained and coached to deliver an evidence-based program and a group that has not learned to deliver the program, then the instrument would possess evidence of discriminative validity (Carroll et al., 2000). This is critical for program evaluation since it affords researchers the ability to determine whether adherence and competence scores differed between groups as expected (e.g., business as usual condition and program condition). Further, measurement sensitivity provides evidence that scores on an instrument can detect changes in program delivery over time (Kazdin, 2016). This is critical for program evaluation because it allows researchers to determine whether adherence and competence scores change over the course of training and coaching (e.g., Did teachers deliver more or less of the evidence-based program over time?; id teachers improve their quality in the delivery of the evidence-based program over time?).

BEST in CLASS Adherence and Competence Scale (BiCACS)

The BEST in CLASS Adherence and Competence Scale (BiCACS; Sutherland et al., 2014) is an integrity instrument that assesses the treatment integrity domains of adherence and competence of the core practices of the BEST in CLASS program. The assessment instrument also evaluates two additional items that assess how a child responds to a teacher's delivery of the program: *Child Engagement* and *Child Responsiveness*. The BEST in CLASS program is a Tier-2 teacher-delivered program that is delivered within early classroom contexts. Similar to recommendations presented in this review, the BiCACS is an observational integrity instrument that utilizes a 7-point Likert-type scale to assess these integrity domains.

The BiCACS has been used to determine the level of adherence and competence of BEST in CLASS from pretest to posttest and whether there are group differences between program conditions in their levels of adherence and competence. Results from research studies have shown that adherence scores for teachers who were trained to deliver BEST in CLASS significantly increased both for adherence and competence scores from pretest to posttest (Sutherland et al., 2018). Results have also demonstrated that teachers in the BEST in CLASS condition delivered the intervention with significantly higher adherence and competence compared to a business-as-usual comparison condition (Conroy et al., 2015). These results speak to the purpose of the BiCACS as an instrument for program evaluation, given that each of these results suggests that teachers in the BEST in CLASS condition are delivering the program more frequently, thoroughly, skillfully, and more responsive than teachers in a comparison condition. Furthermore, the BiCACS has demonstrated initial evidence of reliability and validity to support its use as a program evaluation instrument. Nonetheless, to help support this use of this instrument, the psychometric properties of the BiCACS must be understood and investigated further.

Preliminary Psychometric Evidence

The BiCACS has demonstrated initial evidence of interrater reliability at the item and subscale (i.e., Adherence and Competence subscales) levels, construct validity (i.e., convergent and discriminant validity) at the item and subscale level, discriminative validity at the subscale level, and measurement sensitivity at the subscale level. In particular, a preliminary psychometric investigation of the BiCACS has been conducted using two years of data collected from teachers who participated in the BEST in CLASS development study (*N* = 11; Sutherland et al., 2014). Within this study, a total of 289 observations were conducted of teachers delivering the BEST in CLASS program. Observations were assessed across various phases of the program (i.e., baseline, treatment implementation, posttest, and maintenance). Interrater reliability was evaluated with intraclass correlation coefficients (ICC[2,1]) between two coders and ranged from "fair" to "excellent" for the Adherence items (*M* = .72; *SD* = .15; range, .44 – .91; Sutherland et al., 2014). For the Competence items, ICCs ranged from "poor" to "excellent" with only one item in the "poor" category (*M* = .64; *SD* = .16; range, .39 – .85; Sutherland et al., 2014). The Adherence subscale (ICC = .90) and the Competence subscale (ICC = .85) displayed "good" reliability (Sutherland et al., 2014). These preliminary findings provide initial evidence of interrater reliability of the BiCACS at the item and subscale levels. To enhance the understanding of the interrater reliability as it pertains to the BiCACS, interrater reliability should be assessed using the full double-coded data from a randomized-controlled trial.

Within the context of the preliminary psychometric study (i.e., Sutherland et al., 2014), which included data from the BEST in CLASS development study, the score validity of the

BiCACS was assessed indirectly through a few methods. First, to support the construct validity of the instrument, inter-item and intra-item correlations among the BiCACS items were conducted and included all eight available time points of data collection collectively in analyses. Results indicated that intra-item correlations among the Adherence items, intra-item correlations among the Competence items, and inter-item correlations among the adherence and Competence items were all moderate to strong in strength (Sutherland et al., 2014). Also, in the preliminary psychometric study, correlations among the subscales were also conducted, and findings demonstrated a strong correlation among the Adherence and Competence subscale (*r* = .71; Sutherland et al., 2014). Though, the magnitude of the correlation among the BiCACS subscales suggest that they are redundant (i.e., *r* > .70; Kline, 1979) and may not assess different constructs.

Construct and discriminant validity were further assessed by correlating scores on the Adherence and Competence subscales with scores on two subscales of the Student-Teacher Relationship Scales – Short Form (STRS-SF; Pianta, 2001): Closeness and Conflict. Findings indicated that the Adherence subscale had a strong, positive correlation with the Closeness subscale (*r* = .51, *p* = .026; Sutherland et al., 2014) and the Competence subscale demonstrated a moderate, positive correlation with the Closeness subscale (*r* = .43, *p* = .065; Sutherland et al., 2014). Additionally, the Adherence subscale demonstrated a small, negative correlation with the Conflict subscale (*r* = -.13, *p* = .578; Sutherland et al., 2014), and the Competence subscale demonstrated a similar correlation (*r* = -.18, *p* = .474; Sutherland et al., 2014). Although the magnitude of the correlations is stronger than anticipated, in comparison with the Adherence and Competence subscales, the patterns of the correlations are in line with expectations and support the construct validity of the BiCACS (Furr & Bacharach, 2014). That is, the Adherence and

Competence subscale demonstrated the strongest correlation, followed by the Adherence or Competence subscale with the Closeness subscale, and then the Adherence or Competence subscale with the Conflict subscale.

To investigate the measurement sensitivity of the BiCACS, Sutherland et al. (2014) assessed whether Adherence and Competence subscale scores were distinguishable between different phases of the BEST in CLASS program. In the context of this study, if the subscale scores increased over time, potentially due to the mastery of new intervention components, then the subscale scores would be able to distinguish between these phases (Sutherland et al., 2014). Results demonstrated that the Adherence subscale scores significantly distinguished between four phases of the intervention, but the Competence subscale scores did not have the same level of sensitivity (Sutherland et al., 2014). Overall, the current preliminary psychometric properties of the BiCACS suggest that the instrument demonstrates evidence of interrater reliability, construct validity, discriminative validity, and measurement sensitivity within the context of a program development study. However, to support the purpose of the BiCACS as a program evaluation instrument, the psychometric properties of the instrument need to be investigated further in the context of a rigorous randomized-controlled trial which would include more observations across time and more than one study condition.

Current Study

The current study aimed to investigate the score reliability and validity of the BiCACS to support the instrument's use as a program evaluation instrument for a Tier-2, teacher-delivered evidence-based program (i.e., BEST in CLASS). BiCACS data will be collected by observational coders from teachers participating in two conditions (i.e., BEST in CLASS and business as usual). Teachers randomly assigned to BEST in CLASS were taught how to deliver instructional practices during classroom instructions aimed at children at risk for emotional and behavioral disorders. Teachers were taught these strategies by receiving a manual, attending a one-day didactic training, and meeting with a coach weekly (Conroy et al., 2018; Sutherland et al., 2018). Business as usual condition teachers were not exposed to the behavioral management strategies and served as a comparison condition. More information about these conditions and the coders are provided later.

Study Aims

This study aimed to assess whether the BiCACS possesses evidence of reliability (i.e., interrater reliability), representative validity (i.e., convergent and discriminant validity), and elaborative validity (i.e., discriminative validity and measurement sensitivity). To do so, the study answered the following research questions: (a) can coders reliably code the BiCACS at the item and subscale level?, (b) do the adherence and Competence items on the BiCACS assess what they purport to assess?, (c) can the BiCACS differentiate between intervention conditions (i.e., BEST in CLASS vs. business as usual)?, and (d) are adherence and competence scores on the BiCACS sensitive to changes over time?

Hypotheses

To determine whether the BiCACS demonstrates evidence of score reliability and validity for the purposes of evaluating the BEST in CLASS program, seven hypotheses will be tested.

Hypothesis 1: Interrater Reliability at the Item Level. The BiCACS will demonstrate evidence of interrater reliability at the item level. Using intraclass correlation (ICC), it is hypothesized that BiCACS item scores will demonstrate at least “good” (ICC(2,1) > .60;

Cicchetti, 1994) levels of interrater reliability for each Adherence and Competence item (Sutherland et al., 2014;).

Hypothesis 2: Interrater Reliability at the Subscale Level. The BiCACS will demonstrate evidence of interrater reliability at the subscale level. Using intraclass correlation (ICC), it is hypothesized that subscales scores will demonstrate at least "good" (ICC(2,1) > .60; Cicchetti, 1994) levels of interrater reliability on the Adherence and Competence subscales (Sutherland et al., 2014;).

Hypothesis 3: Construct Validity at Item Level. It is hypothesized that for scores produced by coders: (a) inter-item correlations between the BiCACS Adherence items and the corresponding BiCACS Competence items (e.g., BiCACS Adherence Precorrection and BiCACS Competence Precorrection) will be medium to large in strength (r > .24; Hogue et al., 2008; Sutherland et al., 2014); (b) intra-item correlations among BiCACS Adherence items (e.g., BiCACS Adherence Precorrection and BiCACS Adherence Corrective Feedback) and BiCACS Competence items (e.g., BiCACS Competence Precorrection and BiCACS Competence Corrective Feedback) will be moderately correlated (r = .24 – .36; Hogue et al., 2008; Sutherland et al., 2014); (c) inter-item correlations between non-corresponding BiCACS Adherence and Competence items (e.g., BiCACS Adherence Precorrection and BiCACS Competence Corrective Feedback) will be small to medium in strength (r = .10 – .36; Hogue et al., 2008; Sutherland et al., 2014); and (d) the BICACS Adherence and Competence items will demonstrate small to medium correlations with the BiCACS Engagement and Responsiveness items (r = .10 – .36; Hogue et al., 2008; Sutherland et al., 2014).

Hypothesis 4: Construct Validity at the Subscale Level. It is hypothesized that inter-subscale correlations between (a) the BiCACS Adherence and Competence subscales will be medium to large in strength with each other (r > .24; Carroll et al., 2000; Sutherland et al., 2014); (b) the BICACS Adherence and Competence subscales will demonstrate small to medium correlations with the BiCACS Engagement item (r = .10 – .36; Sutherland et al., 2014); (c) the BICACS Adherence and Competence subscales will demonstrate small to medium correlations with the BiCACS Responsiveness item (r = .10 – .36; Sutherland et al., 2014).

Hypothesis 5: Discriminant Validity. It is hypothesized that BiCACS scores will demonstrate small correlations with scores on the STRS, a teacher self-report instrument. More specifically, it is hypothesized that: (a) the BICACS Adherence and Competence subscale scores will demonstrate a small, positive correlation with scores on the STRS Closeness subscale (r = .10 – .23; Hogue et al., 2008; Sutherland et al., 2014); (b) scores on the BICACS Adherence and Competence subscales will demonstrate a small, negative correlation with scores on the STRS Conflict subscale (r = .10 – .23; Hogue et al., 2008; Sutherland et al., 2014).

Hypothesis 6: Discriminative Validity. The BiCACS Adherence and Competence subscale scores will be sensitive to differences in adherence and competence by demonstrating significant group differences on the BiCACS Adherence and Competence subscales between the BEST in CLASS and business as usual conditions consistent with expected differences between groups (i.e., BEST in CLASS teachers will demonstrate higher scores; Hemmeter et al., 2016).

Hypothesis 7: Measurement Sensitivity. To assess the measurement sensitivity of the BiCACS, it is hypothesized that scores on the BiCACS Adherence and Competence subscales will demonstrate positive change over the course of training and coaching for BEST in CLASS teachers (Sutherland et al., 2014).

Methods

Data Source

Data for the current study were drawn from a larger four-year multi-site cluster randomized efficacy trial hereby referred to as “parent study” (Conroy et al., 2018; Sutherland et al., 2018). The parent study compared the efficacy of the BEST in CLASS program, a Tier-2 intervention delivered by teachers aimed at reducing young children’s problem behaviors in the classroom and increasing positive teacher-child relationships, to a comparison condition (i.e., business-as-usual). Findings of the parent study indicated that children in the BEST in CLASS program had a significant reduction in problematic behavior, increased their social skills, and showed higher classroom engagement compared to business-as-usual from pretest to posttest (Conroy et al., 2018; Sutherland et al., 2018). Teacher-child relationships also improved from pretest to posttest such that observers rated an increase in positive teacher-child interactions and a decrease in negative teacher-child interactions (Sutherland et al., 2018). Additionally, results demonstrated that teachers in BEST in CLASS achieved higher levels of teacher-reported closeness with the child and lower levels of teacher-reported conflict (Sutherland et al., 2018). Lastly, observer-rated adherence and competence of teacher delivery of BEST in CLASS was significantly higher at posttest than pretest for teachers in the BEST in CLASS condition (Sutherland et al., 2018).

Participants

Teachers

A total of 185 teachers participated in the parent study (*n* = 92 in the BEST in CLASS condition; *n* = 93 in the business-as-usual condition). Teachers were eligible to participate in the study if they: (a) taught children aged three to five in an early childhood classroom, (b) had children in their classroom that were eligible to participate, (c) had not previously participated in a BEST in CLASS study, and (d) consented to participate. Teachers from the parent study were included in the current study if BEST in CLASS Adherence and Competence Scale (BiCACS) data were available on the teacher's delivery of BEST in CLASS practices for at least one timepoint and for at least one focal child (i.e., children selected to participate in the parent study who have been identified at risk for emotional and behavioral disorders). This resulted in a sample of 179 teachers (*n* = 89 in the BEST in CLASS condition; *n* = 90 in the business-as-usual condition). Demographic data for the current sample is presented in Table 2.

Children

A total of 465 children at risk for emotional and behavioral disorders participated in the parent study (*n* = 231 in the BEST in CLASS condition; *n* = 234 in the business-as-usual condition). Children were eligible for participation in the parent study if they met the following screening criteria and eligibility requirements: (a) were between the ages of three and five, (b) demonstrated fluency in the English language, (c) their cognitive development fell within the normative range, (d) were at risk of developing an emotional or behavioral disorder, (e) if their caregivers consented to have them participate in the study. Child participants were included in the current study if BiCACS data were available. This resulted in a sample of 416 child participants (*n* = 211 in the BEST in CLASS condition; *n* = 205 in the business-as-usual condition). Demographic data for the current sample is presented in Table 3.

Parent Study Procedures

Recruitment

Participants of the parent study were recruited from early childhood programs located in urban, suburban, and rural communities near Richmond, Virginia, and Gainesville, Florida. The early childhood programs were either located within elementary schools of local school districts or within early childhood education centers. The programs participating in the parent study were mainly federally- or state-funded programs (96%); the rest were privately funded programs (4%).

Child Screening Process

The following screening procedure was conducted to screen in 1-3 target children at risk for emotional and behavioral disorders. First, teachers nominated five children that demonstrated elevated levels of problem behaviors in the classroom. Consent forms were sent home to the nominated children's caregivers if the children were English speakers or demonstrated proficiency in English using the BEST in CLASS English language screener. Children with returned and signed consent forms progressed through the screening process. The nominated children's cognitive developmental level was screened using the *Battelle Developmental Inventory, Second Edition Screener* (*BDI-II Screener*; Newborg, 2005), and risk for an emotional or behavioral disorder was determined by using the *Early Screening Project* (*ESP*; Feil et al., 1998). Children identified as falling within the normal range for cognitive development using the *BDI-II* and at an increased risk for an emotional or behavioral disorder as indicated by the *ESP* were eligible to participate. Based on the amount of returned consent forms and severity of risk for the emotional and behavioral disorder (i.e., scores exceeding one standard deviation from gendered norms on the *ESP*), one to three children with the highest level of emotional and behavioral risk (i.e., the highest score on *ESP*) were selected to participate in the parent study per teacher.

Randomization

Upon consent, teachers were randomly selected to participate in one of the intervention conditions (i.e., BEST in CLASS or business-as-usual) within their recruitment site (i.e., the location of which they were employed). At sites in which more than one teacher consented to the study, teachers were randomly assigned to intervention conditions from within their site, but steps were taken to obtain an equal distribution of teachers across both conditions. For example, at sites with an even number of participating teachers, half of the teachers were randomly assigned to the BEST in CLASS condition, and the other half was then assigned to the business-as-usual condition. At sites with an odd number of participating teachers, using simple random assignment, the extra teacher that was not assigned to a condition was randomly assigned into one of the intervention conditions. At sites in which only one teacher consented, teachers were assigned to the intervention by simple random assignment. Since teachers were randomly selected to intervention condition, the 1-3 focal children that were in their classroom were then assigned to the same condition.

Intervention Conditions

BEST in CLASS. In the BEST in CLASS condition, the teachers were trained to use BEST in CLASS practices geared towards their focal children and received practice-based coaching. Teachers received the program manual and learned the BEST in CLASS practices during a one-day didactic training session that entailed a presentation of BEST in CLASS, modeling of the practices, and opportunities to practice with other teachers. Teachers learned the BEST in CLASS practices through six modules that included rules, precorrection, opportunities to respond, behavior-specific praise, corrective feedback, and instructive feedback. Module information is presented in Table 4. Teachers were also taught via a seventh module, linking and mastery, how to link these six practices together to use them in tandem with each other.

After the training concluded, teachers began receiving practice-based coaching. *Practice-based coaching* is defined as a cyclical process that supports teachers' use of effective evidence-based practices that produce positive child outcomes (Synder et al., 2015). Practice-based coaching typically involves three steps, including (a) assessment of classroom needs, (b) planning and implementation of evidence-based practices in the classroom, and (c) evaluation of the delivery of the evidence-based practices (Synder et al., 2015). Within the context of the BEST in CLASS program, Sutherland et al. (2015) detail six components of practice-based coaching, including (a) facilitated instructed of new skills; (b) shared goals and action planning; (c) guided practice; (d) reflection; (e) direct, focused observation; and (f) reflection and feedback.

During each coaching session, coaches addressed teachers' questions about BEST in CLASS, modeled BEST in CLASS components, and helped the teachers meet their self-determined goals. Practice-based coaching lasted for 14-weeks. Within this timeframe, teachers were provided two weeks of coaching centered around at least one component of BEST in CLASS before focusing on another component. During the first week, coaches and teachers formed an action plan that determined how to deliver a BEST in CLASS component with the target child or children. Then the teacher implemented the plan while the coach observed and provided performance-based feedback. The following week, the coach and teacher met to reflect on the implementation of the BEST in CLASS component. Upon conclusion of the second week, the teacher was considered to have mastered the BEST in CLASS practice, and the subsequent coaching session focused on the next BEST in CLASS practice. The mastery of these practices occurred at the following time points of which time point 1 is a pretest measurement prior to training and coaching: time 2 = rules mastery; time 3 = precorrection mastery; time 4 = opportunities to respond mastery; time 5 = praise mastery; and time 6 = corrective feedback and instructive feedback mastery.

Business-as-usual. Participants in the business-as-usual condition were not exposed to BEST in CLASS components. Business-as-usual is a comparison condition in which teachers and children participate in instructional activities typically offered in early childhood classrooms. Across both conditions, teachers reported the use of both specific behavioral strategies (e.g., token economy, tangible reinforcement, time-out) and manualized early childhood curricula to guide their daily instruction (Sutherland et al., 2018). Most commonly reported manualized early childhood curricula that were implemented in the classrooms included Teaching Strategies Gold (Heroman et al., 2010a), the Creative Curriculum (Heroman et al., 2010b), and High Scope (High Scope Educational Research Foundation, 2014).

Data Collection

At pretest, teachers in both conditions were given a packet of self-report forms to complete, including demographics, a measure of child problem behavior, and a measure of the student-teacher relationship. Children were sent home with demographic forms to be completed by their caregivers. Posttest procedures closely resembled pretest procedures, except demographics were not collected. Participants were compensated for their participation.

Current Study Procedures

BEST in CLASS Adherence and Competence Scale

Overview. The BEST in CLASS Adherence and Competence Scale (BiCACS; Sutherland et al., 2014) is a 14-item direct observational treatment integrity instrument used to assess adherence and competence of the delivery of BEST in CLASS practices in early classroom settings. Coders are asked to rate the teacher’s adherence and competence on six items that are directly related to the BEST in CLASS practices the teachers received training and coaching on. The BiCACS includes a total of 14 items (six Adherence items, six Competence items, and two additional items). A list of the BiCACS items and their definitions can be found in Table 5. The following six items are included on both the Adherence subscale and the Competence subscale: (a) *Teacher reviews rules and addresses rule violations*, (b) *Teacher provides precorrection*, (c) *Teacher provides opportunities to respond*, (d) *Teacher provides behavior-specific praise*, (e) *Teacher provides corrective feedback*, and (f) *Teacher provides instructive feedback*. The two additional child behavior items include: (a) *Child Responsiveness to teacher behavior* (i.e., Child Responsiveness), and (b) *Child Engagement*. Whereas the Adherence and Competence items assess how the teacher delivers the BEST in CLASS program, the Child Responsiveness and Child Engagement items assess how the focal child receives or responds to the teacher delivery of the intervention.

Lastly, it is important to note that some of the BiCACS items were added or changed during the four years of data collection. For instance, the Child Responsiveness and Child Engagement items were added to the BiCACS in year two of the parent study, and therefore only three years of data are available. Furthermore, Opportunities to Respond was accessed via two separate items (i.e., Academic Opportunities to Respond and Social Opportunities to Respond) in year one of data collection. After the first year, these two items were combined to assess "Opportunities to Respond" as defined in Table 5. No further changes were made to the BiCACS during the four years of data collection.

Scoring. The six items that compose the Adherence subscale and the two additional child behavior items are scored using a 7-point extensiveness scale ranging from 1 (*not at all*) to 7 (*extensively*; Hogue, et al., 1996; McLeod & Weisz, 2010; Sutherland et al., 2013). Two dimensions are considered when rating extensiveness: thoroughness and frequency. *Thoroughness* is defined as the intensity or persistence with which the teacher executes a BEST in CLASS practice component. *Frequency* refers to the amount or number of times a specific BEST in CLASS practice is utilized during the observation. By coding adherence using a combination of both thoroughness and frequency, the Adherence subscale provides information regarding the quantity or extent to which the BEST in CLASS practice was delivered by the teacher during the observed session.

Competence items are scored using a 7-point Competence subscale ranging from 1 (very poor) to 7 (excellent; Carroll et al., 2000; Sutherland et al., 2013). Since adherence presupposes competence, competence is only scored if adherence was scored between a 2 and 7. If adherence was scored using "1," then competence is scored using "NO" meaning "no opportunity." Competence ratings using the BiCACS consider the extent to which teachers demonstrate skillfulness and responsiveness when delivering BEST in CLASS. Specifically, coders are expected to consider the following dimensions: (1) expertise, commitment, motivation to promote change in the child (skillfulness); (2) clarity of language and communication when intervening with the child (skillfulness); (3) appropriate timing of actions (responsiveness); and (4) ability to read and respond to where the child appears to be (responsiveness).

Observational Coding Procedures

Coders. The coding team consisted of coaches and data staff. Twenty-six coaches were hired to work on the parent study, and of these 26 coaches, 17 resided in Virginia, and nine resided in Florida. Coach demographics are as follows: 92.3% female, 7.7% male; 73.1% White/European American, 15.4% Black/African American, 11.5% Hispanic/Latinx; 100% attained a Bachelor's degree, 50% were enrolled in a graduate program. Additionally, coaches ranged in age from 25-65 years old. Data staff included individuals who were hired to work on the parent study as raters of treatment integrity and other research tasks (e.g., data collection, data entry, etc.). The majority of data staff attained or were in the process of obtaining their bachelor's degree.

Training. Prior to the start of the parent study, all coders received a copy of the BiCACS coding manual and were asked to read through the manual and memorize the definition for each item. The coders then attended a 2-hour didactic training session where the definitions of the codes and scoring strategy were reviewed. Following, coders practiced scoring videotaped sessions in order to gain practice using the BiCACS. This coding occurred over a two-month period. Coders were trained over this time period to reach adequate pre-study reliability across all BiCACS items on the videotaped sessions (ICC > .59; Cicchetti, 1994). When the training was completed, coders began coding live observations of teachers providing instruction in their classrooms.

Coding Procedures. Live-observational coding occurred in the participating teachers' classrooms during instructional time for a minimum of 10 minutes; if coders could not observe for 10 minutes, then they did not complete the BiCACS. Instructional time consists of any teacher- or child-directed activity that involves instructional opportunities for the focal child to engage in instructional opportunities (e.g., circle time, center time, one-on-one instruction, free-play, small group). Observations focused on the teacher's behavior directed to the focal child or a group in which the focal child is a part of and the focal child's reaction to the teacher's behavior. In classrooms with more than one participating focal child, coders were asked to conduct the observation on one focal child at a time.

Data Collection. Data collection for the BiCACS occurred across 18 weeks at eight time points. These data points include pretest (prior to the start of the intervention), posttest (the last week of the coaching session), five "intervention" timepoints (at week 4, week 6, week 8, week

10, and week 13, and maintenance (3-5 weeks after posttest). The intervention time-points correspond to the week that the teacher was considered to "master" the BEST in CLASS practice they were receiving coaching on.

Reliability Sample. To assess the interrater reliability of the instrument, a second observer completed the BiCACS for a target number of observations in the parent study. A secondary coder completed the BiCACS in classrooms at pretest (target of 20% of classrooms), posttest (target of 30% of classrooms), and maintenance (target of 20% of classrooms). A secondary observer also completed the BiCACS for a target of 20% of classrooms at each intervening time-point (i.e., week 4, week 6, week 8, week 10, and week 13).

Other Measures

Instruments Used in Validity Analyses

The Student-Teacher Relationship Scale-Short Form (STRS-SF). The STRS-SF (Pianta, 2001) is a teacher-report measure of teacher perceptions of their relationships with children in their classroom. Teachers used this instrument to assess their relationship with each child participating in the study at pretest and posttest. The STRS-SF consists of 15 items that are scored on a 5-point Likert scale with the following anchors: 1 (*definitely does not apply*) to 5 (*definitely applies*). Teachers are asked to reflect on the degree to which each of the 15 statements apply to the relationship. Example items include: (a) "I share an affectionate, warm relationship with this child"; (b) "this child and I always seem to be struggling with each other"; (c) "this child easily becomes angry with me"; and (d) "this child openly shares his/her feelings and experiences with me". Two subscales are derived from the STRS-SF: Closeness and Conflict. The full-scale Student-Teacher Relationship Scale (Pianta & Hamre, 2001), of which this instrument is derived from, has demonstrated high internal consistency for both the

Closeness (alpha = .86) and Conflict (alpha = .92) subscales, discriminant validity, concurrent validity, and predictive validity (Hamre & Pianta, 2001; Pianta, 2001). Pretest STRS-SF scores were used in the validity analyses of the current study. Internal consistency of the STRS-SF at pretest for the current study was acceptable for both scales, Closeness (alpha = .72) and Conflict (alpha = .86).

Instruments Used to Describe Sample

Social Skills Improvement System- Rating Scale (SSIS-RS). The SSIS-RS (Gresham & Elliott, 2008) is a teacher-report instrument that assesses children's social skills and problem behaviors. Teachers completed one to three SSIS-RS measures depending on how many children were participating in the study who was a student in their classroom. The SSIS-RS consists of 76 items which are rated using a 4-point frequency scale ranging from 0 (*never*) to 3 (*almost always*). The instrument can be broken into two subscales: Social Skills and Problem Behaviors. Scores on the total scale and subscales are summed and then converted into a standard score. The items on the Problem Behavior and Social Skills subscales at pretest will be used to determine whether children across conditions displayed similar baseline levels of problem behavior and social skills. The SSIS-RS demonstrates evidence of high internal consistency for the Problem Behavior subscale (alpha = .95) and Social Skills scale (alpha =.97; Gresham et al., 2011). Further, it demonstrates evidence of construct validity such that it converges with the SSIS long version (Gresham et al., 2011). Internal consistency of the SSIS Problem Behavior (alpha = .91) and Social Skills (alpha = .94) scales of the current study were acceptable.

Analytic Plan

The purpose of the current study was to investigate the score reliability and validity of the BEST in CLASS Adherence and Competence Scale (BiCACS). To achieve study goals, score reliability, construct validity was examined at the item level. At the subscale level, reliability, construct validity, discriminative validity, and measurement sensitivity were examined. BiCACS item scores for Adherence and Competence were produced by coders observing teachers in two conditions (i.e., BEST in CLASS and business-as-usual). These item scores were then averaged together to produce an Adherence and Competence subscale. Seven timepoints collected across 15 weeks were used for the current study analyses. The maintenance timepoint was dropped from analyses to eliminate possible effects of the removal of intervention supports on the analyses; the investigation of psychometric analyses across the remaining seven timepoints allowed for the examination of the score reliability and validity of the BiCACS while coaching and training were ongoing. As such, the removal of the maintenance timepoint was consistent with study aims.

Preliminary Analyses

Sample Bias

Sample bias analyses were conducted to determine whether the teachers and children in the BEST in CLASS and business-as-usual conditions differed at baseline using chi-square or t-test analyses on child and teacher demographic and pretest variables. Further, sample bias analyses were conducted to determine whether teachers and children included in these study analyses differed in their reported demongraphics from those of the parent study. Lastly, analyses were conducted to determine if children and teachers were included in the reliability sample, a subset of the current sample, differed from those in the current study based on demographics.

Data Screening and Data Distribution

Data were screened for data entry errors. Values were marked as missing if the value could not be remedied. Due to data manual changes between years one and two of data collection, as previously mentioned, a combined Opportunities to Respond item was created for year one data by averaging the Academic Opportunities to Respond and Social Opportunities to Respond scores. Child Responsiveness and the Child Engagement items for year one were marked as missing since these items were not accessed in year one. Descriptive statistics, including means, standard deviations, ranges, frequencies, skewness, and kurtosis were calculated for the BiCACS at the item level to assess for normality. To interpret the distribution of scores, acceptable values for skewness and kurtosis were considered to be values between -2 through +2 (George & Mallory, 2010). Outliers were assessed by computing the raw scores of the BiCACS items into *z*-scores. Outliers were defined as a z-score greater than the absolute value of 3.29 (Tabachnick & Fidell, 2006).

Reliability Analyses

Interrater Reliability: BiCACS Adherence and Competence items

Interrater reliability for each BiCACS item was assessed by calculating intraclass correlation coefficients, ICC (2,1), for every observation that was double-coded**.** The single-rater ICC was used since one observer coded all observations, and only a subset of these observations was coded by a second observer (McLeod et al., 2013). A two-way random-effects model was used as it allows for the generalizability of the results to other similar samples (Koo & Li, 2016; Shrout & Fleiss, 1979). To support the score reliability at the item level, it was expected per Hypothesis 1 that BiCACS Adherence and Competence items would demonstrate at least “good” (ICC(2,1) > .60; Cicchetti, 1994) levels of interrater reliability for the Adherence and

Competence items (Sutherland et al., 2014; Sutherland et al., 2018). To categorize reliability at the item level between coders, guidelines determined by Cicchetti (1994) were used: ICCs less than .40 would be considered "poor" reliability, ICCs between .40 and .59 will be considered "fair" reliability, ICCs between .60 and .74 will be considered "good" reliability, and ICCs greater than .75 will be considered "excellent" reliability.

Interrater Reliability: BiCACS Adherence and Competence subscales

To assess reliability at the subscale level, the BiCACS Adherence and Competence subscales were created using the item level reliability analyses to inform subscale creation. All items on the BiCACS Adherence subscale were averaged together to create the Adherence subscale score using the coder's ratings. For items on the Competence subscale, all scores of "0" were first coded as "missing." Afterward, the remaining items on the Competence subscale were averaged to create the Competence subscale score.

Interrater reliability at the subscale level was then assessed using a similar procedure as the item-level analyses. That is, an ICC(2,1) was computed for the Adherence subscale and the Competence subscale for every observation that was double-coded. It was hypothesized per hypothesis 2 that the Adherence and Competence subscales would demonstrate at least "good" (ICC(2,1) > .60; Cicchetti, 1994) levels of interrater reliability. The same guidelines that were applied to the item-level analyses will be used to categorize reliability at the subscale level.

Validity Analyses

Construct Validity: BiCACS Adherence and Competence items

Construct validity analyses of the BiCACS focused on the magnitude and pattern of correlations among BiCACS Adherence items, the BiCACS Competence items, the Engagement item, and the Responsiveness item. Per hypothesis 3, it was expected that (a) inter-item correlations between the BiCACS Adherence items and the corresponding BiCACS Competence items would be medium to large in strength (r > .24; Carroll et al., 2000; Hogue et al., 2008; Sutherland et al., 2014); (b) intra-item correlations between BiCACS Adherence or Competence items would be moderately correlated (r = .24 – .36; Hogue et al., 2008); (c) inter-item correlations between non-corresponding BiCACS Adherence and Competence items would be small to medium in strength (r = .10 – .36; Hogue et al., 2008); and (d) the BICACS Adherence and Competence items would demonstrate small to medium correlations with the BiCACS Engagement and Responsiveness items (r = .10 – .36; Hogue et al., 2008). To interpret correlations, the following guidelines were used: *r* is small correlation effect if .10 – .23, *r* is medium correlation effect if .24 –.36, *r* is large correlation effect if > .36 (Rosenthal & Rosnow, 1984). It interpret the pattern of the correlations, means of the absolute values of the correlations were computed, and Fisher's r-to-z transformation were used to determine whether these mean correlations significantly differed in magnitude.

Construct Validity: BICACS Adherence and Competence subscales

Construct validity of scores on the BiCACS subscales were assessed by evaluating the magnitude and pattern of correlations among the BiCACS Adherence subscale, BiCACS Competence subscale, Engagement item, and Responsiveness item. Per hypothesis 4, it was hypothesized that: (a) correlations between the BiCACS Adherence and Competence subscales would be medium to large in strength (r > .24; Carroll et al., 2000; Hogue et al., 2008; Sutherland et al., 2014); (b) the BICACS Adherence and Competence subscales would demonstrate small to medium correlations with the BiCACS Engagement item (r = .10 – .36; Hogue et al., 2008); and (c) the BICACS Adherence and Competence subscales would demonstrate small to medium correlations with the BiCACS Responsiveness item (r = .10 – .36;

Hogue et al., 2008). Follow-up contrasts using Fisher's r-to-z transformation were used to determine whether correlations significantly differed in magnitude.

Pearson product-moment correlations using the BiCACS Adherence and Competence subscales, BiCACS Child Engagement and Responsiveness items, and the STRS-SF Closeness and Conflict subscales were assessed to evaluate the construct validity of the BiCACS Adherence and Competence subscales. Although the STRS-SF was collected at pretest and posttest, posttest data were not assessed due to the observed increase in positive teacher-child relationships in BEST in CLASS participants (Sutherland et al., 2018). Eight correlations across the BiCACS subscales or items with the STRS-SF subscales were conducted using pretest data. Per hypothesis 5 it was expected that: (a) the BiCACS Adherence and Competence subscale scores would demonstrate a small, positive correlation with scores on the STRS Closeness subscale (r = .10 – .23; Hogue et al., 2008; Sutherland et al., 2018); (b) scores on the BICACS Adherence and Competence subscales will demonstrate a small, negative correlation with scores on the STRS Conflict subscale (r = .10 – .23; Hogue et al., 2008; Sutherland et al., 2018); (c) item scores of the Engagement and Responsiveness items will demonstrate a small, positive correlation with scores on the STRS Closeness subscale (r = .10 – .23; Hogue et al., 2008; Sutherland et al., 2018); and (d) item scores of the Engagement and Responsiveness items will demonstrate a small, negative correlation with scores on the STRS Conflict subscale (r = .10 – .23; Hogue et al., 2008; Sutherland et al., 2018). Rosenthal and Rosnow's (1984) guidelines were used to interpret the strength of the correlations. Fisher's r-to-z transformation were used to determine whether correlations significantly differed in magnitude from one another.

Discriminative Validity

Discriminative validity was evaluated by comparing between-group differences on the Adherence and Competence subscale scores across the BEST in CLASS and business-as-usual conditions. It was expected that significant group differences on the BiCACS Adherence and Competence subscales would be found between the BEST in CLASS and business-as-usual conditions. In particular, per hypothesis 6, it was expected that the BEST in CLASS condition would demonstrate higher adherence and competence subscale scores than the business-as-usual condition (Hemmeter et al., 2016). To test this hypothesis, group mean differences on the Adherence and Competence subscales at posttest were used to examine the group differences after training and coaching. Independent sample t-tests were conducted to test this mean difference to examine group differences at posttest. Linear regression analyses were then computed using MPlus software Version 8 (Muthen & Muthen, 1998-2017) to test whether the intervention condition predicts differences in posttest scores. This method was used to correct standard errors due to the nested data structure (i.e., children nested within teachers) via the sandwich estimator available in MPlus was used (Diggle et al., 2002).

Measurement Sensitivity

To assess measurement sensitivity over time, linear growth models were computed using MPlus software Version 8 (Muthen & Muthen, 1998-2017). Per hypothesis 7, it was hypothesized that scores on the BiCACS Adherence and Competence subscales would demonstrate positive change over the course of training and coaching for BEST in CLASS teachers (Sutherland et al., 2014; Sutherland et al., 2018). To test this, two growth models were used to evaluate change in scores on the Adherence and Competence subscales, respectively. All seven timepoints were included in analyses with the intercept fixed at pretest to account for baseline scores. Since this research question is concerned with whether the BiCACS Adherence and Competence subscales are sensitive to the training and coaching provided in the BEST in CLASS condition, only BEST in CLASS teachers were included in these two models. Lastly, to control for the effects of nesting (i.e., children within teachers), the sandwich estimator in MPlus was used (Diggle et al., 2002).

Results

Preliminary Analyses

Sample Bias

As found in Tables 2 and 3, chi-square analyses and analyses of variance revealed no statistically significant differences in the characteristics of the teacher and child participants at pretest between the BEST in CLASS and business-as-usual conditions. Further, no statistically significant differences (i.e., all *p-values* > .05) were found between teacher demographic variables (i.e., sex, race/ethnicity, education level, and years of teaching experience) between the current sample and the excluded teacher participants from the parent study. There were also no statistically significant differences in child demographics variables (i.e., sex, race/ethnicity, age, and baseline Social Skills and Problem Behavior SSIS subscales) between the excluded child participants from the parent study and the current sample. Lastly, there were no significant differences found between the aforementioned teacher and child demographic variables from the current sample and the reliability sample, a subset of the current sample. As such, the current sample and reliability sample included in these analyses are representative of the parent study, and the reliability sample is also representative of the current sample.

Data screening and Data Distribution

Means, standard deviations, ranges, skewness, and kurtosis were explored for items on the BiCACS Adherence and Competence subscales across the seven time points (see Table 6). All items on both subscales demonstrated a range of six, indicating that the full range of scores from one to seven was used. Skewness and kurtosis values across all Adherence and Competence items fell within acceptable limits (i.e., between -2 and +2; George & Mallory, 2010). No outliers were detected at the item level (i.e., z-score > |3.29|), and therefore no items were removed for analysis.

Reliability Analyses

Interrater Reliability: BiCACS Adherence and Competence items

Interrater reliability was conducted using the available double-coded data in which the BiCACS was coded by two independent raters during the same live observation of a teacher's classroom instruction. A total of 529 observations (26% of total observations) were available for reliability analyses. All seven timepoints and both conditions were included in the analyses. Table 7 presents ICC(2,1) data including 95% confidence intervals and the frequency of which each item was rated. The single-measure ICC was used since only a subset of the total observations were double-coded (McLeod et al., 2013). For Adherence items, ICCs ranged from "good" to "excellent" (i.e., .67 to .82; *M* = 0.74, *SD* = 0.06). ICCs for the Competence items ranged from "poor" to "fair" (i.e., .29 to .52; *M* = 0.46, *SD* = 0.14). All ICCs for the Competence items fell within the "fair" range, with one exception: the ICC for Behavior Specific Praise (i.e., .29) fell in the "poor" range. Though, the 95% confidence interval of the ICC for Behavior Specific Praise did not include zero (i.e., 95% CI = .17 – .41), and therefore was retained in subsequent analyses. The ICC for the Child Responsiveness item was "good" (i.e.,

.60), and ICC for the Child Engagement item was “fair” (i.e., .53). No ICCs demonstrated a 95% confidence interval including zero. Given that all Adherence item ICCs were above .60, these ICCs were found to be as hypothesized (hypothesis 1) and provide evidence that Adherence item scores are reliable. However, the Competence item ICCs were found to be lower than hypothesized (hypothesis 1) which suggests that item scores on the Competence subscale Adherence subscale may be harder to code than items on the Adherence subscale.

Interrater Reliability: BiCACS Adherence and Competence subscales

Scores on the BiCACS Adherence and Competence subscales were created by averaging together the Adherence or Competence items. For the Competence subscale, items were included in calculating the Competence subscale score only if both observers rated the Competence item a “1” or greater; otherwise, these items were dropped from the Competence subscale computation. The Adherence and Competence subscales were normally distributed and demonstrated a full range. Descriptive data and data distribution information can be found in Table 6.

After creating the subscales, reliability analyses were conducted for each subscale score. The ICC(2,1) coefficients, frequency, and confidence intervals of the Adherence and Competence subscales can be found in Table 7. The Adherence subscale demonstrated an “excellent” ICC score (i.e., .81) and exhibited a 95% confidence interval that did not include zero. This ICC was found to be as hypothesized (hypothesis 2) and suggests that the Adherence subscale score is reliable. However, similar to the item-level analyses, the ICC of the Competence subscale was “fair” (i.e., .43) but the 95% confidence interval did not include zero and therefore the subscale was retained for subsequent analyses. This finding did not support the hypothesis (hypothesis 2) as it was expected for the ICC of the Competence subscale to achieve at least “good” reliability.

Validity Analyses

Construct Validity: BiCACS Adherence and Competence items

To evaluate the score validity of the BiCACS items, correlations between the BiCACS Adherence items, BiCACS Competence items, the Child Responsiveness item, and the Child Engagement item were computed. Table 8 displays correlations of corresponding Adherence and Competence items (e.g., Adherence Precorrection correlated with Competence Precorrection). These correlations ranged from .58 to .69 (*M* = 0.61; *SD* = 0.05). All of the correlations were large (r > .36; Rosenthal & Rosnow, 1984), and no items were redundant (*r* > .70; Kline, 1979). It was expected that the strength of these correlations would range between medium to large, and the current correlations support this hypothesis (hypothesis 3a).

Table 9 displays intra-item correlations (i.e., correlations of items within the same subscale) among the Adherence item scores (e.g., Adherence Precorrection correlated with Adherence Corrective feedback). Intra-item correlations among the Adherence items ranged from .30 to .55 (*M* = 0.40, *SD* = 0.11) and were medium to large in strength. Additionally, intra-item correlations among the Competence items (e.g., Competence Precorrection correlated with Competence Corrective feedback), as seen in Table 10, ranged from .45 to .76 (*M* = 0.43; *SD* = 0.14) and were found to be large. Both of these findings were unexpected as it was hypothesized that all of these correlations would be medium in strength (hypothesis 3b). In total, 47% of the intra-item adherence correlations and none of the intra-item competence correlations were medium in strength. Only two items were found to be redundant with one another (i.e., Competence Instructive Feedback with Competence Corrective Feedback; *r* = .76; Kline, 1979).

As seen in Table 8, inter-item correlations among noncorresponding Adherence and Competence items (e.g., Adherence Precorrection correlated with Competence Corrective feedback) ranged from .18 – .46 (*M* = 0.32; *SD* = 0.07) and were small to large in strength. None of the correlations were redundant. It was hypothesized that these correlations would range from small to medium in strength (hypothesis 3c), which was not fully supported by the data. However, 73% of the noncorresponding correlations fell within the expected range.

Correlations between the Child Responsiveness or Child Engagement items with the Adherence or Competence items are found in Tables 9 and 10, respectively. Correlations among the Adherence items with the Child Responsiveness and Child Engagement items ranged between .19 and .30 (*M* = 0.25; *SD* = 0.03) and were small to medium in strength. This was found to be in line with the hypothesized correlation magnitude (hypothesis 3d). Correlations between the Competence items and the Child Responsiveness and Child Engagement items ranged from .22 and .39 (*M* = 0.29; *SD* = 0.05) and were found to be medium to large in magnitude. Though, only one correlation (i.e., Child Responsiveness and Opportunities to Respond) was large in magnitude (i.e., *r* = .39), which provides evidence that these findings partially support the hypothesis.

The mean of the absolute value of the correlations for each group of analyses was computed to help determine the pattern of the correlations. Means of the various groups of correlations were then compared to determine whether the mean correlations were statistically different with one another. These follow-up contrasts were computed using Fisher r-to-z transformations. Fisher r-to-z transformations revealed that the mean of the absolute value of the inter-item correlations between corresponding BiCACS Adherence and Competence items (*M* = 0.61; *SD* = 0.05) was significantly higher than (a) the mean of the intra-item correlations among the BiCACS Adherence items (*M* = 0.40; *SD* = 0.10; *Z* = ±9.21, *p* < .001), (b) the mean of the intra-item correlations among the BiCACS Competence items (*M* = 0.43; *SD* = 0.14; *Z* = ±8.23, *p* < .001), (c) the mean of the BiCACS Adherence and Competence non-corresponding items (*M* = 0.32; *SD* = 0.07; *Z* = ±12.47, *p* < .001, (d) the mean of the correlations between the BiCACS Adherence items and the Engagement and Responsiveness items (*M* = 0.25 ; *SD* = 0.03; *Z* = ±13.05, *p* < .001), and (e) the mean of the correlations between the BiCACS Competence items and the Engagement and Responsiveness items (*M* = 0.29; *SD* = 0.05; *Z* = ±11.80, *p* < .001). This suggests that the inter-item correlations between corresponding BiCACS Adherence and Competence items were found to be significantly larger than all other BiCACS correlations.

As expected, Fisher's r-to-z transformations revealed that the mean of the absolute value of the intra-item correlations among the BiCACS Adherence items (*M* = 0.40; *SD* = 0.10) was not found to be significantly larger than the mean of the intra-item correlations among BiCACS Competence items (*M* = 0.43; *SD* = 0.14; *Z* = ±0.95, *p* = .172). The mean of the absolute value of the correlations among the BiCACS Adherence items (*M* = 0.36; *SD* = 0.11) was larger than (a) the mean of the BiCACS Adherence and Competence non-corresponding items (*M* = 0.32; *SD* = 0.07; *Z* = ±3.31, *p* < .001), (b) the mean of the correlations between the BiCACS Adherence items and the Engagement and Responsiveness items (*M* = 0.25; *SD* = 0.03; *Z* = ±4.90, *p* < .001), and (c) the mean of the correlations between the BiCACS Competence items and the Engagement and Responsiveness items (*M* = 0.29; *SD* = 0.05; *Z* = ±3.65, *p* < .001). The mean of the absolute value of the intra-item correlations between the BiCACS Competence items (*M* = 0.43; *SD* = 0.14) was larger than (a) the mean of the BiCACS Adherence and Competence non-corresponding items (*M* = 0.32; *SD* = 0.07; *Z* = ±4.24, *p* < .001), (b) the mean of the correlations between the BiCACS Adherence items and the Engagement and Responsiveness items (*M* = 0.25; *SD* = 0.03; *Z* = ±5.73, *p* < .001), and (c) the mean of the correlations between the BiCACS Competence items and the Engagement and Responsiveness items (*M* = 0.29; *SD* = 0.05; *Z* =

±4.47, *p* < .001). These comparisons suggest that the correlations among the Adherence items and among the Competence items are similar in magnitude. Further, these correlations are significantly larger than the correlations among non-corresponding BiCACS Adherence and Competence items and the correlations with the Engagement and Responsiveness items.

Fisher's r-to-z transformations revealed that the mean of the absolute value of the inter-item correlations between BiCACS Adherence and Competence non-corresponding items (*M* = 0.32; *SD* = 0.07) was significantly larger than the mean of the correlations between the BiCACS Adherence items and the Engagement and Responsiveness items (*M* = 0.25; *SD* = 0.03; *Z* = ±1.96, *p* = .025), but was not significantly different than the mean of the correlations between the BiCACS Competence items and the Engagement and Responsiveness items (*M* = 0.29; *SD* = 0.05; *Z* = ±0.70, *p* = .241). This suggests that the inter-item correlations between BiCACS Adherence and Competence non-corresponding items are significantly larger than the correlations between the BiCACS Adherence items and the Engagement and Responsiveness items, but not the correlations between the BiCACS Competence items and the Engagement and Responsiveness items. Lastly, the mean of the absolute value of the correlations between the BiCACS Adherence items and the Engagement and Responsiveness items (*M* = 0.25; *SD* = 0.03) was not significantly different than (a) the mean of the correlations between the BiCACS Competence items and the Engagement and Responsiveness items (*M* = 0.29; *SD* = 0.05; *Z* = ±1.14, *p* = .127). This suggests that the correlations between the Engagement and Responsiveness items with the Adherence items are not significantly larger or smaller than their correlations with the Competence items. Overall, the pattern of correlations provide support for the construct validity of the BiCACS Adherence and Competence items. More specifically, the pattern of the correlations display overlap among items that were expected to display overlap among each other (i.e., convergent validity), but also provided evidence that the items were largely distinct (i.e., discriminant validity)

Construct Validity: BICACS Adherence and Competence subscales

Construct validity was also assessment at the subscale level. To evaluate the score validity, correlations among the Adherence subscale, Competence subscale, Child Responsiveness item, and the Child Engagement item were computed. Correlations can be found in Table 11. To create the Adherence and Competence subscale score for validity analyses, Adherence or Competence items were averaged together to create the corresponding subscale score. The correlation among the BiCACS Adherence and Competence subscales was large in magnitude (r = .47, p <.01, n = 2159), which supports the hypothesis (hypothesis 4a). As expected (hypothesis 4b), (a) a medium correlation was found the BiCACS Adherence subscale and the Child Engagement item (r = .34, p < .01, n = 1393), and (b) a small correlation was found between the BiCACS Competence subscale and the Child Engagement item (r = .20, p < .01, n = 1393). A medium correlation was found between the BiCACS Competence subscale and the Child Responsiveness item (r = .32, p <.01, n = 1393), which was hypothesized (hypothesis 4c). However, unexpectedly, a large correlation was found between the BiCACS Adherence subscale and the Child Responsiveness item (r = .38 p <.01, n = 1393). These findings are contradictory to the proposed hypotheses as it was expected for these correlations to be small to medium in magnitude (hypothesis 4c). Follow-up contrasts were computed using the Fisher r-to-z transformation to determine the overall pattern of the correlations. It was found that the correlation between the BiCACS Adherence and Competence subscales (r = .47, p <.01) was significantly larger than the correlations between the (a) BiCACS Adherence subscale and Child Responsiveness item (r = .38, p <.01; Z = ±3.17, p = .001), (b) BiCACS Competence subscale and Child Responsiveness item (*r* = .32, *p* <.01*; Z* = ±5.20, *p* < .001), (c) BiCACS Adherence subscale and Child Engagement item (*r* = .34, *p* <.01*; Z* = ±4.81, *p* < .001), and (d) the Competence subscale and Child Engagement item (*r* = .20, *p* <.01*; Z* = ±9.05, *p* < .001).

To further assess the construct validity of the BiCACS Adherence and Competence subscales, Person-product moment correlations were computed between the BiCACS Adherence and Competence subscales with the two subscales of the Student-Teacher Relationship Scale (STRS): Closeness and Conflict. These correlations were evaluated using pretest data as previously explained. The correlations among these subscales can be found in Table 12. The correlation between the BiCACS Adherence and Competence subscales at pretest was large in magnitude (*r* = .39, *p* = < .001, *n* = 397). The correlations between the STRS Closeness subscale with the Adherence subscale (*r* = .11, *p* = .149, *n* = 167) and Competence subscale (*r* = .13, *p* = .096, *n* = 167) were small in magnitude. Additionally, the correlations between the STRS Conflict subscale with the Adherence subscale *(r* = -.06, *p* = .446, *n* = 164) and Competence subscale (*r* = -.09, *p* = .233, *n* = 164) were small in magnitude. All correlations were found to be in the expected direction and range (hypothesis 5a and 5b).

Follow-up contrasts were computed using the Fisher r-to-z transformation to determine the overall pattern of the correlations. It was found that the correlation between the BiCACS Adherence and Competence subscales (*r* = .39, *p* <.01) were significantly larger than the correlations between the (a) Adherence subscale and the STRS Closeness subscale (*r* = .11, *p* = .149; *Z* = 3.20, *p* < .001), (b) Competence subscale and the STRS Closeness subscale (*r* = .13, *p* = .096; *Z* = 3.01, *p* < .001), (c) Adherence subscale and the STRS Conflict subscale (*r* = -.06, *p* = .446; *Z* = 5.02, *p* < .001), and (d) Competence subscale and the STRS Conflict subscale (*r* = -.09, *p* = .233; *Z* = 5.39, *p* < .001). This suggests that the correlation between the BiCACS Adherence and Competence subscale is significantly larger in magnitude than all the correlations among the BiCACS subscales with the STRS subscales.

Together, these results indicate that the overall pattern and magnitude of the correlations presented largely support the hypotheses (hypothesis 4 and 5). In particular, the correlation between the Adherence and Competence subscales was found to be large and also significantly larger than the correlations between the Adherence or Competence subscale with the (a) Child Responsiveness, (b) Child Engagement items, (c) STRS Closeness subscale, and (d) STRS Conflict subscale. Together, these correlations provide evidence of the construct validity of the BiCACS at the subscale level.

Discriminative Validity

Independent samples t-tests between BEST in CLASS and business-as-usual teachers scores on the Adherence and Competence subscales at posttest were used to assess the discriminative validity of scores on the BiCACS Adherence and Competence subscales. Results indicated significant differences between group conditions for the BiCACS Adherence and Competence subscales. The BEST in CLASS condition (*M* = 4.32, *SD* = 1.36) was found to have higher Adherence subscale scores at posttest than the business-as-usual condition (*M* = 2.06, *SD* = 0.70), t(375) = 20.12, *p* < .001, Cohen's *d* = 2.09 (Lakens, 2013). Similarly, the BEST in CLASS condition (*M* = 5.46, *SD* = 0.98) was also found to have higher Competence subscale scores at posttest then the business-as-usual condition (*M* = 4.29, *SD* = 1.05), t(361) = 11.21, *p* < .001, Cohen's *d* = 1.15. To determine if nesting (i.e., children nested within teachers) influenced group comparisons, linear regression analyses using the sandwich estimator (Diggle et al., 2002) were computed using MPlus software Version 8 (Muthen & Muthen, 1998-2017). The sandwich estimator was used to control for the effects of nesting (i.e., children within teachers). Results indicated that group condition significantly predicted both the Adherence (β = .72, p < .001) and Competence (β = 1.17, p < .001) subscale scores. Taken together, results indicate that the BEST in CLASS condition demonstrated statistically significant higher scores on the BiCACS Adherence and Competence subscales than business-as-usual condition at posttest. These results were as expected, such that it was hypothesized that the BEST in CLASS condition would demonstrate higher scores than business-as-usual at posttest due to receiving training and coaching on the BEST in CLASS program (hypothesis 6). Since this hypothesis was as expected, this finding provides support of the discriminative validity of the BiCACS at the subscale level.

Measurement Sensitivity

Measurement sensitivity was assessed by computing two linear growth models using MPlus version 8 (Muthen & Muthen, 1998-2017). Analyses were designed to determine whether a positive linear growth was observed for teachers undergoing training and coaching in BEST in CLASS. Therefore, BiCACS Adherence and Competence subscale scores from the BEST in CLASS condition were used. To test whether there was a positive growth over time after assessing baseline levels of adherence and competence, the intercept in each model was fixed at pretest. As hypothesized, results of the linear growth models indicated a significant positive growth over time for the BiCACS Adherence subscale (b = .326, p < .001; y = 2.621, p < .001) and the BiCACS Competence subscale (b = .226, p < .001; y = 4.201, p < .001). Though, goodness of fit statistics for the Adherence (χ^2(23) = 136.01, p < .001; RMSEA = .153; CFI = .438; TLI = .487) and Competence (χ^2(23) = 80.92, p < .001; RMSEA = .109; CFI = .791; TLI = .809) subscale growth models were poor. Due to the poor fit, two quadratic growth models were computed using the BiCACS Adherence and Competence subscales. Similar to the linear models, results of the quadratic growth models indicated a significant growth for the BiCACS

Adherence subscale (*b* = .659, *p* < .001; *y* = 2.353, *p* < .001; *q* = -.058, *p* <.001) and BiCACS Competence subscale (*b* = .474, *p* < .001; *y* = 3.903, *p* < .001, *q* = -.036, *p* <.001). Fit statistics indicated that the quadratic growth models were a poor fit for the BiCACS Adherence (χ^2(19) = 57.64, *p* < .001; RMSEA = .098; CFI = .808; TLI = .788) and Competence (χ^2(19) = 52.75, *p* < .001; RMSEA = .092; CFI = .878; TLI = .866) subscales. However, Santorra-Bentler Scaled Chi-Square Difference Test (Santorra & Benter, 2010) indicated that the quadratic models of the Adherence ($\Delta\chi^2_{\mathrm{SB}}$ = 74.94, Δ*df* = 4.00, *p* < .001) and Competence ($\Delta\chi^2_{\mathrm{SB}}$ = 25.98, Δ*df* = 4.00, *p* < .001) subscales fit better than their respective linear growth models. Taken together, results of each model were as hypothesized (hypothesis 7) such that the models revealed a positive slope in the BiCACS Adherence and Competence subscales over the course of training and coaching for the BEST in CLASS teachers. This suggests that the BiCACS displays evidence of sensitivity over the course of training and coaching.

Discussion

With the increased use of treatment integrity instruments that assess how much (i.e., adherence) and how well (i.e., competence) an intervention was delivered in early childhood settings, these instruments must be evaluated to ensure that they adequately capture the constructs that they are designed to assess. Treatment integrity instruments are critical to determining whether an intervention was delivered as intended, which is needed to accurately interpret intervention outcomes of randomized-controlled trials (Sanetti et al., 2020; Sutherland et al., 2021). However, treatment integrity instruments that assess adherence and competence, and have been rigorously evaluated for score reliability and validity are lacking (Sanetti et al., 2020; Sutherland et al., 2013; Sutherland et al., 2021). As such, the present study aimed to assess the psychometric properties of a treatment integrity instrument used to evaluate adherence and competence of an intervention (i.e., BEST in CLASS) delivered in early childhood settings.

The present study examined the score reliability and validity of the BEST in CLASS Adherence and Competence Scale (BiCACS; Sutherland et al., 2014), an instrument designed to assess the adherence and competence of the core practices of the BEST in CLASS program. Findings generally supported the score reliability and validity of the BiCACS. Adherence item and subscale scores produced by independent raters demonstrated at least “good” reliability, whereas the Competence item and subscale scores demonstrated primarily “fair” reliability. The scores on the BiCACS items and subscales also revealed evidence of construct validity. In particular, the documented pattern of correlations between the Adherence and Competence items and subscales provided evidence of both convergent and discriminant validity. When correlating the BiCACS subscale scores with an instrument that assessed a different construct (i.e., student-teacher relationships), the correlations were weaker than the correlations among the BiCACS subscales, providing evidence of discriminant validity. Group differences in the level of Adherence and Competence scores at posttest between BEST in CLASS and business-as-usual teachers provided evidence of discriminative validity, suggesting that the BiCACS can detect differences in treatment integrity between intervention conditions. Lastly, scores on the BiCACS demonstrated a positive growth over the course of training, meaning that scores on the BiCACS appear to be sensitive to change in treatment integrity across time. Overall, these findings suggest that scores on the BiCACS Adherence and Competence items and subscales predominately demonstrated evidence of score reliability and validity across various domains.

Reliability

Descriptive data for the BiCACS Adherence and Competence items indicated that the items demonstrated the full ranges of scores (i.e., 1 to 7), and no items were found to be skewed or kurtotic. Single-measure interrater reliability was assessed for all Adherence and Competence items and their related subscale. Intra-class correlation (ICC) coefficients of the BiCACS items ranged from “poor” to “excellent” (Cicchetti, 1994); none of the ICCs had a confidence interval that included zero. The level of reliability between Adherence and Competence was discrepant such that the ICC scores fell into different ICC classifications. All Adherence items demonstrated at least “good” interrater reliability (*M ICC* = .74; *SD* = .06; range, .67 – .82), but Competence item scores fell in the “poor” and “fair” categories (*M ICC* = .46; *SD* = .14; range, .29 – .52). A similar pattern in reliability was found when ICCs were computed at the subscale level. The Adherence subscale score demonstrated “excellent” interrater reliability (ICC = .81), whereas the Competence subscale score fell in the “fair” range (ICC = .43). The difference in the level of reliability achieved by the Adherence scores versus the Competence scores suggests that competence may be harder to code.

The item-level reliability estimates for the Adherence items and subscales are similar to previous research with the BiCACS. Sutherland et al. (2014) found that during live observations, the single-measure reliability of the Adherence items ranged from “fair” to “excellent” (*M ICC* = .72; *SD* = .15; range, .44 – .91), and the Adherence subscale displayed “excellent” reliability (ICC[2,1] = .90). In contrast, the current study found that the score reliability of the Competence items and subscale were lower than previous findings. Sutherland et al. (2014) reported that the Competence items fell within the “poor” to “excellent” ranges (*M ICC* = .64; *SD* = .16; range .39 – .85), and the Competence subscale displayed “excellent” reliability (ICC[2,1] = .85).

However, the discrepancy between adherence and competence reliability has also been seen in research conducted in mental health at the item (Hogue et al., 2008) and subscale level (Carroll et al., 2000). For example, Hogue et al. (2008) reported Adherence item reliability to span from “fair” to “excellent” (range = .56 – .83) and competence reliability to span from “poor” to “good” (range = .01 – .63). At the subscale level, Carroll et al. (2000) found that six Adherence subscales displayed “excellent” reliability (range = .80 – .95), whereas the Competence subscales were at “good” to “excellent” (range = .71 – .97). This further suggests that competence may be harder to code than is adherence.

One reason for the lower estimates for the Competence items and subscales could be because the current study evaluated the reliability of the BiCACS using single-measure ICC estimates. Single-measure ICC was appropriate for the current study because only a subset of all observations was coded by a second observer. Single-measure ICC estimates produce lower estimates than the average-measure ICC estimates (Koo & Lee, 2016; McLeod et al., 2013). This means that if all observations were double-coded, then it is possible the reliability estimates for BiCACS items would be higher. For example, the single-measure ICC for the Competence subscale was found to be in the “fair” range (i.e., ICC = .43), but the average-measure ICC produces an estimate in the “good” range (i.e., ICC = .61). This average-measure ICC is undoubtedly better than the single-measure ICC; however previous investigations of the BiCACS have reported the average-measure ICC for the Competence subscale to fall in the “excellent” range (i.e., ICC[2,2] = .84; Sutherland et al., 2014).

Another reason the Competence items and subscale may have demonstrated lower estimates than the Adherence items and subscale is the BiCACS rating procedures. When coders rate the BiCACS, they produce adherence ratings for every observation but only rate competence when a teacher delivered a BEST in CLASS practice (i.e., if adherence was scored above a “1”). This rating procedure results in fewer observations for the Competence items than the Adherence items. For example, in the current study the interrater reliability for *Adherence Precorrection* was based on 529 observations, whereas the interrater reliability for *Competence Precorrection* was based on 144 observations. Since a larger number of observations typically results in better reliability estimates (Koo & Lee, 2016), this rating procedures may explain the lower estimates for Competence. In conclusion, the findings suggest that independent raters are able to reliably score the BiCACS Adherence and Competence items. Though, researchers who want to use individual Competence items in analyses may need to consider double coding all observations produce higher reliability estimates.

Construct Validity

Findings provide evidence supporting the construct validity of the BiCACS items and subscales. Evidence of convergent validity at the item level was found. For example, the strength of the correlations among corresponding Adherence and Competence items (e.g., Adherence Precorrection correlated with Competence Precorrection) were large, and the mean of the correlations among these items was significantly stronger than the mean correlations among other relations between items (e.g., Adherence items, Competence items, noncorresponding adherence and Competence items). This suggests that the correlations among the corresponding BiCACS Adherence and Competence items evidenced a stronger relation than correlations among the Adherence items, Competence items, noncorresponding Adherence and Competence items, and with the Child Responsiveness and Engagement items. In addition, the correlations among the Adherence items and the correlations among the Competence items were medium to large, and the mean of these correlations was not statistically significant from each other. This suggests that these items overlap more with items on their respective subscale, as the within-subscale items measure a similar treatment integrity component (e.g., adherence or competence). At the subscale level, the correlation among the Adherence and Competence subscales was large (i.e., *r* = .47), and the strength of this correlation was statistically stronger than correlations among the Adherence or Competence subscale with a measure of another construct (i.e., child responsiveness, child engagement, student-teacher relationship quality). These findings provide evidence of convergent validity by demonstrating that the relation among the Adherence and Competence subscale is stronger than the relation among the BiCACS subscales with measures that assess child responsiveness, child engagement, or teacher-child relationship quality. Previous research in education (i.e., *rs* = .38 – .82, Sutherland et al., 2014) and mental health (e.g., *r* = .38, Carroll et al., 2000; *r* = .46, McLeod et al., 2018; *r* = .42, Hogue et al., 2008) have also reported large correlations among adherence and competence at the item and subscale level. However, in contrast to the current study, Sutherland et al. (2014) reported a larger correlation among the Adherence and Competence subscale (*r* = .71), one that suggests these subscales may be redundant (i.e., *r* > .70; Kline, 1979) and assess similar constructs. This discrepancy may have been due to variability in the sample as the current study included both BEST in CLASS and business-as-usual teachers and included a larger sample size. Though, it is promising that the correlation among the BiCACS subscales are closer to those reported in the mental health literature. In sum, correlations among the BiCACS items and subscales provide evidence of convergent validity at the item and subscale-levels.

Findings also supported the discriminant validity of the BiCACS at the item and subscale-level. First, the correlations among the BiCACS items suggest that the items are distinct. Only one correlation across all BiCACS items was found to be redundant (*r* < .70;

Kline, 1979), with the rest of the inter-item correlations ranging from small to large, suggesting that the items are not redundant. Second, the mean correlation among the Adherence (*r* = .40) and the Competence (*r* = .43) items was significantly larger than the correlations among the noncorresponding Adherence and Competence items (*r* = .32) as well as with the Child Responsiveness and Child Engagement items (*rs* = .25 – .29). Previous research supports this finding as the relation among client involvement (i.e., the degree to which a client is involved in therapeutic activities) with Adherence and Competence scores ranges from small to medium (i.e., *rs* = .13 – .24; McLeod et al., 2018). A similar pattern was seen for the BiCACS subscales. As previously mentioned, the correlation among the Adherence and Competence subscale was large (*r* = .47), but the magnitude also suggests that the subscales are distinct from one another (*r* < .70; Kline, 1979). Correlations among the Adherence or Competence subscales with the Child Responsiveness and Child Engagement items were small to large in magnitude (i.e., .20 – .38). Though, follow-up contrasts revealed that these correlations are significantly smaller than the correlation among the Adherence and Competence subscale. Also, the correlations among the Adherence or Competence subscales with the subscales of an instrument of teacher-child relationships (i.e., Closeness and Conflict) were small in magnitude (i.e., *rs* = ± .06 – .13). Follow-up contrasts revealed that the correlation between the Adherence and Competence subscales was significantly larger than the correlations among the Closeness and Conflict subscales with the BiCACS subscales. This indicates that the BiCACS subscales are distinct from the Closeness and Conflict subscales, and therefore these instruments measure distinct constructs and provide further evidence of discriminant validity for the BiCACS subscales. Previous investigations of the BiCACS supports these findings as Sutherland et al. (2014) also found that the correlation among the BiCACS Adherence and Competence subscales was significantly higher than the correlations among the Adherence or Competence subscale with the Closeness and Conflict subscale. This finding is also consistent with previous research conducted in mental health such that the magnitude of the correlations among the adherence or competence scores with alliance, a construct that assesses the quality of the therapeutic relationship between a therapist and client, is shown to be smaller than the correlation between adherence and competence (Carroll et al., 2000; Hogue et al., 2008). Together, the pattern of findings supports the discriminant validity of scores on the BiCACS items and subscales.

In sum, scores on the BiCACS items and subscales evidenced a pattern of correlations that support convergent and discriminant validity. The evidence of the current study provides the strongest support for the discriminant validity of the BiCACS subscales. It is important for treatment integrity instruments to demonstrate evidence of both convergent and discriminant validity so researchers can confidently assume that the instrument is assessing the accurate treatment integrity constructs they wish to assess (Cross et al., 2015; Cross & West, 2011). Treatment integrity instruments that are found to assess the constructs they purport to measure allow researchers to make sound conclusions about the efficacy of the interventions they are developing and evaluating.

One factor that limits confidence in the convergent validity is that the current study did not have another treatment integrity instrument that assesses adherence and competence. Convergent validity is best evaluated by comparing scores of the same construct across two or more different measurement instruments (Foster & Cone, 1995; Hill & Lambert, 2004). Since no additional treatment integrity instrument was used in the parent study, the construct validity should continue to be investigated as research progresses. However, the patterns and magnitude of the correlations among the items and subscales presented in the current study prove promising in supporting the construct validity of the BiCACS.

Discriminative Validity

Findings support the discriminative validity of the BiCACS Adherence and Competence subscale scores. There was evidence of discriminative validity demonstrated by differences between intervention conditions in the level of Adherence and Competence subscale scores at posttest. It was expected that these two groups would significantly differ after training was completed. It was found that scores between the Adherence and Competence subscale scores predicted treatment conditions after controlling for data nesting. It was also found that both the Adherence (Cohen's d = 2.09) and Competence (Cohen's d = 1.15) subscale scores of the BEST in CLASS condition were significantly higher than the subscale scores of the business-as-usual condition. Taken together these findings suggests that Adherence and Competence subscale scores can detect differences between teachers who have been trained in BEST in CLASS from those teachers who have not been trained in the intervention.

The difference in treatment integrity subscale scores between intervention conditions is supported by previous research in education (Hemmeter et al., 2016) and mental health (McLeod et al., 2019; Smith et al., 2017). Hemmeter et al. (2016) conducted a randomized-controlled trial of the Pyramid Model program which aims to promote positive social-emotional outcomes and decrease problem behavior in early childhood settings. They found that teachers in the Pyramid Model condition demonstrated higher adherence (*ES* = 1.95) scores at posttest than the business-as-usual teachers; though, competence data was not assessed. Additionally, McLeod et al. (2019) demonstrated that a treatment integrity instrument assessing adherence and another assessing competence were able to detect differences across therapists providing a cognitive-behavioral therapy intervention in research and community settings. They found that therapists providing therapy in research settings demonstrated higher levels of adherence to cognitive-behavioral skills (*ES* = .57), adherence to exposure interventions (*ES* = 1.17), and overall competence (*ES* = 1.13) than therapists in the community setting. Further, Smith et al. (2017) found that therapists providing cognitive-behavioral therapy in research settings had higher adherence ratings in the middle and end phase of data collection compared to therapists providing therapy in community settings; both of these conditions demonstrated higher level of adherence than therapists in a control condition (i.e., usual care). In sum, previous research supports the findings of the current study by providing evidence that treatment integrity instruments have been known to detect changes across treatment conditions. As such, current study findings suggest that the BiCACS subscales can detect differences between intervention groups. However, more research is needed in the education field to support the discriminative validity of the Competence subscale. This finding implies that the BiCACS can be used to evaluate the efficacy of a program, especially in determining whether the level of adherence and competence supports that the intervention was delivered as intended compared to a control group.

Measurement Sensitivity

The measurement sensitivity of the BiCACS subscale scores was assessed by examining growth models. These analyses were aimed at evaluating if the BiCACS subscale scores captured change in adherence and competence over the course of the study. Results indicated that the Adherence and Competence subscale scores evidence positive growth during the study. This suggests that teacher adherence and competence increased over the course of the study, and that the BiCACS subscales were able to capture this change. This is an important finding because it implies that the BiCACS can capture the variability of adherence and competence over time.

Previous investigations in education have also found that the average adherence score increased across phases of data collection (Hemmeter et al. 2016; Sutherland et al., 2014). However, Sutherland et al. (2014) did not find that teacher competence changed over time, which is discrepant from the current study, and Hemmeter et al. (2016) did not report competence data. This discrepancy may be due to differences in sample size. The current sample used a larger sample which may have provided a more precise estimate of adherence and competence.

Similarly, research conducted in the mental health field have also found Adherence subscale scores to change over time. Smith et al. (2017) found that adherence to cognitive-behavioral therapy increased over time for therapists trained in a cognitive-behavioral therapy intervention; though, change in competence score was not tested in this study. Additionally, McLeod et al. (2019) tested whether scores on a subscale designed to assess adherence changed across two phases of treatment delivery (i.e., Skills and Exposure) for a cognitive-behavioral therapy intervention. Results indicated that during the Exposure phase, adherence to skill-building decreased as adherence to exposure therapy increased (McLeod et al., 2019). McLeod et al. (2019) did not find any changes over time in level of competence across data collection. In sum, previous research supports the finding that the BiCACS Adherence subscale can detect change over time. However, future research will need to replicate the current findings to help address the discrepant finding for the BiCACS Competence subscale across the current study and previous research.

Summary

The current study aimed to establish whether the BEST in CLASS Adherence and Competence Scale (i.e., BiCACS; Sutherland et al., 2014) demonstrated evidence of score reliability and validity. Findings suggest that the Adherence items and subscale produced at least “good” reliability estimates across coders. However, all Competence items may need to be double-coded to achieve at least “good” reliability at the item and subscale-level. It is important for treatment integrity scores to exhibit good reliability at the item level so researchers can run analyzes that examine the effects of distinct intervention practices. Good reliability at the subscale level is also important in order for researchers to draw conclusions about intervention efficacy. If reliability evidence is not present, then the item and subscale scores of the treatment integrity instrument are essentially inaccurate (Sanetti & Kratochwill, 2009; Sanetti et al., 2020).

Construct validity evidence was also apparent in the current study, such that the correlations among BiCACS items and subscales produced a pattern associated with convergent and discriminant validity evidence. Findings also displayed evidence of discriminant validity through small to medium correlations among the BiCACS subscales with the subscales of an instrument that assesses the quality of student-teacher relationships. These findings are promising as it is critical for treatment integrity instruments to evidence construct validity when evaluating interventions. More specifically, for researchers to derive accurate conclusions about whether an intervention is efficacious and improves targeted child outcomes, they will need to understand whether the intervention was delivered as intended (Sanetti & Kratochwill, 2009; Sanetti et al., 2020). When scores on treatment integrity instruments show evidence of construct validity, it means that the treatment integrity instrument assesses the distinct treatment integrity components (e.g., adherence, competence) that it was intended to measure (Hill & Lambert, 2004). This is true for both the item and subscale scores because where the items can help researchers determine whether distinct core practices of an intervention drive treatment outcomes, the subscale score can help determine whether adherence and competence of the intervention as a whole was delivered as intended (Perepletchikova et al., 2007; Sanetti et al., 2009; Sutherland et al., 2013). Though, when construct validity is questionable, it implies that researchers cannot accurately ascertain whether an intervention produced results because it was delivered as designed or due to another factor.

The results of the current study also add to treatment integrity research in the field of education. Since a review of the literature suggests that researchers under-report validity evidence of the treatment integrity instruments they utilize in randomized-controlled trials, this current study adds to the literature by reporting these findings. That is, as the literature grows and more studies report their results, the field may be able to estimate the extent to which treatment integrity constructs overlap with one another (Sanetti et al., 2020). This will be helpful information as it will provide a guideline for which researchers can base their construct validity findings and make conclusions about their treatment integrity instruments.

Lastly, this current study provides evidence of elaborative validity (i.e., whether the scores produced by an instrument have utility; Foster & Cone, 1995). In particular, results imply that the BiCACS can accurately distinguish between intervention groups (i.e., discriminative validity). This is helpful for researchers in deciphering whether an intervention was delivered as intended compared to a control group that is presumed to achieve lower levels of treatment integrity. It is also helpful information for researchers in establishing the level of post-intervention adherence and competence across intervention conditions to conclude the level of treatment integrity needed to produce favorable treatment outcomes. Additional evidence of elaborative validity in the current study is that the instrument displayed sensitivity to change in adherence and competence over time. In particular, the Adherence and Competence subscale scores increased over time which suggests that the BiCACS is capable of capturing variability in these treatment integrity domains across training and coaching (Sutherland et al., 2014). It is essential for treatment instruments to capture this variability because it can provide researchers with pertinent information about the impacts of the training and coaching supports intended to improve adherence and competence (Cross et al., 2015; Cross & West, 2011; Fiske, 2008).

Limitations

Although the current study displayed its strengths, it also has some limitations. First, the present study examined the psychometrics of a treatment integrity instrument designed to assess the adherence and competence of the core practices of the BEST in CLASS program. As such, teachers in the current study included those implementing the BEST in CLASS program and a control group. Therefore, these findings may not generalize to other early childhood interventions. Still, these findings have merit in providing estimates of score reliability and validity that other researchers can use to establish guidelines for interpreting the psychometric properties of their treatment integrity instruments.

Additionally, although evidence of the convergent and discriminant validity is promising, findings cannot conclude whether the BiCACS actually assesses the treatment integrity domains of adherence and competence. Results imply that some items and subscales either assess similar or different constructs. However, the current study cannot conclude that the items and subscales assess adherence and competence because the present study did not include other treatment integrity instruments to act as a comparator. Comparing the BiCACS to an independent measure that assesses adherence and competence could provide further evidence of convergent and discriminant validity. For example, if the additional instrument assessed adherence and competence, the correlations could be run across this instrument and the BiCACS to evaluate the overlap among similar constructs. If these correlations were large in magnitude, and even considered redundant, it would be more accurate to assume that the BiCACS assesses adherence and competence (i.e., convergent validity). To further assess discriminant validity, if findings suggested that the BiCACS and an established instrument that assesses additional domains of treatment integrity (e.g., Responsiveness) were distinct, one could conclude that the BiCACS does not capture another treatment integrity component.

Lastly, measurement sensitivity findings demonstrated increased growth in adherence and competence over time. However, the goodness of fit statistics indicated that neither the linear nor the quadratic growth models were a good fit for the data. This means that a linear or quadratic shape may not adequately capture change in Adherence and Competence scores.

Future Directions

Due to the limitations mentioned above and general guidelines for assessing psychometric properties of treatment integrity instruments, there are several avenues for future research. First, it is important to replicate the current study's findings across multiple samples, studies, and contexts. BiCACS was designed to evaluate the adherence and competence of the BEST in CLASS program, so replication of this study using a different teacher-delivered intervention would not be appropriate. However, future studies could attempt to replicate the current study with a different sample of participants or assessment of another context (e.g., web-based delivery of the BEST in CLASS; Conroy et al., 2020). This would help to further ground the reliability and validity of the BiCACS while also assessing any factors that may influence its psychometric properties (Martinez et al., 2014). For example, these studies could address further factors that may have been associated with the discrepancy between the adherence and competence reliability estimates.

As previously delineated, future research should investigate the construct validity of the BiCACS with another established treatment integrity instrument. As such, future research should include multiple treatment integrity instruments of similar and different constructs to evaluate the convergent and discriminant validity across these instruments. Since sound treatment integrity instruments used in early educational constructs are still largely in development (Sanetti et al., 2020; Sutherland et al., 2021), it would be beneficial for researchers to use these multiple instruments to inform the recommendations for establishing instruments with score validity in the field.

Lastly, future investigations should replicate the measurement sensitivity analyses with the BiCACS subscales due to discrepant findings with previous research. The aim of this would be to reconcile whether the BiCACS Competence scores reflect a positive growth over time as was found in the current study or no change over time as was found in Sutherland et al. (2014). Additionally, future work could also determine recommendations to assess the measurement sensitivity analyses as the growth curve models computed in the current study were not a good fit for the data.

In sum, as more researchers develop and utilize treatment integrity instruments, they must examine the score reliability and validity of the instrument. This information is important as it establishes whether scores produced by the instrument are accurate, intend to measure the constructs it was designed to, and has evidence of utility. The psychometric evaluation of the BiCACS may prove to be a helpful example for researchers conducting assessments on treatment integrity instruments. Such instruments that display evidence of these psychometric domains of reliability and validity are needed to guide conclusions about the delivery and efficacy of evidence-based interventions delivered in early childhood contexts.

Reference List

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms & profiles*. Burlington, VT: University of Vermont: Research Center for Children, Youth, & Families.

Barnett, W. S., Jung, K., Yarosz, D. J., Thomas, J., Hornbeck, A., Stechuk, R., & Burns, S. (2008). Educational effects of the Tools of the Mind curriculum: A randomized trial. *Early Childhood Research Quarterly*, *23*(3), 299–313. https://doi.org/10.1016/j.ecresq.2008.03.001

Basten, M., Tiemeier, H., Althoff, R. R., Van de Schoot, R., Jaddoe, V. W., Hofman, A., Hudziak, J. J., Verhulst, F.C. & Van der Ende, J. (2016). The stability of problem behavior across the preschool years: An empirical approach in the general population. *Journal of Abnormal Child Psychology*, *44*(2), 393-404. https://doi.org/ 10.1007/s10802-015-9993-y

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., … Gill, S. (2008). Promoting academic and social-emotional school readiness: The Head Start REDI program. *Child Development*, *79*(6), 1802–1817. https://doi.org/10.1111/j.1467-8624.2008.01227.x

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C., & Domitrovich, C. E. (2008). Executive functions and school readiness intervention: Impact, moderation, and mediation in the Head Start REDI program. *Development and Psychopathology*, *20*(3), 821–843. https://doi.org/10.1017/S0954579408000394

Brennan, L. M., Shaw, D. S., Dishion, T. J., & Wilson, M. (2012). Longitudinal predictors of school-age academic achievement: Unique contributions of toddler-age aggression, oppositionality, inattention, and hyperactivity. Journal of Abnormal Child Psychology, 40(8), 1289–1300. http://dx.doi.org/10.1007/ s10802-012-9639-2

Campbell, S. B., Shaw, D. S., & Gilliom, M. (2000). Early externalizing behavior problems: Toddlers and preschoolers at risk for later maladjustment. *Development and Psychopathology*, *12*(3), 467–488. https://doi.org/10.1017/s0954579400003114

Carroll, K. M., & Nuro, K. F. (2002). One size cannot fit all: A stage model for psychotherapy manual development. *Clinical Psychology: Science and Practice*, *9*(4), 396–406. https://doi.org/10.1093/clipsy/9.4.396

Carroll, K. M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., … 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*(4), 674–680. https://doi.org/10.1037/a0033021

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Conroy, M. A., Sutherland, K. S., Algina, J. J., Wilson, R. E., Martinez, J. R., & Whalon, K. J. (2015). Measuring teacher implementation of the BEST in CLASS intervention program and corollary child outcomes. *Journal of Emotional and Behavioral Disorders*, *23*(3), 144–155. https://doi.org/10.1177/1063426614532949

Conroy, M. A., Sutherland, K. S., Algina, J., Werch, B., & Ladwig, C. (2018). Prevention and treatment of 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

Cross, W., & West, J. (2011). Examining implementer fidelity: Conceptualising and measuring adherence and competence. *Journal of Children's Services*, *6*(1), 18–33. https://doi.org/10.5042/jcs.2011.0123

Cross, W., West, J., Wyman, P. A., Schmeelk-Cone, K., Xia, Y., Tu, X., … Forgatch, M. (2015). Observational measures of implementer fidelity for a school-based preventive intervention: Development, reliability, and validity. *Prevention Science*, *16*(1), 122–132. https://doi.org/10.1007/s11121-014-0488-9

DeVellis, R. F. (2017). *Scale development: Theory and applications*. Thousand Oaks, CA: SAGE Publications, Inc.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K. Y., Heagerty, P. J., & Zeger, S. (2002). *Analysis of longitudinal data*. United Kingdom: Oxford University Press.

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., … Ialongo, N. S. (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

Domitrovich, C. E., Cortes, R. C., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the preschool PATHS curriculum. *The Journal of Primary Prevention*, *28*(2), 67–91. https://doi.org/10.1007/s10935-007-0081-0

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly*, *25*(3), 284–298. https://doi.org/https://doi.org/10.1016/j.ecresq.2010.04.001

Driscoll, K. C., & Pianta, R. C. (2010). Banking time in Head Start: Early efficacy of an intervention designed to promote supportive teacher-child relationships. *Early Education and Development*, *21*(1), 38–64. https://doi.org/10.1080/10409280802657449

Dunlap, G., Strain, P., Lee, J. K., Joseph, J., & Leech, N. (2018). A randomized controlled evaluation of Prevent-Teach-Reinforce for young children. *Topics in Early Childhood Special Education*, *37*(4), 195–205. https://doi.org/doi.org/10.1177/0271121417724874

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

Feil, E. G., Frey, A., Walker, H. M., Small, J. W., Seeley, J. R., Golly, A., & Forness, S. R. (2014). The efficacy of a home-school intervention for preschoolers with challenging behaviors: A randomized controlled trial of Preschool First Step to Success. *Journal of Early Intervention, 36*(3), 151–170. https://doi.org/10.1177/1053815114566090

Feil, E. G., Severson, H. H., & Walker, H. M. (1998). Screening for emotional and behavioral delays: The Early Screening Project. *Journal of Early Intervention*, *21*(3), 252–266. https://doi.org/10.1177/ 105381519802100306.

Finsaas, M. C., Bufferd, S. J., Dougherty, L. R., Carlson, G. A., & Klein, D. N. (2018). Preschool psychiatric disorders: Homotypic and heterotypic continuity through middle childhood and early adolescence. *Psychological Medicine*, *48*(13), 2159–2168. https://doi.org/10.1017/S0033291717003646

Fiske, K. E. (2008). Treatment integrity of school-based behavior analytic interventions: A review of the research. *Behavior Analysis in Practice*, *1*(2), 19–25. https://doi.org/10.1007/BF03391724

Fleiss, J. L. (1981). Balanced incomplete block designs for interrater reliability studies. *Applied Psychological Measurement*, *5*(1), 105–112. https://doi.org/10.1177/014662168100500115

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment*, *7*(3), 248–260. https://doi.org/10.1037/1040-3590.7.3.248

Foundation, H. S. E. R. (2014). *High Scope Curriculum*. Ypsilanti, MI: High Scope Educational Research Foundation.

Furr, R. M., & Bacharach, V. R. (2014). Estimating and evaluating convergent and discriminant validity evidence. In *Psychometrics: An introduction* (pp. 221–268). Thousand Oaks, CA: SAGE Publications, Inc.

George, D., & Mallery, M. (2010). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a ed.) Boston: Pearson.

Gresham, F. M. (2009). Evolution of treatment integrity concept: Current status and future directions. *School Psychology Review*, *38*(4), 533–540.

Gresham, F., & Elliott, S. N. (2008). *Social Skills Improvement System (SSIS-RS) rating scales*. Bloomington, MN: Pearson Assessments.

Gresham, F. M. (2014). Measuring and analyzing treatment integrity data in research. In L. M. H. Sanetti & T. R. Kratochwill (Eds.), *Treatment integrity: A foundation for evidence-based* *practice in applied psychology* (pp. 109–130). Washington, DC: American Psychological Association.

Gresham, F. M., Elliott, S. N., Vance, M. J., & Cook, C. R. (2011). Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly*, *26*(1), 27–44. https://doi.org/10.1037/a0022662

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*(2), 625–638. https://doi.org/10.1111/1467-8624.00301

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2012). Promoting young children's social competence through the preschool PATHS curriculum and MyTeachingPartner professional development resources. *Early Education and Development*, *23*(6), 809–832. https://doi.org/10.1080/10409289.2011.607360

Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, *25*(3), 329–347. https://doi.org/10.1016/j.ecresq.2009.07.002

Han, S. S., Catron, T., Weiss, B., & Marciel, K. K. (2005). A teacher-consultation approach to social skills training for pre-kindergarten children: Treatment model and short-term outcome effects. *Journal of Abnormal Child Psychology*, *33*(6), 681–693. https://doi.org/doi.org/10.1007/s10802-005-7647-1

Hansen, W. B., Graham, J. W., Wolkenstein, B. H., & Rohrbach, L. A. (1991). Program integrity as a moderator of prevention program effectiveness: Results for fifth-grade students in the adolescent alcohol prevention trial. *Journal of Studies on Alcohol*, *52*(6), 568–579. https://doi.org/10.15288/jsa.1991.52.568

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: Using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology*, *27*(5), 711–731. https://doi.org/10.1023/a:1022194005511

Hemmeter, M. L., Hardy, J. K., Schnitz, A. G., Adams, J. M., & Kinder, K. A. (2015). Effects of training and coaching with performance feedback on teachers' use of Pyramid Model practices. *Topics in Early Childhood Special Education*, *35*(3), 144–156. https://doi.org/10.1177/0271121415594924

Hemmeter, M. L., Snyder, P. A., Fox, L., & Algina, J. (2016). Evaluating the implementation of the Pyramid Model for promoting social-emotional competence in early childhood classrooms. *Topics in Early Childhood Special Education*, *36*(3), 133–146. https://doi.org/10.1177/0271121416653386

Heroman, C., Burts, D. C., Berke, K., & Bickart, T. S. (2010). *Teaching Strategies GOLD objectives for development & learning: Birth through kindergarten*. Bethesda, MD: Teaching Strategies, LLC.

Heroman, C., Trister Dodge, D., Berke, K., Bickart, T. S., Colker, L., Jones, C., & Dighe, J. (2010). *The Creative Curriculum for Preschool* (5th ed.). Bethesda, MD: Teaching Strategies, LLC.

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J.Lambert (Ed.), *Handbook of psychotherapy and behavior change* (pp. 84– 135). New York, NY: Wiley.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, *33*(2), 258–270. https://doi.org/10.1080/02796015.2004.12086247

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inclan, J., … 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., 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), 332–345. https://doi.org/10.1037/0033-3204.33.2.332

Kam, C.-M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science*, *4*(1), 55–63. https://doi.org/10.1023/A:1021786811186

Kazdin, A. E. (2016). Single-case experimental research designs. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 459–483). Washington, DC, US: American Psychological Association.

Keller, T. E., Spieker, S. J., & Gilchrist, L. (2005). Patterns of risk and trajectories of preschool problem behaviors: A person-oriented analysis of attachment in context. *Development and Psychopathology*, *17*(2), 349–384. https://doi.org/10.1017/S0954579405050170

Kline, P. (1979). Psychometrics and psychology. London, UK: Academic Press.

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

Lakens D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Little, J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.2307/2290157

Martinez, R. G., Lewis, C. C., & Weiner, B. J. (2014). Instrumentation issues in implementation science. Implementation Science, 9, 118. https://10.1186/s13012-014-0118-8

McLeod, B. D., Islam, N. Y., & Wheat, E. (2013). Designing, conducting, and evaluating therapy process research. In J. Comer & P. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 142–164). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793549.013.0009

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., Rodríguez, A., Quinoy, A. M., Arnold, C. C., Kendall, P. C., & Weisz, J. R. (2018). Development and initial psychometrics for a therapist competence instrument for CBT for youth anxiety. *Journal of Clinical Child & Adolescent Psychology*, *47*(1), 47-60. https://doi.org/10.1080/15374416.2016.1253018

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review*, *38*(4), 541–546.

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

Mesman, J., Bongers, I. L., & Koot, H. M. (2001). Preschool developmental pathways to preadolescent internalizing and externalizing problems. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(5), 679–689. https://doi.org/10.1017/S0021963001007351

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340. https://doi.org/10.1177/109821400302400303

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén

Newborg, J. (2005). *Battelle Developmental Inventory*. Rolling Meadows, IL: Riverside Publishing.

Nix, R. L., Bierman, K. L., Domitrovich, C. E., & Gill, S. (2013). Promoting children's social-emotional skills in preschool can enhance academic and behavioral functioning in kindergarten: Findings from Head Start REDI. *Early Education and Development*, *24*(7), 1000–1019. https://doi.org/10.1080/10409289.2013.825565

Noell, G. H., & Gansle, K. A. (2014). Research examining the relationships between consultation procedures, treatment integrity, and outcomes. In W. P. Erchul & S. M.

Sheridan (Eds.), *Handbook of research in school consultation* (2nd ed., Vol. 2, pp. 386–408). New York, NY: Routledge.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., & Duhon, G. J. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review*, *34*(1), 87–106. https://doi.org/10.1080/02796015.2005.12086277

Pas, E. T., Waasdorp, T. E., & Bradshaw, C. P. (2015). Examining contextual influences on classroom-based implementation of positive behavior support strategies: Findings from a randomized controlled effectiveness trial. *Prevention Science*, *16*(8), 1096–1106. https://doi.org/10.1007/s11121-014-0492-0

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

Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology*, *18*(2), 148–153. https://doi.org/10.1111/j.1468-2850.2011.01246.x

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. (2001). *Student-Teacher Relationship Scale: Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Pianta, R. C., & Rimm-Kaufman, S. (2008). The social ecology of the transition to school: Classrooms, families, and children. In K. McCartney & D. Phillips (Eds.), *Blackwell* *handbook of early childhood development* (pp. 490–507). Wiley-Blackwell. https://doi.org/10.1002/9780470757703.ch24

Raver, C. C., Jones, S. M., Li-Grining, C. P., Metzger, M., Champion, K. M., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly*, *23*(1), 10–26. https://doi.org/https://doi.org/10.1016/j.ecresq.2007.09.001

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *77*(2), 302–316. https://doi.org/10.1037/a0015302

Rescorla, L. A., Achenbach, T. M., Ivanova, M. Y., Bilenberg, N., Bjarnadottir, G., Denner, S., … Verhulst, F. C. (2012). Behavioral/emotional problems of preschoolers: Caregiver/teacher reports from 15 societies. *Journal of Emotional and Behavioral Disorders*, *20*(2), 68–81. https://doi.org/10.1177/1063426611434158

Sanetti, L. M. H., Charbonneau, S., Knight, A., Cochrane, W. S., Kulcyk, M. C., & 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

Sanetti, L. M. H., & Fallon, L. M. (2011). Treatment integrity assessment: How estimates of adherence, quality, and exposure influence interpretation of implementation. *Journal of Educational and Psychological Consultation, 21*(3), 209–232. https://doi.org/10.1080/10474412.2011.595163

Sanetti, L. M. H., 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/10.1177/0143034313476399

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.

Satorra, A., & Bentler, P. M. (2009). Ensuring Positiveness of the Scaled Difference Chi-square Test Statistic. Psychometrika, 75(2), 243–248. doi:10.1007/s11336-009-9135-y

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*(1), 32–43. https://doi.org/10.1007/s10488-010-0321-0

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review*, *38*(4), 460–475

Settanni, M., Longobardi, C., Sclavo, E., Fraire, M., & Prino, L. E. (2015). Development and psychometric analysis of the Student-Teacher Relationship Scale – Short Form. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00898

Sheridan, S. M., Swanger-Gagné, M., Welch, G. W., Kwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review*, *38*(4), 476–495.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

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

Snyder, P. a, Hemmeter, M. L., Fox, L., Bishop, C. C., & Miller, M. D. (2013). Developing and gathering psychometric evidence for a fidelity instrument: The Teaching Pyramid Observation Tool–Pilot Version. *Journal of Early Intervention*, *35*(2), 150–172. https://doi.org/10.1177/1053815113516794

Standing, L. G. (2004). Halo effect. In M. S. Lewis-Black, A. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods, volume 1* (p. 452). Thousand Oaks, CA: Sage Publications, Inc.

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*(1), 31–43. https://doi.org/10.1016/j.ecresq.2017.08.001

Sutherland, K. S., Conroy, M. A., Vo, A., & Ladwig, C. (2015). Implementation integrity of practice-based coaching: Preliminary results from the BEST in CLASS efficacy trial. *School Mental Health*, *7*(1), 21–33. https://doi.org/10.1007/s12310-014-9134-8

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. (2013). Measuring implementation of evidence-based programs targeting young children at risk for emotional/behavioral disorders. *Journal of Early Intervention*, *35*(2), 129–149. https://doi.org/10.1177/1053815113515025

Tabachnik, B. G., & Fidell, S. L. (2007). *Using multivariate statistics.* Boston, MA: Pearson Education Inc.

Tsigilis, N., & Gregoriadis, A. (2008). Measuring teacher-child relationships in the Greek kindergarten setting: A validity study of the Student-Teacher Relationship Scale–Short Form. *Early Education and Development*, *19*(5), 816–835. https://doi.org/10.1080/10409280801975826

Webster-Stratton, C., Jamila Reid, M., & Stoolmiller, M. (2008). Preventing conduct problems and improving school readiness: Evaluation of the Incredible Years teacher and child training programs in high-risk schools. *Journal of Child and Adolescent Psychiatry*, *49*(5), 471–488. https://doi.org/10.1111/j.1469-7610.2007.01861.x

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2001). Preventing conduct problems, promoting social competence: A parent and teacher training partnership in Head Start. *Journal of Clinical Child Psychology*, *30*(3), 283–302. https://doi.org/10.1207/S15374424JCCP3003_2

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child & Adolescent Psychology*, *33*(1), 105–124. https://doi.org/10.1207/S15374424JCCP3301_11

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of*

Clinical Child & Adolescent Psychology, *33*(1), 105–124. https://doi.org/10.1207/S15374424JCCP3301_11

Wolery, M. (2011). Intervention research: The importance of fidelity measurement. *Topics in Early Childhood Special Education*, *31*(3), 155–157. https://doi.org/10.1177/0271121411408

Appendix A

Tables

Table 1

Treatment Integrity Data Reported in Randomized-Controlled Trials for Teacher-delivered Evidenced-Based Programs for Early Classroom Settings

Program	RCT	Adh?	Comp?	Metrics?	Notes on Instrument
BEST in CLASS	Conroy et al. (2018)	**YES**	**YES**	**YES ***	Coach and/or research staff report (7-pt scale); Interrater reliability, construct validity
	Sutherland et al. (2018)	**YES**	**YES**	**YES ***	Coach and/or research staff report (7-pt scale); Interrater reliability, construct validity
Chicago School Readiness Project	Raver et al. (2008)	NO	NO	NO	
	Raver et al. (2009)	**YES**	**YES**	NO	Consultant report (response format unknown)
Incredible Years	Webster-Stratton et al. (2001)	NO	NO	NO	
	Webster-Stratton et al. (2004)	NO	NO	NO	
	Webster-Stratton et al. (2008)	NO	NO	NO	
Head Start REDI	Bierman et al. (2008)	**YES**	**YES**	NO	Teacher report (3-pt scale); Trainer report (6-pt scale)
	Nix et al. (2013)	NO	NO	NO	
Preschool PATHS	Domitrovich et al. (2007)	NO	**YES**	NO	Coordinator report (3-pt scale)
	Hamre et al. (2012)	**YES**	NO	NO	Observer report (checklist)
Preschool First Step (PFS) to Success	Feil et al. (2014)	**YES**	**YES**	**YES**	Observer report (5-pt scale); Interrater reliability
Prevent-Teach-Reinforce	Dunlap et al. (2018)	**YES**	NO	NO	Research staff (checklist)
Pyramid Model	Hemmeter et al. (2016)	**YES**	NO	**YES ***	Observer report (checklist); Interrater reliability; convergent validity
RECAP	Han et al. (2005)	NO	NO	NO	
Tools of the Mind	Barnett et al. (2008)	**YES**	NO	**YES**	Observer report (checklist); Reliability data available

Note. RCT = randomized-controlled trial; Adh = RCT reported adherence; Comp = RCT reported competence; Metrics = RCT reported psychometrics. * Additional studies that examine the psychometric properties for these instruments exist (see: Sutherland et al., 2014; Snyder et al., 2013)

Table 2

Teacher Demographics for Current Study

Variable	BiC (n = 89)	BAU (n = 90)	Total (N = 179)	F or χ^2	*p*
Sex (%)				1.00	.317
Female	99	97	97.8		
Male	1	3	2.2		
Race/Ethnicity (%)				0.65	.957
White/European American	49	46	47.2		
Black/African American	46	50	47.8		
Hispanic/Latinx	2	2	2.2		
Asian American/Pacific Islander	1	1	1.1		
No Report	2	1	1.7		
Highest level of education (%)				2.56	.769
High School	1	3	2.2		
Associate Degree	30	29	29.61		
Bachelor’s Degree	41	39	39.7		
Master’s Degree	25	28	26.6		
Doctoral Degree	1	0	0.6		
Other	2	1	1.7		
Average years of teaching experience (SD)	11.5 (9.1)	12.9 (10.1)	12.2 (9.6)	0.95	.330

Note. Analysis of variance was conducted with continuous variables, and chi-square analyses were conducted with categorical variables. Demographic comparisons presented in this table were conducted between the two study conditions. BiC = BEST in CLASS; BAU = business-as-usual.

Table 3

Child Demographics for Current Study

Variable	BiC (n = 211)	BAU (n = 205)	Total (N = 416)	F or χ^2	*p*
Sex (%)				0.04	.846
Male	65	65	64.9		
Female	35	35	35.1		
Race/Ethnicity (%)				0.98	.324
White/European American	17	16	16.3		
Black/African American	66	69	67.8		
Hispanic/Latinx	4	4	4.1		
Asian American/Pacific Islander	0	1	0.2		
Native American	1	0	0.2		
Other	7	6	6.7		
No report	5	4	4.6		
Average Age at Entry (SD)	4.3 (0.6)	4.4 (0.6)	4.36 (0.56)	2.25	.813
Average Baseline SSIS-RS (SD)					
Problem Behavior	120.6 (15.1)	120.7 (16.9)	120.7 (16.0)	0.01	.941
Social Skills	77.6 (12.7)	78.7 (12.0)	78.2 (12.3)	0.85	.358

Note. Analysis of variance was conducted with continuous variables, and chi-square analyses were conducted with categorical variables. Demographic comparisons presented in this table were conducted between the two study conditions. BiC = BEST in CLASS; BAU = business-as-usual. SSIS = Social Skills Improvement Scale (Gresham & Elliott, 2008).

Table 4

Definitions of BEST in CLASS Modules

Module/Practice	Content
Rules	How to implement intentional, frequent, and targeted guidelines that provide structure to help the focal child learn what is expected during activities in the classroom
Precorrection	How to communicate and set specific instructions or prompts that tell the focal child the expected behavior before a challenging behavior is likely to occur
Opportunities to Respond	How to engage focal children through different techniques such as questions, prompts, or signals during instructional activities
Behavior Specific Praise	How to use frequent, targeted, and specific praise statements during instructional activities with the focal child
Corrective Feedback	How to increase the use of providing feedback following a challenging behavior or incorrect response that teaches the focal children an appropriate alternative response or behavior
Instructive Feedback	How to increase focal children’s engagement by providing extra instructional information following a correct response or answer
Linking and Mastery	How to combine and use the practices sequentially while interacting with the focal child

Table 5

BEST in CLASS Adherence and Competence Scale Item Definitions

Adherence/Competence item	Definition
1. Rules	Teacher provides a statement that includes a classroom rule.
2. Precorrection	Teacher provides instruction or prompts to remind child of appropriate classroom behavior
3. Opportunities to Respond	Teacher provides question, prompt, or signal to child seeking a preacademic, social, or behavioral Child Responsiveness
4. Behavior Specific Praise	Teacher provides a statement to a child that indicates specific, labeled approval of a child behavior
5. Corrective Feedback	Teacher provides information to child after a preacademic or behavioral error occurs
6. Instructive Feedback	Teacher provides extra instructional information to child after a correct preacademic or appropriate behavioral response occurs
Child Items	**Definition**
1. Child Responsiveness	Focal child reacts to teacher’s attempts at using BEST in CLASS practices
2. Child Engagement	Focal child participates appropriately and actively on assigned or approved classroom activity

Table 6

BiCACS Item-level Descriptive Data

Item	N	Range	*M*	*SD*	Skewness	Kurtosis
Adherence						
Teacher Reviews Rules	2150	6	3.10	2.21	0.51	-1.28
Precorrection	2155	6	2.55	2.07	1.04	-0.40
Opportunities to Respond	2156	6	5.20	1.57	-0.63	-0.36
Behavior Specific Praise	2158	6	2.58	1.97	1.11	-0.06
Corrective Feedback	2159	6	2.23	1.67	1.34	0.77
Instructive Feedback	2159	6	2.18	1.66	1.45	1.15
Subscale	2159	6	2.97	1.33	0.70	-0.23
Competence						
Teacher Reviews Rules	1240	6	5.01	1.56	-0.6	-0.41
Precorrection	994	6	4.65	1.57	-0.41	-0.6
Opportunities to Respond	2115	6	5.09	1.37	-0.55	-0.22
Behavior Specific Praise	1175	6	4.79	1.61	-0.49	-0.61
Corrective Feedback	1053	6	4.33	1.53	-0.26	-0.73
Instructive Feedback	1032	6	4.39	1.57	-0.29	-0.67
Subscale	2123	6	4.60	1.26	-0.40	-0.30
Additional Child Items						
Child Responsiveness	1393	6	5.46	1.28	-0.86	0.41
Child Engagement	1393	6	5.74	1.24	-1.01	0.73

Note. Number of observations vary across Adherence items due to missing data due to data entry error and vary across Competence items due to scoring procedure (i.e., a score of “0” is marked as “not observed” and was not included in analyses as it is captured by adherence scores).

Table 7

BiCACS Interrater Reliability Data

			95% CI	
Item	N	ICC(2,1)	Lower	Upper
Adherence				
Rules	526	.82	.80	.85
Precorrection	529	.70	.67	.74
Opportunities to Respond	529	.75	.71	.78
Behavior Specific Praise	529	.80	.77	.83
Corrective Feedback	529	.67	.63	.72
Instructive Feedback	529	.70	.66	.74
Adherence subscale	529	.81	.78	.84
Competence				
Rules	256	.49	.39	.58
Precorrection	144	.47	.33	.59
Opportunities to Respond	505	.52	.45	.58
Behavior Specific Praise	222	.29	.17	.41
Corrective Feedback	182	.41	.28	.52
Instructive Feedback	178	.42	.29	.53
Subscale	506	.43	.36	.50
Additional Child Items				
Child Responsiveness	342	.60	.52	.66
Child Engagement	341	.53	.45	.60

Note. Number of observations vary across Adherence items due to missing data due to data entry error and vary across Competence items due to scoring procedure (i.e., a score of “0” is marked as “not observed” and was not included in analyses as it is captured by adherence scores).

Table 8

Inter-item Correlations Between BICACS Adherence and Competence items

	Competence item					
Item	1	2	3	4	5	6
Adherence						
1. Teacher Reviews Rules	**.69****	.46**	.38**	.35**	.25**	.28**
2. Precorrection	.31**	**.59****	.31**	.22**	.18**	.21**
3. Opportunities to Respond	.38**	.35**	**.58****	.25**	.29**	.30**
4. Behavior Specific Praise	.31**	.32**	.34**	**.55****	.28**	.27**
5. Corrective Feedback	.30**	.26**	.36**	.35**	**.62****	.44**
6. Instructive Feedback	.30**	.26**	.37**	.36**	.43**	**.63****

Note. The bolded numbers in the diagonal represent correlations among corresponding Adherence and Competence items (e.g., Adherence Precorrection and Competence Precorrections). Non-bolded numbers represent correlations among non-corresponding Adherence and Competence items (e.g., Adherence Precorrection and Competence Corrective Feedback); ** = p < .01.

Table 9

Intra-item Adherence Correlations and Correlations Between BiCACS Adherence/Child Items

	Adherence item					
Item	1	2	3	4	5	6
Adherence						
1. Rules	—					
2. Precorrection	.55**	—				
3. Opportunities to Respond	.37**	.34**	—			
4. Behavior Specific Praise	.54**	.45**	.35**	—		
5. Corrective Feedback	.35**	.32**	.32**	.42**	—	
6. Instructive Feedback	.38**	.34**	.30**	.42**	.61**	—
Child						
7. Child Responsiveness	.28**	.27**	.26**	.30**	.25**	.26**
8. Child Engagement	.24**	.22**	.19**	.29**	.24**	.24**

Note. ** = p < .01

Table 10

Intra-item Adherence Correlations and Correlations Between BiCACS Adherence/Child Items

	Competence item					
Item	1	2	3	4	5	6
Competence						
1. Rules	—					
2. Precorrection	.57**	—				
3. Opportunities to Respond	.53**	.49**	—			
4. Behavior Specific Praise	.50**	.55**	.45**	—		
5. Corrective Feedback	.46**	.52**	.47**	.57**	—	
6. Instructive Feedback	.49**	.54**	.52**	.63**	.76**	—
Child						
7. Child Responsiveness	.34**	.27**	.39**	.27**	.34**	.29**
8. Child Engagement	.25**	.24**	.33**	.27**	.29**	.23**

Note. ** = p < .01

Table 11

Correlations among the Adherence and Competence subscales and with the Child Items

Subscale/Item	1	2	3
1. Adherence subscale	—		
2. Competence subscale	.47**	—	
3. Child Responsiveness	.38**	.32**	—
4. Child Engagement	.34**	.20**	.75**

Note. ** = p < .01

Table 12

Correlations Between the BiCACS and STRS subscales

Subscale/Item	1	2	3
1. Adherence subscale	—		
2. Competence subscale	.39**	—	
3. STRS Closeness	.11	.13	—
4. STRS Conflict	-.06	-.09	-.19*

Note. ** = p < .01; * = p < .05.

Appendix B

Figures

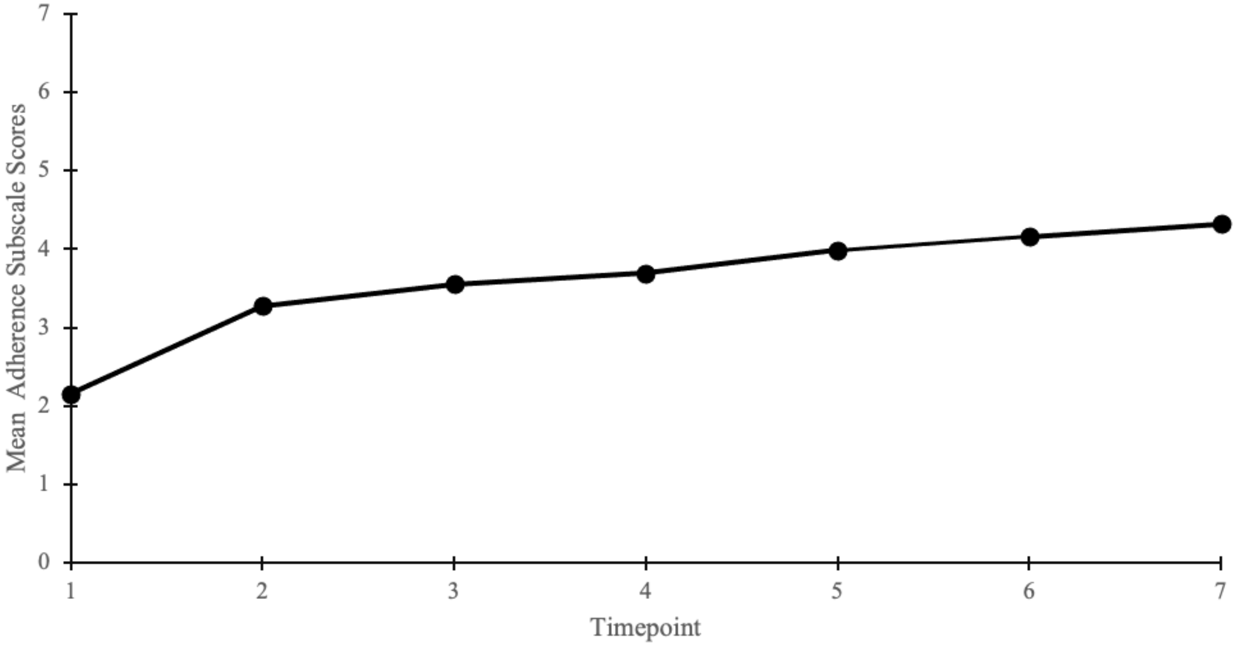

Figure 1. Change in Average Adherence Subscale Score over Time

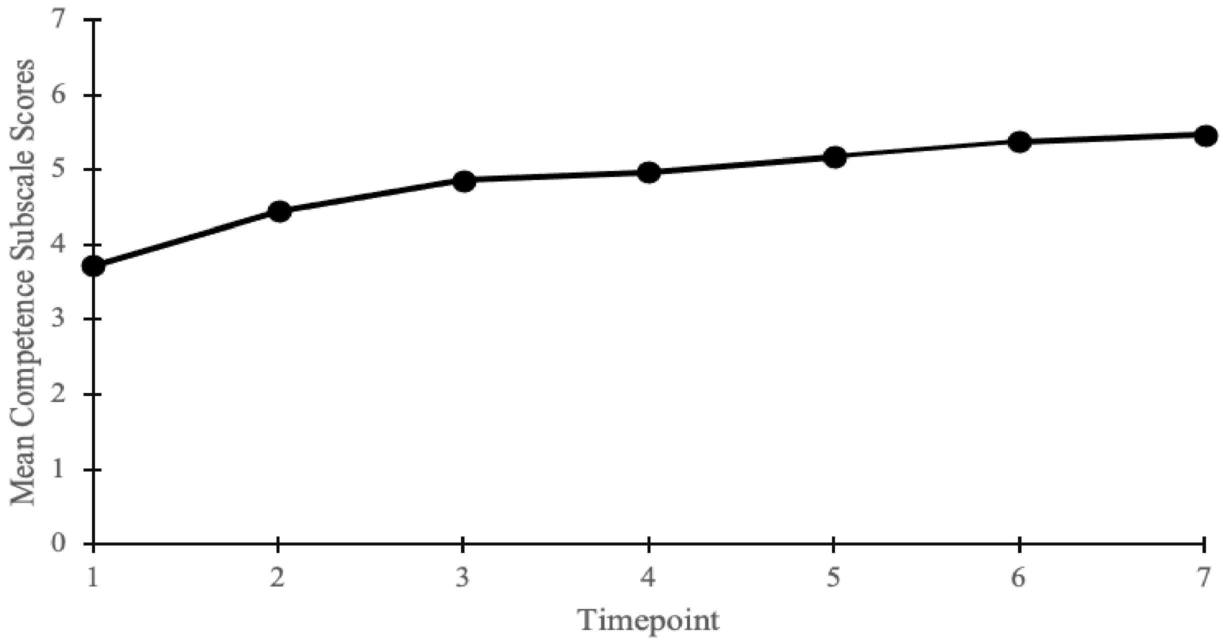

Figure 2. Change in Average Competence Subscale Score over Time

Vita

Katrina Markowicz was born on May 22, 1992, in Detroit, Michigan. She graduated with Honors from Wayne State University in December of 2014 with a Bachelor of Science in Psychology. During her undergraduate experience, she was a Research Assistant in the Family Resilience Laboratory at Wayne State University led by Dr. Erika Bocknek. Upon graduation, she simultaneously dedicated her time as a Research Assistant in three Research Laboratories at Wayne State University and was employed as a Behavior Technician by the Judson Center, where she delivered Applied Behavior Analysis intervention to youth diagnosed with Autism Spectrum Disorder. She then commenced her graduate education in the Clinical Psychology Program at Virginia Commonwealth University in Richmond, Virginia, in 2016.