Hammill Institute on Disabilities

Journal of Emotional and Behavioral Disorders
2014, Vol. 22(4) 249–259
© Hammill Institute on Disabilities 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1063426613497258
jebd.sagepub.com
SAGE

Preliminary Psychometric Properties of the BEST in CLASS Adherence and Competence Scale

Kevin S. Sutherland, PhD[1], Bryce D. McLeod, PhD[1], Maureen A. Conroy, PhD[2], Lisa M. Abrams, PhD[1], and Meghan M. Smith, BA[1]

Abstract

The measurement of treatment integrity is critical to evaluate the efficacy and effectiveness of evidence-based programs (EBPs) designed to improve the developmental outcomes of young children at risk of emotional/behavioral disorders. Unfortunately, the science of treatment integrity measurement lags behind the development and evaluation of EBP for young, high-risk children. This article describes the development and preliminary psychometric properties of the BEST in CLASS Adherence and Competence Scale (BiCACS), designed to measure the adherence and competence of delivery of the BEST in CLASS prevention program. Independent observers coded videotaped (*n* = 116) and live (*n* = 289) observations of teachers delivering the BEST in CLASS program. The BiCACS showed good interrater reliability and analyses provided some support for the validity of the measure. Implications for future research and integrity measurement work are discussed.

Keywords

treatment integrity, teacher implementation

Many young children who attend early childhood programs display high levels of problem behaviors. The severity and intensity of these children's problem behaviors negatively impact their learning (Driscoll & Pianta, 2010; Quesenberry, Hemmeter, & Ostrosky, 2011) and their teachers find them difficult to manage (Hemmeter, Corso, & Cheatham, 2006). Fortunately, there are evidence-based programs (EBPs) available to address the needs of young children in early childhood classrooms who are at elevated risk for the development of emotional and behavioral disorders (EBD). For example, Incredible Years (Webster-Stratton, Reid, & Hammond, 2004) and Preschool PATHS (Bierman et al., 2008; Domitrovich, Cortes, & Greenberg, 2007) have produced promising outcomes in multiple randomized trials. However, it is a challenge to transport and implement EBPs in authentic early childhood classrooms (Domitrovich, Moore, & Greenberg, 2012). One reason the field struggles with EBP implementation is most programs lack comprehensive professional development tools and procedures for training and supporting early childhood educators in their use.

Carroll and Nuro (2002) identify four elements that are required to develop EBPs and evaluate their effectiveness: (a) a standardized treatment model (e.g., treatment manual); (b) a well-defined target population; (c) documented and standardized procedures for selecting, training, and supervising interventionists; and (d) tools to monitor treatment integrity. Each element is designed to help investigators interpret study findings as well as aid the transportability (i.e., scale-up) of EBPs. Although some existing EBPs (e.g., Incredible Years; Preschool PATHS) meet three of the four requisites needed for this work (i.e., treatment manuals, target population, training protocols), most EBPs lack validated treatment integrity measures designed to support program evaluation and teacher training (Hagermoser Sanetti, Dobey, & Gritter, 2012; Schulte, Easton, & Parker, 2009).

Treatment integrity refers to the degree to which an EBP was delivered as intended. When developing and evaluating an EBP, it is important to develop tools to assess two components of treatment integrity (Carroll & Nuro, 2002): *treatment adherence* and *competence*. *Treatment adherence* refers to the extent to which an EBP is delivered as designed (i.e., delivery of prescribed interventions) whereas *competence* refers to the level of skill and degree

[1]Virginia Commonwealth University, Richmond, USA
[2]University of Florida, Gainesville, USA

Corresponding Author:
Kevin S. Sutherland, Department of Special Education and Disability Policy, Virginia Commonwealth University, 1015 W. Main St, PO Box 842020, Richmond, VA 23284, USA.
Email: kssuther@vcu.edu

of responsiveness demonstrated by a teacher when delivering the prescribed interventions. Assessing these integrity components allows researchers to answer key questions related to internal validity needed to support program evaluation (Carroll et al., 2000). Indeed, assessing adherence and competence allows researchers to establish whether (a) a program was delivered as designed, and (b) there was variation in delivery across teachers, classrooms, and/or schools. Treatment integrity tools also allow researchers to engage in process-outcome analyses that can help optimize the impact and delivery of an EBP (Carroll et al., 2000; Durlak, 2010). By investigating linkages between intervention components and outcomes, important questions can be asked: Were particular intervention components linked to outcomes? Do various children, classrooms, and/or settings require different intervention components to maximize outcomes? Psychometrically strong integrity measures can therefore play an important role in the development and evaluation of EBPs.

Integrity measures also play an important role in efforts to establish and maintain treatment integrity, which is integral for implementation research (Southam-Gerow & McLeod, 2013). Establishing and maintaining treatment integrity via teacher training and coaching is critical in implementation research and integrity measurement plays a significant role in this process. Treatment integrity measures can be used to (a) establish adherence and competence benchmarks used to guide teacher training efforts, and (b) assess the outcome of teacher training and coaching efforts. Integrity measures also play an important role in interpreting findings in implementation research by determining whether EBPs are implemented as designed (McLeod, Southam-Gerow, Bair, Rodriguez, & Smith, 2013). For these reasons, integrity measures are considered to be critical for implementation research.

The purpose of this article is to describe our attempt to develop and validate an integrity measure designed to support program evaluation and teacher training for a program targeting the reduction of young, high-risk children's problem behavior. Specifically, we describe the development of the BEST in CLASS Adherence and Competence Scale (BiCACS; Sutherland & McLeod, 2010). The BiCACS is an observational treatment integrity measure designed to support the development and evaluation of the BEST in CLASS program. BEST in CLASS is a theoretically driven program based on evidence-based instructional practices that target problem behaviors of young children at high risk for the development of EBD (Conroy, Sutherland, Vo, Carr, & Ogston, 2013; Sutherland, Conroy, Abrams, & Vo, 2010; Vo, Sutherland, & Conroy, 2012). Conceptualized as a "value-added" model, BEST in CLASS is designed to increase the quantity and quality of specific instructional practices that have been demonstrated to prevent and reduce the occurrence of young children's problem behaviors. Two pilot investigations of the BEST in CLASS program have provided promising initial data on the model (Conroy et al., 2013; Vo et al., 2012).

In this report, we describe the development and report on the psychometric properties of the BiCACS. We first describe the development of the BiCACS. Then, we report data from two studies designed to evaluate the psychometric properties of the BiCACS relevant to using the measure for program implementation and evaluation. First, we evaluate whether trained coders can reliably code BiCACS items using videotaped recordings of the implementation of BEST in CLASS. Second, we evaluate whether coaches can reliably code the BiCACS items. This information is important for program implementation and refinement because coaches are commonly used to establish and maintain treatment integrity in school-based trials. We therefore wanted to evaluate whether coaches could achieve adequate reliability on the BiCACS items under normal training conditions. We also examined the potential for the BiCACS to inform evaluation efforts by assessing the validity of the measure. Specifically, we examined the construct validity of the measure by examining its sensitivity to change over time as well as its relation to a measure of teacher–child relationships.

Method

BEST in CLASS

The BiCACS was developed in the context of an Institute for Education Sciences Development and Innovation Goal 2 project (see Vo et al., 2012). All data reported below were pulled from this parent project. BEST in CLASS, conceptualized as a Tier-2 program, is a manualized classroom-based program that targets the reduction of problem behaviors demonstrated by young children at risk of EBD. Teachers are trained to deliver the BEST in CLASS program via a 6-hr professional development workshop and 14 weeks of performance-based coaching provided by trained coaches (see Vo et al., 2012 for a description of the program development process). Training and coaching focus on eight learning modules: (a) Basics of Behavior and Development; (b) Rules, Expectations, and Routines; (c) Behavior-Specific Praise; (d) Precorrection and Active Supervision; (e) Opportunities to Respond and Instructional Pacing; (f) Instructive and Corrective Feedback; (g) Home–School Communication; and (h) Linking and Mastery. While the efficacy of BEST in CLASS is currently being investigated in a multisite randomized controlled trial, preliminary data from the Goal 2 project were promising. Specifically, two pilot, nonexperimental investigations of BEST in CLASS suggest that the program had an impact on observed teacher instructional and child behaviors (Conroy et al., 2013) as well as standardized measures of child behavior (Vo et al., 2012).

Table 1. BiCACS Adherence and Competence Items.

1. Teacher reviews rules, addresses rule violations—teacher statement that includes classroom rule.
2. Teacher uses clear routines (within and between activities)—teacher uses procedures and activities to provide structure.
3. Teacher maintains brisk instructional pace—rate at which teacher provides instruction.
4. Teacher provides precorrection—instruction or prompt to remind child of appropriate behavior.
5. Teacher uses proximity control and visual monitoring—teacher visually monitors child and positions herself in close proximity for child exhibiting problem behavior.
6. Teacher provides preacademic OTR—question, prompt or signal by teacher that seeks an active, observable, and specific child preacademic response.
7. Teacher provides social/behavioral OTR—question, prompt or signal by teacher that seeks an active, observable, and specific child social/behavioral response.
8. Teacher provides behavior specific praise—verbal approval statement that tells child the specific behavior for which they are being praised.
9. Teacher provides corrective feedback—specific information provided to child after error occurs.
10. Teacher provides instructive feedback—teacher statement that provides extra instructional information when responding to child's correct response or appropriate behavior.

Note. BiCACS = BEST in CLASS Adherence and Competence Scale; OTR = opportunities to respond.

Development of the BiCACS

The BiCACS is a 20-item scale designed to assess the adherence and competence of the core BEST in CLASS program components. The BiCACS was developed via a three-step process.

Step 1: Item development. Our first step was to develop items for the two BiCACS subscales: Adherence and Competence subscales. First, we identified all prescribed content from the BEST in CLASS treatment manual that represented the core BEST in CLASS interventions (see Vo et al., 2012). Second, the developers of the BEST in CLASS program reviewed the list of items to ensure that all interventions essential to the theory underlying the BEST in CLASS program were included (see Vo et al., 2012). The program developers provided feedback about item content, including suggestions for additional items. The resulting item pool was checked against the treatment manual. In all, this process generated 10 items for the Adherence and Competence subscales, respectively (see Table 1).

Step 2: Scoring strategy. Next, we determined the appropriate scoring strategy for the BiCACS subscales. For the Adherence subscale, we used a scoring strategy based on past treatment integrity research in the child psychotherapy literature (e.g., Hogue, Liddle, & Rowe, 1996; McLeod & Weisz, 2010) that involves macroanalytic extensiveness ratings. This scoring strategy requires coders to estimate the extent to which teachers engage in each intervention during an observation using a 7-point Likert-type scale with the following anchors: 1 = *not at all*, 3 = *somewhat*, 5 = *considerably*, and 7 = *extensively*. Extensiveness ratings are comprised of two key components: thoroughness and frequency. Thoroughness refers to the depth, complexity, or persistence with which the teacher engages in a given intervention. Frequency refers to the number of times throughout an observation that a given intervention is executed (regardless of the thoroughness of the intervention in any particular segment). Thoroughness and frequency are considered in making an extensiveness rating on each item; therefore, extensiveness ratings provide quantity, or dosage, information about each BEST in CLASS program.

For the Competence subscale, we adopted a scoring strategy that involves macroanalytic competence ratings that estimate the technical quality of interventions (skillfulness) and their timing and appropriateness for the given child and situation (responsiveness). This scoring strategy is used in exemplar competence coding systems developed for youth (Hogue et al., 2008) and adult (Carroll et al., 2000) psychotherapy. In assessing competence, coders are asked to make ratings on a 7-point Likert-type scale with the following anchors: 1 = *very poor*; 3 = *acceptable*; 5 = *good*; 7 = *excellent*. For each item, coders are asked to consider the extent to which a teacher demonstrated the following dimensions: (a) expertise, commitment, motivation; (b) clarity of language; (c) appropriate timing of interventions and actions (responsiveness); and (d) ability to read and respond to where the child appears to be (responsiveness).

Step 3: Scoring manual. Once the items were developed, a draft of the scoring manual was produced. The scoring manual was intended to promote interrater reliability by providing coders with clear scoring procedures, item definitions, exemplars, and item distinctions for the Adherence and Competence subscales (see Hogue et al., 1996). Coders used the scoring manual to code videotaped sessions of the BEST in CLASS program being delivered by teachers in early childhood classrooms. Feedback from the pilot coding was used to produce a revised version of the BiCACS scoring manual.

Study 1

It is important for treatment integrity measures to demonstrate reliability at the item level to support program evaluation and inform efforts to optimize the impact and delivery of an EBP. The first study was therefore conducted to evaluate the initial reliability of the BiCACS items and subscales. Trained coders rated recordings of teachers delivering the BEST in CLASS program in early childhood classrooms.

Participants

Child participants. In Years 2 and 3 of the development project, the BEST in CLASS program was implemented in 25 state- or federally funded classrooms serving high-risk children with 47 focal children within one suburban and one urban school district on the east coast. As part of program development, videotaped observations were collected in 19 classrooms (*n* = 32 child participants). These recordings were used in Study 1 to examine the initial reliability of the BiCACS.

Multiple measures were used to screen child participants for inclusion in the study. All focal children were between 3 and 5 years of age and enrolled in state and federally funded early childhood programs designed to provide services for children at elevated risk. The first two stages of the *Early Screening Project* (*ESP*; Walker, Severson, & Feil, 1995) were used to identify potential focal children. In the first stage, teachers nominated up to five children in their classrooms who demonstrated the most severe and chronic problem behaviors, and consent from parents or guardians of all nominated child participants was sought. Next, to confirm risk for EBD, teachers completed the *Externalizer Questionnaire* of the *ESP* on each of the child participants for whom consent was obtained. Children were then assessed with the *Battelle Developmental Inventory, second edition screener* (*BDI II* Screener; Newborg, 2005) and if children demonstrated average or above average cognitive/intellectual abilities, they were retained in the sample. Following this screening process, children with the top two most extreme scores on the ESP were included in the sample as the focal children.

The 32 children (24 males and 7 females; data missing for one child) in Study 1 all qualified for free and reduced lunch and averaged 3.97 years of age (*SD* = 0.32; range from 3–5). Of the children, 25 were African American, 2 were Caucasian, 1 was Latino, 1 was Asian/Pacific Islander, and 1 was Other (biracial); data were missing for two children. All children scored as "at risk" for future development of EBD on the ESP and scored within the normal range on the BDI II Total Screening Score (Newborg, 2005).

Teacher participants. Of the 19 teachers (1 male and 18 females) who volunteered to participate (6 had a bachelor's degree and 13 had a master's degree), 13 were Caucasian, 5 were African American, and 1 was Latina. They averaged 9.84 years of experience working with preschool-aged children (*SD* = 9.77; range 0–34 years).

Coders. The coding team consisted of two research assistants who were both Caucasian females. One coder was completing her BA in psychology, and one was completing her master's degree in school counseling.

BiCACS Scoring and Session Sampling Procedures

Recordings (*n* = 116) from Years 2 and 3 of the BEST in CLASS development project were selected for coding. These recordings were collected as part of the development work to examine teachers' use of BEST in CLASS interventions with focal children during instructional activities. The sessions occurred during teacher-led instructional activities (e.g., circle time, small group), ranged in length from 3 min and 19 s to 16 min and 18 s (*M* = 13.62; *SD* = 3.40), and represented times when the teacher and focal child were in the classroom. Session length was determined a priori and was based on several factors. First, indicators of high quality early childhood programs suggest that teacher-led activities should be relatively brief (e.g., 10–20 min; Cate, Diefendorf, McCullough, Peters, & Whaley, 2010). In addition, behavioral observers have found that given an adequate base rate of responses, a 10- to 20-min observation typically results in a representative sample of behavior (Thompson, Felce, & Symons, 2000). The coders trained over a 2-month period to reach adequate prestudy reliability on all BiCACS items (intraclass correlation coefficient [ICC] > .59; Cicchetti, 1994). Training consisted of reading the scoring manual, review of specific observation segments, and practice scoring of observations. Once scoring commenced, the observations were randomly assigned to coders and regular reliability assessments were performed. The results of these assessments were discussed in weekly meetings with the first two authors to prevent coder drift (Margolin et al., 1998).

Results

We first evaluated whether the Study 1 sample was comparable with the original sample. The Study 1 sample of children (*n* = 32) did not differ significantly from the original sample (*n* = 47) in demographic characteristics (e.g., age, gender, race/ethnicity). Similarly, the Study 1 sample of teachers (*n* = 19) did not significantly differ from the original sample (*n* = 25) in demographic or training characteristics (e.g., age, gender, race/ethnicity, degree, years of experience).

Table 2. Item Scores and Interrater Reliability of the BiCACS—Adherence Subscale.

	Study 1				Study 2			
Item description	*M* (*SD*)	Minimum	Maximum	ICC	*M* (*SD*)	Minimum	Maximum	ICC
Rules	2.55 (2.27)	1	7	.92	3.63 (2.28)	1	7	.90
Clear routines	5.43 (1.83)	1	7	.80	5.01 (1.32)	1	7	.44
Brisk instructional pace	6.18 (1.40)	1	7	.88	4.99 (1.42)	1	7	.59
Precorrection	3.72 (2.11)	1	7	.77	4.45 (2.33)	1	7	.87
Proximity control	6.56 (1.03)	2	7	.73	6.18 (1.25)	1	7	.77
Preacademic OTR	6.57 (1.26)	1	7	.82	5.89 (1.81)	1	7	.63
Social OTR	1.80 (1.19)	1	7	.66	6.07 (1.45)	1	7	.67
Behavior specific praise	1.93 (1.29)	1	7	.81	4.08 (2.36)	1	7	.91
Corrective feedback	3.78 (2.84)	1	7	.76	3.15 (1.77)	1	7	.70
Instructive feedback	1.41 (0.89)	1	7	.64	2.69 (1.72)	1	7	.74

Note. BiCACS = BEST in CLASS Adherence and Competence Scale; ICC = intraclass correlation coefficient; OTR = opportunities to respond.

Table 3. Item Scores and Interrater Reliability of the BiCACS—Competence Subscale.

	Study 1				Study 2			
Item description	*M* (*SD*)	Minimum	Maximum	ICC	*M* (*SD*)	Minimum	Maximum	ICC
Rules	6.43 (0.91)	4	7	.95	5.20 (1.59)	1	7	.84
Clear routines	6.40 (0.91)	3	7	.81	5.14 (1.34)	1	7	.42
Brisk instructional pace	6.59 (0.71)	4	7	.90	5.13 (1.40)	1	7	.67
Precorrection	6.18 (1.01)	2	7	.72	5.20 (1.42)	1	7	.76
Proximity control	6.67 (0.80)	3	7	.69	6.01 (1.17)	1	7	.39
Preacademic OTR	6.74 (0.61)	4	7	.72	5.65 (1.19)	2	7	.55
Social OTR	5.57 (1.18)	2	7	.56	5.62 (1.14)	2	7	.63
Behavior specific praise	6.08 (0.87)	4	7	.87	5.08 (1.47)	1	7	.85
Corrective feedback	6.11 (0.82)	4	7	.64	4.69 (1.38)	1	7	.68
Instructive feedback	5.33 (1.15)	4	7	.51	4.61 (1.38)	1	7	.65

Note. BiCACS = BEST in CLASS Adherence and Competence Scale; ICC = intraclass correlation coefficient; OTR = opportunities to respond.

To determine if coders could reliably code items on the BiCACS, interrater reliability was calculated using ICC (Shrout & Fleiss, 1979; see Tables 2 and 3). The ICC provides an estimate of the ratio of the true score variance to total variance. These correlations therefore provide a reliability estimate that allows for generalizability of the results to other samples. Following Cicchetti (1994), ICCs less than .40 reflect "poor" agreement, ICCs from .40 to .59 reflect "fair" agreement, ICCs from .60 to .74 reflect "good" agreement, and ICCs of .75 and higher reflect "excellent" agreement.

Interrater reliability was first calculated for each item on the Adherence and Competence subscales. According to Cicchetti's (1994) criteria, interrater reliability for the Adherence items ranged from "good" to "excellent" (ICCs ranged from .64 to .92 [*M* = .78, *SD* = .09], see Table 2), with 7 of the 10 items in the "excellent" range (ICC > .74), and three in the "good" range. The Adherence subscale was highly reliable (ICC = .93). The Competence items ranged from "fair" to "excellent" (ICCs ranged from .51 to .95 [*M* = .74, *SD* = .15], see Table 3), with 4 of the 10 items in the "excellent" range, 4 in the "good" range, and the remaining 2 items in the "fair" range (see Cicchetti, 1994). The Competence subscale was also highly reliable (ICC = .84). In sum, the Adherence and Competence subscales along with the items comprising the subscales demonstrated adequate reliability.

Theoretical and empirical work has thus far not clarified the amount of overlap between the adherence and competence integrity components (see Barber, Sharpless, Klostermann, & McCarthy, 2007), so we examined the degree of overlap between the Adherence and Competence subscales. Subscale scores were produced by calculating the mean score for all observations from each case on the 10 BiCACS items then averaging together the items on each subscale. The Adherence and Competence subscales evidenced moderate overlap ($r = .43$, $p < .001$). These findings suggest that there is moderate overlap among the BiCACS Adherence and Competence subscales, indicating that the subscales measure distinct content.

Study 2

The second study was conducted to evaluate whether coaches could reliably code each BiCACS item under typical training conditions. Coaches involved in training the teachers produced scores on the BiCACS items following live observations of teachers delivering the BEST in CLASS program in early childhood classrooms. For treatment integrity measures to contribute to efforts to establish and maintain treatment integrity, it is important to demonstrate reliability at the item level as well as demonstrate sensitivity to changes in treatment integrity over the course of program delivery. Study 2 was therefore designed to evaluate the preliminary reliability and construct validity of the BiCACS.

Participants

Child participants. In Year 3 of the development project, the BEST in CLASS program was implemented with 23 focal children in 11 state or federally funded classrooms serving high-risk children within one suburban and one urban district on the east coast (see above for procedure details). The 23 children (15 males and 8 females) in Study 2 averaged 3.95 years of age (*SD* = 0.38; range from 3 to 5). Child participants included 16 African American, 2 Caucasian, 1 Asian/Pacific Islander, and 1 Latino (data were not provided for 3 children). All children were enrolled in state or federally funded early childhood programs, qualified for free and reduced lunch, scored as "at risk" for future development of EBD as indicated by the ESP, and scored within the normal range on the BDI II.

Teacher participants. Of the 11 female teachers (3 had a bachelor's degree and 8 had a master's degree), 5 were African American, 5 were Caucasian, and 1 was Latina. They averaged 9.27 years of experience with preschool-aged children (*SD* = 10.31; range 0–34 years).

Coders. The primary coding team consisted of six coaches (five females). Five coaches were Caucasian and one was Latina. All coaches had completed their bachelor's degree, and four had completed a master's degree. Secondary coders used for reliability analyses consisted of four data collectors (three of whom were also coaches). Three secondary coders were Caucasian and one was Latina. All secondary coders had completed their bachelor's degree, and two had completed a master's degree.

BiCACS Scoring and Session Sampling Procedures

Live observations of the *BEST in CLASS* program were conducted weekly by coaches during their regular coaching sessions with teachers across three phases of the program (i.e., baseline, treatment implementation, maintenance). A total of 289 observations were conducted of which 54 (18.68%) were independently coded by a second observer for reliability purposes. During these observations, the coaches observed 15 min of instructional time during teacher-led activities and then completed the BiCACS. Coders' training on the BiCACS consisted of reading the scoring manual and a 2-hr didactic training session.

Measure for Validity Analyses

The Student Teacher Relationship Scale (*STRS*; Pianta & Hamre, 2001) assesses teacher perceptions of relationships with children and was used in Study 2 to assess the validity of the BiCACS. The STRS consists of 15 items measured on a 5-point Likert-type scale, where 1 represents *definitely does not apply* and 5 represents *definitely applies*. Two subscales, Closeness and Conflict, are derived from the STRS. For the current sample, the internal consistency for both factors was acceptable (Cronbach's α = .78 and .89 for closeness and conflict, respectively). The STRS has demonstrated validity with regard to predicting academic and social functioning in prekindergarten through the elementary grades (Hamre & Pianta, 2001; Pianta, La Paro, Payne, Cox, & Bradley, 2002) and has been used extensively in studies of preschool and elementary-age children (e.g., Birch & Ladd, 1997, 1998; Howes & Hamilton, 1992; Howes & Ritchie, 1999). The STRS has been validated with low-income and minority samples (Hamre & Pianta, 2001).

Results

To determine whether coaches could reliably code items on the BiCACS under typical training conditions, interrater reliability for data collected during live observations was calculated using ICCs (Shrout & Fleiss, 1979; see Tables 2 and 3). A coach and a secondary observer scored 18.68% of the sessions (*n* = 54) for reliability. Because a single coach coded all observations and reliability with another coder was calculated on a subset of the observations, the appropriate ICC estimate was single rater (McLeod, Islam, & Wheat, 2013). Interrater reliability for the Adherence items ranged from "fair" to "excellent" (ICCs ranged from .44 to .91 [*M* = .72, *SD* = .15], see Table 2), with four of the items in the "excellent" range, four in the "good" range, and two in the "fair" range. The Adherence subscale was highly reliable (ICC = .90). The interrater reliability for the Competence items ranged from "poor" to "excellent" (ICCs ranged from .39 to .85 [*M* = .64, *SD* = .16], see Table 3), with three of the items in the "excellent" range, four in the "good" range, two items in the "fair" range, and one item in the "poor" range (see Cicchetti, 1994). The Competence subscale was also highly reliable (ICC = .85). In sum, the interrater reliability for the items were slightly lower than

Table 4. Bivariate Correlations Among BiCACS Adherence and Competence Items.

Measure	1	2	3	4	5	6	7	8	9	10
1. Rules	__.74**__	.52**	.49**	.44**	.36**	.20**	.17**	.53**	.41**	.27**
2. Clear routines	.68**	__.82**__	.78**	.44**	.45**	.28**	.22**	.35**	.50**	.33**
3. Brisk instructional pace	.66**	.82**	__.87**__	.45**	.62**	.27**	.24**	.34**	.56**	.35**
4. Precorrection	.63**	.59**	.66**	__.60**__	.60**	.37**	.31**	.61**	.38**	.30**
5. Proximity control	.50**	.56**	.60**	.52**	__.69**__	.29**	.43**	.42**	.39**	.33**
6. Preacademic OTR	.41**	.56**	.57**	.39**	.41**	__.49**__	.63**	.30**	.02	.02
7. Social OTR	.45**	.57**	.65**	.42**	.52**	.81**	__.38**__	.29**	.08	.05
8. Behavior specific praise	.62**	.61**	.69**	.61**	.52**	.52**	.59**	__.55*__	.24**	.31**
9. Corrective feedback	.32**	.44**	.40**	.36**	.23**	.28**	.32**	.30**	__.64**__	.64**
10. Instructive feedback	.44**	.50**	.47**	.47**	.38**	.32**	.40**	.46**	.70**	__.60**__

Note. The top half of the triangle contains correlations among BiCACS Adherence items whereas the bottom half of the triangle contains correlations among BiCACS Competence items; the bolded diagonal contains correlations between BiCACS Adherence–Competence items. BiCACS = BEST in CLASS Adherence and Competence Scale; OTR = opportunities to respond.
*p < .05. **p < .01.

those reported in Study 1, but interrater reliability for most items remained in the acceptable range. The single rater ICC does produce a lower reliability estimate, compared with the average rater used in Study 1. Thus, these findings suggest that coaches can produce reliable ratings using the BiCACS under typical training conditions.

Next, we examined the amount of interitem overlap among the adherence and competence items (see Table 4). The interitem correlations among the adherence items were moderate to strong in strength and in the expected direction. This suggests that the teachers tended to use the interventions together and none of the items were redundant (r > .85). The interitem correlations among the competence items were also moderate to strong in strength and in the expected direction. In fact, the competence items evidenced stronger correlations than the adherence items, though none of the items were redundant. The correlations between the adherence and competence items were moderate (Social Opportunities to Respond, r = .38, p < .001) to strong (Clear Routines, r = .82, p < .001). The strength of these correlations are similar to amount of overlap observed between the Adherence and Competence subscales (r = .71, p < .001). The correlations between corresponding Adherence and Competence items were generally stronger than the correlations between noncorresponding items on the Adherence and Competence subscales, which suggests that the adherence and competence ratings covaried. This finding is consistent with the BEST in CLASS program, which aims to increase the dosage and quality of the specific interventions. In sum, the interitem correlations were in the expected direction and generally support the construct validity of the items and subscales. The findings do, however, suggest that there was moderate to strong overlap among the BiCACS Adherence and Competence items and subscales.

The relation of the BiCACS Adherence and Competence subscales to standardized measures of the student–teacher relationship are also relevant to the discriminant validity of the subscales. We used the STRS Closeness (STRS-CL) and Conflict (STRS-CO) scales to represent the student–teacher relationship. The Adherence subscale evidenced a strong positive correlation to the STRS-CL (r = .51, p = .026) and a small negative correlation to the STRS-CO (r = −.13, p = .578). The Competence subscale demonstrated a similar pattern; the Competence subscale had a moderate positive relationship with the STRS-CL (r = .43, p = .065), and a small negative relationship with the STRS-CO (r = −.18, p = .474). These correlations are in the expected direction, consistent with past treatment integrity research, and support the discriminant validity of the Adherence and Competence subscales (Carroll et al., 2000; Hogue et al., 2008).

Finally, we examined the construct validity of the BiCACS Adherence and Competence subscales. As construct validity cannot be assessed directly, an indirect approach was used to determine whether the measure could identify expected differences within groups (Lambert & Hill, 1994). We evaluated whether the Adherence and Competence subscale scores could distinguish between different phases of the BEST in CLASS program. It was expected that Adherence and Competence would increase over time due to the sequential introduction of modules with weekly performance-based coaching. For this analysis, treatment was characterized as being comprised of four phases. The first phase represented all baseline data collection (3 weeks), the second represented the first half of treatment (Weeks 1−7 of the program), the third represented the second half of treatment (Weeks 8−14 of the program), and the last time point represented maintenance data collected 1 month after the end of the program (3 weeks). For the Adherence subscale, our analyses indicated the scores varied across the four phases, F(3, 72) = 10.03, p < .001. Independent-sample t tests showed the following:

Table 5. BiCACS Adherence and Competence Subscale Scores Across Treatment Phases.

	Phase 1	Phase 2	Phase 3	Phase 4
Subscale	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Adherence	3.78 (1.01)	4.55 (0.75)	5.21 (0.71)	4.91 (0.90)
Competence	5.19 (0.97)	5.08 (0.72)	5.42 (0.70)	5.37 (0.89)

Note. BiCACS = BEST in CLASS Adherence and Competence Scale.

Adherence ratings were higher at Phase 3 than Phase 2 (*M*s = 5.21 and 4.55, *SD*s = 0.71 and 0.71, $t(36) = 2.79$, $p = .008$) and Phase 1(*M*s = 5.21 and 3.78, *SD*s = 0.71 and 1.01, $t(36) = 5.06$, $p < .001$); Adherence ratings were higher at Phase 4 than Phase 1 (*M*s = 4.91 and 3.78, *SDs* = 0.90 and 1.01, $t(36) = 3.64$, $p < .001$); Adherence ratings were higher at Phase 2 than Phase 1 (*M*s = 4.55 and 3.78, *SD*s = 0.75 and 1.01, $t(36) = 2.67$, $p = .007$). For the Competence subscale, our test indicated the scores did not vary across the treatment phases, $F(3, 72) = 0.71$, $p = .547$ (see Table 5). These findings suggest that the scores on the Adherence subscale are sensitive across treatment phases in a direction that would be expected given the sequential introduction of the components of the BEST in CLASS program; the scores on the Competence subscale did not indicate the same level of sensitivity.

Discussion

The purpose of this study was to describe the development of the BiCACS and report preliminary psychometric data for the measure. The analyses provide evidence to support the initial reliability and validity of the BiCACS. The BiCACS items and subscales demonstrated fair to strong reliability. The findings also generally supported the validity of the BiCACS: The pattern of correlations among the items and subscales was in the expected direction and the subscales were distinct from a teacher-report measure of the child–teacher relationship. The analyses also indicated that the Adherence subscale is sensitive to changes in adherence over the course of the BEST in CLASS program. Thus, the BiCACS appears to be a promising treatment integrity measure that may contribute to efforts designed to evaluate and implement the BEST in CLASS program in early childhood classrooms.

Results from the current study are an important first step in establishing the psychometric properties of the BiCACS. Our findings suggest that trained coders and coaches can achieve adequate reliability at the item level with the BiCACS across videotaped and live observations. While the reliability of integrity measures for teacher-delivered programs targeting young children has not been reported, the interrater reliability of the BiCACS was comparable with the reliability at the item (e.g., Barber, Mercer, Krakauer, & Calvo, 1996; Hogue et al., 2008) and subscale (e.g., Carroll et al., 2000) level for treatment integrity measures in the psychotherapy field. Importantly, we evaluated the reliability of the BiCACS across different conditions (videotaped and live observations) and observer types (trained coders and coaches). This means that the BiCACS may be flexible enough to be reliably used by different types of observers, which bodes well for its use in efficacy and effectiveness research.

Compared with the adherence items and subscale, the interrater reliability for the competence items and subscales was lower. This is consistent with previous findings (e.g., Hogue et al., 2008), suggesting that competence may be harder to code than adherence to a treatment protocol. Although few studies have attempted to measure competence in the prevention field (see Hagermoser Sanetti et al., 2012; Webb, DeRubeis, & Barber, 2010), researchers have begun to recognize that competence may have an important relation to treatment outcomes (Harn, Parisi, & Stoolmiller, 2013). There is considerable debate within the field about what level of training is needed for coders to rate competence (e.g., Southam-Gerow & McLeod, 2013; Waltz, Addis, Koerner, & Jacobson, 1993). Some assert that model experts should code competence whereas others argue that trained graduate students are capable of coding competence. Even though undergraduate and graduate students and trained coaches were able to reliably code competence in the present study, it is possible that model experts may have obtained higher interrater reliability estimates. Thus, more empirical work is needed to determine what level of training is needed to generate reliable and valid competence ratings.

However, it is important to note that these reliability estimates should be considered conservative. We used a measure designed to code the integrity of the BEST in CLASS program to rate observations of teachers only delivering this program. This within-condition approach can limit variability in intervention delivery and thus represents a conservative approach to estimating reliability (Startup & Shapiro, 1993). Taken together, the results reported across both studies suggest that the BiCACS demonstrates adequate reliability.

Our findings suggest that trained coders and coaches can reliably use the BiCACS to code videotaped and live observations, which has important implications for future research applications of the measure. These data indicate that the BiCACS can be used by (a) trained coders as a manipulation check to aid interpretation of findings, and (b) coaches as a tool for informing teacher training efforts. Reliability at the item level makes it possible for researchers to investigate process–outcome relations at the intervention level, which could help identify the core ingredients of the BEST in CLASS program. Researchers have highlighted the need for reliable measures of integrity to advance the science of prevention (Durlak, 2010; Hagermoser Sanetti & Kratochwill, 2009; Wolery, 2011), and the reliability data at the subscale and item level for the BiCACS are a promising step in this direction.

Our findings also provide preliminary support for the validity of the BiCACS items and subscales. At the item level, the associations among the Adherence and Competence items were all in the expected direction, supporting the construct validity of the items. At the subscale level, the Adherence and Competence subscales demonstrated overlap with a measure of the quality of the child–teacher relationship in the expected direction (Hogue et al., 2008). Moreover, the magnitude of the relationship between the subscales and the affective scale was consistent with past research investigating this association (Carroll et al., 2000). These findings therefore provide preliminary support for the discriminant validity of the BiCACS subscales.

We also investigated the amount of overlap among the Adherence and Competence items and subscales. Across the two studies, the Adherence and Competence items and subscales evidenced moderate to strong overlap. These findings are consistent with past studies that have examined these associations at the item (Hogue et al., 2008) and subscale level (Carroll et al., 2000). Moreover, these findings suggest that the adherence and competence of delivery of the BEST in CLASS interventions were associated in the expected direction.

However, it is important to note that theoretical and empirical work has not clarified the amount of overlap between the adherence and competence integrity components (see Barber et al., 2007). The degree of overlap between adherence and competence in the psychotherapy field has ranged from moderate (*r*s = .31 to .40; Carroll et al., 2000) to high (*r*s = .77 to .90; Barber et al., 1996). Moreover, the relationship between these components in school-based prevention work remains unclear as well (Hagermoser Sanetti & Kratochwill, 2009; Schulte et al., 2009). The Adherence and Competence subscales evidenced moderate to strong overlap, which is to be expected given the focus of BEST in CLASS on increasing the frequency of component delivery as well as the quality of implementation. Of course, we cannot rule out the potential impact response bias may have had on our findings; however, our findings do suggest that the strength in the relation may vary across observer type.

Our findings also indicate that the Adherence subscale may be capable of measuring variability in treatment implementation. Scores on the Adherence subscale increased across time, which is to be expected given the sequential introduction of coaching associated with learning modules, providing some preliminary evidence that the subscale can measure variability in treatment adherence. However, the Competence subscale did not demonstrate significant increases across time.

Our findings supporting the validity of the BiCACS subscales add to the literature in a few important ways. Few studies have examined the validity of treatment integrity measures, so the current findings represent an initial step to address this issue. Moreover, the ability to assess variability is particularly important when interpreting findings from a randomized controlled trial. Simply put, it is important to determine whether treatment adherence varies across time, teachers, or schools as this may account for differences in outcomes (Hagermoser Sanetti, Gritter, & Dobey, 2011; Wolery, 2011). Being able to identify variability in treatment integrity can also inform teacher training efforts, especially when the integrity measure is used by coaches as part of a training system. Therefore, our findings help demonstrate that observational integrity measures used in school settings can be sensitive to changes in treatment adherence over time, which means that the measures may have some utility in implementation research.

Although our study has a number of strengths, there are a few limitations that should be kept in mind as readers interpret the results. First, in addition to having a small sample, our videotaped sample was not randomly selected as the collection of these videotapes was part of the program development process. While we have no reason to believe that our videotaped sample was not representative of implementation of BEST in CLASS by all teachers in Years 2 and 3 of the development project, readers should take this into consideration as they interpret our findings. Second, the data reported in this article are from an integrity measure developed to assess adherence and competence for BEST in CLASS and therefore are not generalizable to other programs. That said, the process used to develop the BiCACS as well as the preliminary findings may be of use to researchers developing integrity measures for their own early intervention programs that target reduction in problem behavior of young, high-risk children. Third, Study 2 had secondary observations conducted on approximately 19% of the total observations, and this is lower than the minimally recommended 20% of total observation sessions (Kennedy, 2005). Finally, in the BEST in CLASS development project we did not have comparison classrooms. Therefore, we were not able to establish the criterion validity of the BiCACS or provide meaning to the BiCACS scores by comparing them with levels in business-as-usual classrooms.

In sum, as the field increasingly moves toward effectiveness research, integrity measures are needed to guide interpretation of findings, thereby helping to identify implementation and evaluation barriers to transporting EBPs to new settings, as well as identifying active ingredients of EBPs. The development of the BiCACS is a promising first step in this direction, and hopefully will serve as a stepping stone to further development work in integrity research in early childhood classrooms as well as provide a blueprint for other researchers conducting prevention research in classroom settings.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A080074 to Virginia Commonwealth University as well as an award from the Presidential Research Initiation Program at Virginia Commonwealth University.

References

Barber, J. P., Mercer, D., Krakauer, I., & Calvo, N. (1996). Development of adherence/competence rating scale for individual drug counseling. *Drug and Alcohol Dependence*, *43*, 125–132. doi:10.1016/S0376-8716(96)01305-1

Barber, J. P., Sharpless, B. A., Klostermann, S., & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. Professional Psychology: *Research and Practice*, *38*, 493–500. doi:10.1037/0735-7028.38.5.493

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. G., Welsh, J. A., Greenberg, M. T., & Gill, S. (2008). Promoting academic and social-emotional school readiness: The head start REDI program. *Child Development*, *79*, 1802–1817. doi:0009-3920/2008/7906-0019

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

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

Carroll, K. M., Nich, C., Sifty, R. L., Nuro, K. F., Frankfurter, T. L., Ball, S. A., & Rounsaville, B. J. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research. *Drug and Alcohol Dependence*, *57*, 225–238. doi:10.1016/S0376-8716(99)00049-6

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*, 396–406. doi:10.1093/clipsy/9.4.396

Cate, D., Diefendorf, M., McCullough, K., Peters, M. L., & Whaley, K. (Eds.). (2010). *Quality indicators of inclusive early childhood programs/practices: A compilation of selected resources*. Chapel Hill: The University of North Carolina, FPG Child Development Institute, National Early Childhood Technical Assistance Center.

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

Conroy, M. A., Sutherland, K. S., Vo, A. K., Carr, S. E., & Ogston, P. (2013). Early childhood teachers' use of effective instructional practices and the collateral effects on young children's behavior. *Journal of Positive Behavior Interventions*. Advance online publication. doi:10.1177/1098300713478666

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*, 67–91. doi:10.1007/s10935-007-0081-0

Domitrovich, C. E., Moore, J. E., & Greenberg, M. T. (2012). Maximizing the effectiveness of social-emotional interventions for young children through high-quality implementation of evidence-based interventions. In B. Kelly & D. F. Perkins (Eds.), *Handbook of implementation science for psychology in education* (pp. 207–229). Cambridge, UK: Cambridge University Press.

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*, 38–64. doi:10.1080/10409280802657449

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*, 348–357. doi:10.1016/j.ecresq.2010.03.003

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

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

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009). Treatment integrity assessment in the schools: An evaluation of the treatment integrity planning protocol. *School Psychology Quarterly*, *24*, 24–35. doi:10.1037/a0015431

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

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, *79*, 181–193.

Hemmeter, M. L., Corso, R., & Cheatham, G. (2006, February). *Issues in addressing challenging behaviors in young children: A national survey of early childhood educators*. Paper presented at the Conference on Research Innovations in Early Intervention, San Diego, CA.

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, *76*, 544–555. doi:10.1037/0022-006X.60.1.73

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

Howes, C., & Hamilton, C. E. (1992). Children's relationships with caregivers. *Child Development*, *63*, 859–866. doi:10.2307/1131238

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

Kennedy, C. H. (2005). *Single case designs for educational research*. Boston, MA: Allyn & Bacon.

Lambert, M., & Hill, C. (1994). Assessing psychotherapy outcomes and processes. In A. Bergin & S. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 72–113). New York, NY: John Wiley & Sons.

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

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.

McLeod, B. D., Southam-Gerow, M. A., Bair, C. E., Rodriguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychological treatment quality indicator. *Clinical Psychology: Science and Practice*, *20*, 14–32. doi:10-1111/cpsp.12020

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*, 436–443. doi:10.1080/15374411003691750

Newborg, J. (2005). *Battelle Developmental Inventory, 2nd edition. Examiner's manual*. Rolling Meadows, IL: Riverside.

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

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

Quesenberry, A. C., Hemmeter, M. L., & Ostrosky, M. M. (2011). Addressing challenging behaviors in head start: A closer look at program policies and procedures. *Topics in Early Childhood Special Education*, *30*, 209–220. doi:10.1177/0271121410371985

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*, 460–475.

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

Southam-Gerow, M. A., & McLeod, B. D. (2013). Advances in applying treatment integrity research for dissemination and implementation science: Introduction to special issue. *Clinical Psychology: Science and Practice*, *20*, 1–13. doi:10-1111/cpsp.12019

Startup, M., & Shapiro, D. A. (1993). Therapist treatment fidelity in prescriptive vs. exploratory psychotherapy. *British Journal of Clinical Psychology*, *32*, 443–456.

Sutherland, K. S., Conroy, M., Abrams, L., & Vo, A. (2010). Improving interactions between teachers and young children with problem behavior: A strengths-based approach. *Exceptionality*, *18*, 70–81. doi:10.1080/09362831003673101

Sutherland, K. S., & McLeod, B. D. (2010). *Treatment adherence and competence measure for classroom based interventions*. Richmond: Virginia Commonwealth University.

Thompson, T., Felce, D., & Symons, F. J (Eds.). (2000). *Behavioral observation: Technology and applications in developmental disabilities*. Baltimore, MD: Brookes Publishing.

Vo, A. K., Sutherland, K. S., & Conroy, M. A. (2012). Best in class: A classroom-based model for ameliorating problem behavior in early childhood settings. *Psychology in the Schools*, *49*(5), 402–415.doi:10.1002/pits.21609

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

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. E. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*, 620–630. doi:10.1037/0022-006X.61.4.620

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78*, 200–211. doi:10.1037/a0018912

Webster-Stratton, C., Reid, J. M., & 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*, 105–124. doi:10.1207/S15374424JCCP3301_11

Wolery, M. (2011). Intervention research: The importance of fidelity measurement. *Topics in Early Childhood Special Education*, *31*, 155–157. doi:10.1177/0271121411408621