American Educational Research Journal
December 2017, Vol. 54, No. 6, pp. 1378–1413
DOI: 10.3102/0002831217726522
© 2017 AERA. http://aerj.aera.net

Validating a Fidelity Scale to Understand Intervention Effects in Classroom-Based Studies

Pamela Buckley
University of Colorado Boulder
Brooke Moore
Fort Hays State University
Alison G. Boardman
University of Colorado Boulder
Diana J. Arya
Andrew Maul
University of California, Santa Barbara

K–12 intervention studies often include fidelity of implementation (FOI) as a mediating variable, though most do not report the validity of fidelity measures. This article discusses the critical need for validated FOI scales. To illustrate our point, we describe the development and validation of the

Pamela Buckley, PhD, is a research associate with the Institute of Behavioral Sciences, Blueprints for Healthy Youth Development and Crime Prevention at the University of Colorado Boulder, 483 UCB, Boulder, CO 80309; e-mail: *Pamela.Buckley@colorado.edu*. Her expertise includes intervention research, research design, program evaluation and quantitative data analysis.

Brooke Moore, PhD, is an assistant professor of special education at Fort Hays State University. Her research focuses on helping educators create equitable and inclusive learning environments for all students.

Alison G. Boardman, PhD, is an associate professor of Education at the University of Colorado Boulder. Her research focuses on effective professional development models, reading comprehension instruction, and project-based learning in classrooms with emergent bilingual learners and students with disabilities.

Diana J. Arya, PhD, is an assistant professor of Education at the University of California, Santa Barbara. Her research interests focus on language and literacy practices of K–12 and postsecondary populations within science and engineering contexts.

Andrew Maul, PhD, is an assistant professor of Education at the University of California, Santa Barbara. His scholarship broadly focuses on research methods and the philosophy of social science, and in particular on the theory and practice of educational and psychological measurement.

Implementation Validity Checklist (IVC-R), an observation tool for measuring FOI of a research-based instructional reading approach, Collaborative Strategic Reading. Following Kane (2006), Wilson (2004), and the guidelines of the Standards for Educational and Psychological Testing (Standards: AERA, APA, & NCME, 2014), findings suggest the IVC-R is a valid instrument for measuring fidelity to CSR. We hope this process will provide an informative model for the validation of FOI observation tools in future classroom-based efficacy studies.

Keywords: Fidelity, Fidelity of implementation, Measurement of program fidelity, Observation protocols, Program integrity

Introduction

An important purpose of education research is to evaluate whether programs are effective in supporting student learning and improving achievement, though intervention effects cannot be fully explained without evidence of implementation success or failure (Munter, Wilhelm, Cobb, & Cordray, 2014). Examining fidelity of implementation (FOI), or the degree to which an intervention is implemented as intended (Dane & Schneider, 1998; O'Donnell, 2008), is necessary for understanding the relationship between components of an intervention and program outcomes (Munter et al., 2014). While the randomized controlled trial (RCT) is widely regarded as a valuable method for estimating average treatment effects of educational interventions (Rosenbaum, 1995), such RCTs cannot account for differences between the design of an intervention and its actual implementation, thus resulting in a lack of clarity about the effectiveness of the intervention model (Hulleman & Cordray, 2009). Understanding variance in implementation, however, is not the only reason to measure FOI. Cordray & Pion (2006) contend that assessing FOI also helps with: "(a) specifying the amount or 'magnitude' of the treatment to be delivered, (b) understanding the coherence of the intervention or treatment, (c) enhancing the validity of inference that can be derived from studies of treatment effectiveness, and (d) optimizing the use of the results to guide interventions" (p. 105). Results drawn from valid measures of FOI can therefore give researchers assurance in attributing outcomes to the intervention and practitioners' confidence in implementing the chosen intervention as it was intended. Accurate measurement of FOI also reveals important considerations for the field regarding how to translate evidence-based findings into practice.

Though the importance of rigorous measurement of implementation fidelity is well documented in the literature, most educational intervention studies nevertheless provide limited information on their FOI measures and implementation results (Nelson, 2013; O'Donnell, 2008). Lack of time and financial resources to develop and describe sound FOI measurement practices is likely a driving force. Indeed, we are unable to locate any education intervention studies (to date) that include a full account of the development and validation of fidelity measures. We therefore sought guidance from the instrument development literature on creating high-quality educational measures (e.g., Kane, 2006; Wilson, 2004) in adapting the validation process for a new domain—fidelity. This application is necessary in that it provides a model for developing valid FOI measures that should make the process more efficient, less costly, and thus more feasible, which ultimately will improve understanding of treatment effectiveness.

While there are multiple means for collecting fidelity data (e.g., surveys, interviews, focus groups, analysis of student artifacts), Heck, Steigelbauer, Hall, & Loucks (1981) argue that classroom observation best captures the interactive process of teacher-student exchanges and thus sharpens the picture of the actual implementation practices used by teachers. In the present study, we describe the development and validation of a revised version of the Implementation Validity Checklist (IVC) classroom observation tool designed to measure the FOI of a research-based model of reading strategies instruction, Collaborative Strategic Reading (CSR; Klingner, Vaughn, Boardman, & Swanson, 2012). We discuss the IVC to provide a case study illustrating how the validation can be applied to the development of sound FOI observation measures. We do this by first briefly reviewing contemporary scholarship on the validation of measures (AERA, APA, NCME, 2014; Kane, 2006). Next we list the criteria for measuring FOI frequently cited across different disciplines and how these criteria relate to classroom-based K–12 intervention studies in general, and our case study in particular. We then describe the lack of research in the area of FOI measurement—particularly in the education literature—and the difficulties in validating fidelity assessments that may explain this shortcoming, followed by examples of FOI measurement from the implementation science literature that offer both similar and distinct processes in accomplishing goals similar to ours to inform how the design of these measures align with components and principles of FOI.

Frameworks for the Validation of Assessment Procedures

Drawing from both the implementation science and prevention research literature, Schoenwald and colleagues (2011) recommend that when developing an observational measure to assess FOI, researchers should refer to the Standards for Educational and Psychological Testing (*Standards*;

AERA, APA, and NCME, 2014), as they offer “a unifying element for any test development or evaluation efforts” (p. 36) by providing elaboration on the types of evidence relevant to the evaluation of the validity of an assessment procedure. In particular, the *Standards* discuss the value of five strands of validity evidence, based on (1) test content (e.g., expert reviews of how well the content of the assessment represents its intended purpose, and the relevance, clarity, and importance of each item on the assessment); (2) response processes (e.g., how raters collect and interpret data); (3) internal structure (including, for example, evaluation of the psychometric properties of individual items and the assessment procedure as a whole); (4) relations to other variables (e.g., correlational or experimental studies examining the extent to which scores obtained with an instrument accurately predict outcome variables); and (5) the consequences of testing (e.g., studies of whether anticipated benefits and/or unanticipated negative consequences of testing are realized). Studies that use researcher-developed observation tools to assess FOI typically give little evidence of validity beyond relations to specific external variables and/or statistics related to interrater reliability and internal consistency (e.g., Dane & Schneider, 1998), thereby limiting how the validation process described by scholars such as Kane (2006) and Wilson (2004) applies to fidelity measurement.

Contemporary argument-based frameworks for validation (e.g., Kane, 2006) emphasize the need to begin the validation process with a clear statement of the proposed uses and interpretations of test scores, and then to construct an evidence-based argument to defend the adequacy and appropriateness of the test for its intended uses. For example, in the present case, the claim is that scores on the IVC are interpreted as measurements of fidelity to the CSR instructional model and can be used in any research setting in which understanding FOI of CSR is of interest. Although the argument-based approach to validation articulated by Kane and codified in the *Standards* are perhaps most commonly associated with traditional educational testing, they may apply in principle to any situation in which an instrument (such as an observational protocol) is used to make claims that stretch beyond that which is immediately observed. Measures of fidelity such as the IVC-R use direct observational evidence to support inferences about a broader attribute of teachers and classrooms, and as such can (and, we believe, should) be vetted using the same rigorous standards as are found elsewhere in educational and psychological testing and measurement.

Applying such validation processes to our case study, we ask the following four research questions: Does the revised version of the IVC show evidence of validity based on (1) test content; (2) rater response processes; (3) internal structure; and (4) relations to other variables? Research questions related to the consequences of testing are not of primary interest in FOI research since whether or not an intervention is implemented as intended is typically not used to make consequential (e.g., promotional) decisions about individual teachers, such as with our case study.

Criteria for Assessing Fidelity of Implementation (FOI)

Fidelity of implementation is typically defined in terms of *dosage* (the frequency of program delivery), *adherence* (whether program components are delivered as prescribed), *quality of delivery* (how well the program material is implemented), *participant responsiveness* (how well the instruction is received or perceived), and/or *program differentiation* (the degree of contrast between treatment and control strategies and/or activities; Dane & Schneider, 1998; Mowbray, Holter, Teague, & Bybee, 2003). In classroom-based K–12 intervention studies, FOI is often considered in terms of instructional quality, commonly operationalized as the amount of improvement or change that occurs in a teacher's practice, though overall fidelity is often considered synonymous with adherence and integrity (O'Donnell, 2008). Initially, fidelity to CSR (Klingner et al., 2012) was assessed via the IVC (e.g., Boardman et al., 2016a, 2016b; Boardman, Klingner, Buckley, Annamma, & Lasser, 2015; Vaughn et al., 2013) designed to measure both procedural fidelity and quality of implementation, though little evidence was available for its validity. Our case study illustrates how the FOI tool was revised with a focus on *adherence* (i.e., implementation as inscribed) to the CSR model by following recommendations from the *Standards* (AERA, APA, and NCME, 2014) and applying Kane's (2006) contemporary argument-based framework for validation of an assessment procedure.

Difficulties of Measuring FOI

Though FOI is important to unpack because it is a complex phenomenon, it can be very difficult to measure. There is no perfect process for accurately measuring FOI (Schoenwald et al., 2011). Sound measurement practices informing the development and use of practical FOI observational tools, however, will ensure that an intervention's delivery is accurately assessed. The need to further understand FOI has been identified across several fields, including the case management (Drake & Resnick, 1998), healthcare (Greenhalgh, Robert, MacFarlane, Bate, & Kyriakidou, 2004), education (O'Donnell, 2008), and implementation science (Carroll et al., 2007) intervention literature. This gap in understanding, according to many scholars, encompasses how to (1) define the FOI concepts to be measured; and (2) develop empirically-validated measures that assess FOI for different interventions. FOI measures must also be user-friendly and appropriate for different contexts; when linked with outcomes, they become "an important tool in building evidence-based practices" (George & Childs, 2012, p. 197).

While there is a need to understand how instructional models are implemented in classrooms, claims about FOI based on observations of instruction can be problematic for a number of reasons. O'Donnell (2008) described how many difficult-to-measure features of instruction outside the prescribed intervention may influence student learning. Examples include the extent to which "good teaching" (in general) interacts with implementation of the (specific) instructional model under investigation (Shulman, 1990), and the impact of adaptations to the prescribed model or materials (Rogers, 2003). Given this challenge, Mowbray et al. (2003) noted the considerable effort it takes to develop protocols that result in consistent, accurate, and interpretable observations. Despite researchers' best efforts, results are commonly influenced by idiosyncratic rater variance "even if response scales are well anchored" (Mowbray et al., 2003, p. 330). In addition, measuring FOI involves translating observed behaviors into numeric form (Schoenwald et al., 2011). Two issues are therefore of critical importance: (1) accurately coding reported operations in alignment with the intervention components; and (2) effectively transforming the ordinal or categorical ratings into interval scales from which scores are derived to interpret FOI. Observational coding systems with trained raters have effectively been used in many efficacy trials because they have the potential to provide objective and highly specific information about a program's implementation (Mowbray et al., 2003). Developing methods for rating observations, however, is challenging given the substantial time and expense that goes into designing protocols, hiring and training raters, scoring observations, reviewing scores, and recording, analyzing and disseminating data. Throughout this process, ensuring adherence of raters to coding protocols and high interrater reliability is crucial, and scores must be appropriately analyzed and accurately reported (Schoenwald et al., 2011). The aforementioned difficulties may explain why the validation process for fidelity measures is generally overlooked or underemphasized in education intervention studies, and why clearly communicated methods and procedures are needed to help make this process more efficient, less costly, and thus more practical to adopt.

Process of FOI Validation in Related Fields

Though minimal research has evaluated ways to marry effective and efficient fidelity measurement (Schoenwald et al., 2011), there are some studies in the prevention science research base (which is tightly linked with the implementation science literature) that describe the process for developing a scientifically validated FOI measure. Below are two examples, each of which provides both similar and distinct processes to ours in accomplishing a common goal of improving the interpretation of program outcomes in the context of experimental studies.

School-Wide Positive Behavior Interventions and Support (SWPBIS) provides a systems approach to promoting a social culture with behavioral supports needed for students to experience social and academic success

(Lewis & Sugai, 1999). Research conducted over the past 30 years has shown SWPBIS to be effective in decreasing school-wide behavior problems, improving academic achievement, and creating a positive school climate (Bradshaw, Mitchell, & Leaf, 2010; Horner, Sugai, & Anderson, 2010). Two empirically validated FOI measures of SWPBIS include (1) School-Wide Evaluation Tool (SET; Horner et al., 2004) and (2) Benchmarks of Quality (BoQ; Cohen, Kincaid & Childs, 2007).

The SET is administered by a trained external evaluator and focuses on initial implementation activities of SWPBIS. As per Messick (1988), the measurement theorist who maintained that inferences generated from test scores often require different types of evidence but not different validities, researchers first created a conceptual logic to (1) guide the structure and intended use of the SET; and (2) provide a framework for the ongoing assessment of the validity of the instrument. The content, item format, and scoring of the SET took place over 3 years and involved the teachers and administrators of 150 elementary and middle schools. Trained observers conducted 1–2 hours of interviews at each school with administrators, teachers, staff members, and students. Raters also reviewed archival documents such as school policies, training curricula, and SWPBIS meeting minutes, and they examined systems used to collect and store SWPBIS data. Multiple analyses conducted to assess the psychometric properties of the SET show that SET scores demonstrate adequate central measures of tendency and variability for sensitivity at the item, subscale, and total levels. In addition, internal consistency and interrater observer agreement were both high. Findings also show the SET has adequate test-retest reliability, yields a valid index of SWPBIS as defined by Lewis and Sugai (1999), and can document change in implementation levels (Freeman et al., 2016).

The BoQ is a self-report FOI measure of SWPBIS (Childs, Kincaid, George, & Gage, 2016; Cohen et al., 2007). Similar to the iterative procedure for revising the IVC FOI measure of Collaborating Strategic Reading described in our case study and as per Messick (1988) who pointed out that instrument development and validation is an ongoing process, but specifically following guidelines described by McKennel (1974), the BoQ was developed in three stages: (1) a qualitative pilot to develop the instrument content; (2) another pilot to test the scale structure; and (3) development of the main survey derived from a conceptual network that includes assessing the reliability and validity of the instrument (Cohen et al., 2007). Items on the BoQ were developed from a training manual that is based on the critical elements of SWPBIS (Lewis & Sugai, 1999). The protocol for the scoring guide was generated from the implementation goals spelled out in the SWPBIS training manual. Once the items were generated, 20 observers (i.e., trainers from across several states who are experts in Positive Behavioral Support) rated the importance of each item to the PBS process on a scale from 1 (minimally important) to 3 (critically important). These ratings were then used to establish point values for each item (Cohen et al., 2007). Next, cognitive interviews were conducted to investigate sources of response error in individual survey items (Schechter, Blair, & Hey, 1996). The BoQ was piloted in Florida with 10 SWPBS coaches and teams, and feedback was provided on unclear items or directions. Additional revisions to the instrument were made using these qualitative data (Cohen et al., 2007). Multiple studies validating the BoQ provide evidence of strong internal consistency, interrater reliability, and test-retest reliability (e.g., Cohen et al., 2007; Horner et al., 2004). In addition, Horner et al. (2004) showed moderate concurrent validity between the BoQ and the SET. Meanwhile, Cohen et al. (2007) conducted another assessment of concurrent validity using data from 720 schools completing both the SET and the BoQ with results showing a significant relationship.

Unlike the IVC described in our case study, research describing the BoQ does not specify which FOI construct (i.e., adherence, dosage, quality, participant responsiveness, or program differentiation) the instruments measure. Later studies (e.g., Bradshaw et al., 2010) claim the instruments measure implementation quality. The SET and BoQ have been used in several studies to assess the relationship between fidelity and outcomes (Childs et al., 2016; Freeman et al., 2016)—a chief rationale for validating the revised version of the IVC. Our case study therefore builds on studies that follow sound procedures for evaluating the reliability and validity of FOI measures (e.g., Kazdin & Kendall, 1998; Schoenwald et al., 2011). While we could find no education studies that provide much detail in validating FOI measures, we were able to draw ideas from a few (e.g., Hamre, Pianta, Mashburn, & Downer, 2007; Piburn & Sawada, 2000; Sawada et al., 2002) in following the guidelines offered by Schoenwald et al. (2011) and Kane (2006) to describe the iterative process for developing and validating a classroom observation tool that includes expert and trained raters and a standardized scoring system for assessing FOI.

Case Study: Collaborative Strategic Reading

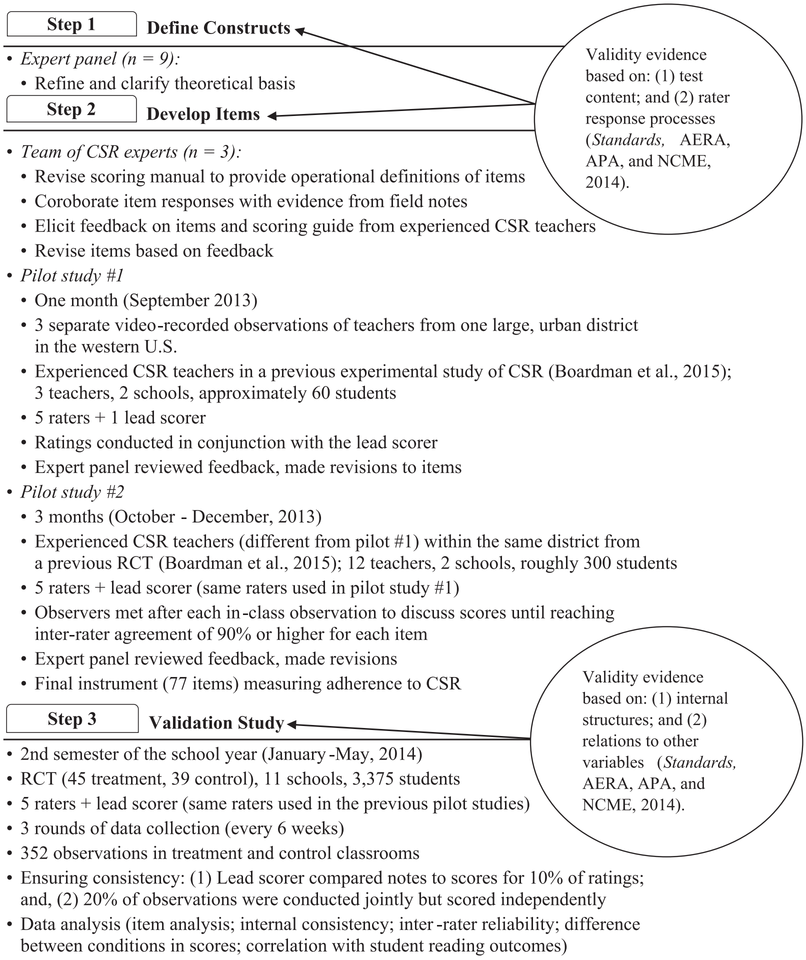
Collaborative Strategic Reading (CSR; Klingner et al., 2012), the instructional model described in our case study, provides students with metacognitive knowledge and self-regulation skills needed to successfully read complex content-related texts. Before reading, students begin with the “Preview” strategy (teacher introduces text and students brainstorm and set a purpose for reading). During reading, students work together in small, heterogeneous, collaborative groups to “Click and Clunk” (monitor understanding by identifying and figuring out unknown words and ideas) and “Get the Gist” (determine main ideas of designated sections of texts). After reading, students engage in the “Question” (asking/answering questions about the reading) and “Review” strategies (summarizing and justifying key ideas). Finally, the teacher brings the class back together for a whole-class wrap-up.

Fidelity and CSR

In previous studies, CSR has benefitted a variety of learners (Boardman et al., 2015; Klingner, Vaughn, & Schumm, 1998; Vaughn et al., 2011) with especially beneficial results for struggling readers in mixed-ability classrooms (Boardman et al., 2016a, 2016b; Kim et al., 2006; Klingner, Vaughn, Arguelles, Hughes, & Leftwich, 2004). In several of these studies, fidelity was assessed the IVC (e.g., Boardman et al., 2015, 2016a, 2016b; Vaughn et al., 2013). This previous version of the IVC contained items scored on a five-point rating scale (e.g., not observed, low quality, mid-low, mid-high, very high quality), with each of these items combining teacher and student behaviors. Although the IVC exhibited acceptable levels of interrater agreement (around 90%; Boardman et al., 2015, 2016a, 2016b; Vaughn et al., 2013) and internal consistency (.80 –.91.; Boardman et al., 2016a), like many FOI assessments, items asked for holistic judgments and thus required a considerable amount of subjectivity. Additionally, there was little evidence of validity beyond the reliability checks referenced above. Finally, observers have noted forms of variation in the quality of CSR implementation, as well as features that also frequently occur in non-CSR classrooms (e.g., main idea strategies)— observations that helped clarify what is unique to CSR. For these reasons, a new version of the IVC was sought, with items specifically linked to concrete observable behaviors and with robust checks on the validity of the assessment. The following hypotheses were explored as assumptions of CSR that would be demonstrated through a validated IVC-R instrument: (1) Student and teacher behaviors embedded in the CSR model can be consistently observed across raters, classrooms, and schools; (2) CSR implementation will necessarily vary across teachers; (3) CSR lessons are distinct from business-as-usual lessons, even though some practices might be shared across lesson types (e.g., student collaboration); and (4) higher quality of CSR implementation will be associated with higher student reading outcomes.

Methods

Guidelines that emphasize the importance of examining fidelity (Mowbray et al., 2003; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012; O'Donnell, 2008) as well as work on developing classroom observation instruments (Fish & Dane, 2000; Hamre et al., 2007; Hill et al., 2012; Pianta & Hamre, 2009) were drawn from in developing the IVC-Revised (or IVC-R). Additionally, guidance was sought from the general literature on creating high-quality educational measures (e.g., Kane, 2006) and empirically validated FOI assessments (e.g., Cohen et al., 2007; Horner et al., 2004; Schoenwald et al., 2011). Roughly following the procedural guidelines established by the Berkeley Evaluation and Assessment Research (BEAR) Assessment System (Wilson, 2004), the steps for revising the IVC-R consisted of: (1) defining the constructs of fidelity to CSR



Step 1 **Define Constructs**

- *Expert panel (n = 9):*
  - Refine and clarify theoretical basis

Step 2 **Develop Items**

- *Team of CSR experts (n = 3):*
  - Revise scoring manual to provide operational definitions of items
  - Coroborate item responses with evidence from field notes
  - Elicit feedback on items and scoring guide from experienced CSR teachers
  - Revise items based on feedback
- *Pilot study #1*
  - One month (September 2013)
  - 3 separate video-recorded observations of teachers from one large, urban district in the western U.S.
  - Experienced CSR teachers in a previous experimental study of CSR (Boardman et al., 2015); 3 teachers, 2 schools, approximately 60 students
  - 5 raters + 1 lead scorer
  - Ratings conducted in conjunction with the lead scorer
  - Expert panel reviewed feedback, made revisions to items
- *Pilot study #2*
  - 3 months (October - December, 2013)
  - Experienced CSR teachers (different from pilot #1) within the same district from a previous RCT (Boardman et al., 2015); 12 teachers, 2 schools, roughly 300 students
  - 5 raters + lead scorer (same raters used in pilot study #1)
  - Observers met after each in-class observation to discuss scores until reaching inter-rater agreement of 90% or higher for each item
  - Expert panel reviewed feedback, made revisions
  - Final instrument (77 items) measuring adherence to CSR

Validity evidence based on: (1) test content; and (2) rater response processes (*Standards,* AERA, APA, and NCME, 2014).

Step 3 **Validation Study**

- 2nd semester of the school year (January -May, 2014)
- RCT (45 treatment, 39 control), 11 schools, 3,375 students
- 5 raters + lead scorer (same raters used in the previous pilot studies)
- 3 rounds of data collection (every 6 weeks)
- 352 observations in treatment and control classrooms
- Ensuring consistency: (1) Lead scorer compared notes to scores for 10% of ratings; and, (2) 20% of observations were conducted jointly but scored independently
- Data analysis (item analysis; internal consistency; inter-rater reliability; difference between conditions in scores; correlation with student reading outcomes)

Validity evidence based on: (1) internal structures; and (2) relations to other variables (*Standards,* AERA, APA, and NCME, 2014).

Figure 1. **Steps for validating the Implementation Validity Checklist-Revised (IVC-R).**

by drawing from prior work and through additional conceptual analysis; (2) designing new items and developing scoring guides though an iterative combination of team-based discussions and analysis of field notes collected by classroom observers; and (3) validating the IVC-R by evaluating the psychometric properties of the FOI measure. The first two of these steps generated validity evidence based on test content and on rater response processes, while the third step generated validity evidence based on internal structure and relations to other variables. Figure 1 illustrates this process.

Case Study Context

Data for the case study were drawn from a 5-year investigation of CSR implemented across 21 middle schools in one large urban school district in the western United States. Theory formation (i.e., defining constructs; step 1, Figure 1) and item development (step 2, Figure 1) occurred with middle school teachers who had previously participated in an experimental study of CSR (e.g., Boardman et al., 2015). The student population in this district is 20% white, 59% Hispanic, and 15% African American. Seventy-two percent of the students are eligible for the free or reduced-price lunch program and 35% are English language learners. A different group of teachers from the same district involved in a separate RCT investigating the efficacy of CSR was included in the validation study (step 3, Figure 1). For each experimental study, teachers who were assigned to the CSR condition received 2 days of up-front professional development, along with 6 hours of follow-up professional development throughout the school year. They were also offered individual coaching sessions approximately one time per month. Treatment teachers received lesson plans and student materials needed to implement the CSR model. Teachers in the control group did not receive professional development in CSR. The level of professional development ensured that teachers in the treatment group understood the CSR process and were supported in implementing CSR with fidelity. Inasmuch as teachers were well supported in their professional learning and implementation of CSR according to models of effective professional development (e.g., Garet, Porter, Desimone, Birman, & Yoon, 2001), CSR teachers were nevertheless subject to the common demands of teaching including time constraints, balancing multiple initiatives, and uneven commitment to implementation across schools.

Defining the Constructs of Fidelity to CSR

The CSR model was developed based on the extant literature in the domains of cognitive psychology (Flavell, 1979; Palincsar & Brown, 1984), evidence-based reading comprehension practices (Scammacca et al., 2007), and evidence-based pedagogical practices such as cooperative learning (e.g., Johnson & Johnson, 1999; Kagan, 1986). The original IVC captured the practices outlined in CSR. For the case study, an expert panel of reviewers helped refine and clarify the theoretical basis of the original IVC in developing the IVC-R. The expert panel consisted of nine individuals with diverse expertise and included the authors of this study, a developer of CSR, a measurement expert, a statistician, trained classroom observers, and former classroom teachers with extensive experience with CSR and reading instruction.

The expert panel defined the following two constructs underlying the IVC-R: (1) procedural fidelity, which refers to fidelity *across* the five CSR components, and (2) pedagogical fidelity, which refers to fidelity *within* each CSR component.

Procedural Fidelity

This construct represents the five components of the CSR model (Preview, Click and Clunk, Get the Gist, Questions, and Review), and items are specific to reading comprehension strategies observed in each CSR component. For example, a teacher implementing CSR with fidelity will engage in the observable behavior as inscribed by the following item: "Attends to quality of students' questions by discussing question types, asking questions related to correctness [of textual interpretations], or guiding students to reread or re-work a question that isn't quite right," during the Questions portion of a lesson.

Pedagogical Fidelity

This construct represents teaching approaches embedded in the CSR process that incorporate discussion about texts into instruction as a means of increasing engagement and comprehension (Lawrence & Snow, 2010). It includes(1) fostering collaboration, in which teachers encourage students working in small collaborative groups to build upon one another's ideas in making sense of the reading; and (2) managing the learning environment, where teachers facilitate pacing and participation so that all students have opportunities to share and to hear ideas from others. These items are observable behaviors that may occur consistently throughout a lesson. For example, a teacher may "keep students engaged and participating" during each portion of the CSR lesson, so that item, which is part of the "managing the learning environment" domain, will be assessed within each CSR component. These items are not unique to CSR and may be observed in non-CSR lessons. However, because they are integral to the CSR model, they are included in the observation tool.

Item Development

Item development started with a team of CSR experts, including the principal investigator and two postdoctoral fellows with expertise in reading (each of whom were also members of the expert panel) drafting descriptions of specific observable teacher and student behaviors or actions that characterize high-quality CSR implementation. A list of possible items was presented to experienced CSR teachers (not involved in the pilot or validation studies, described further in the next section) and to CSR coaches for review. These teachers and coaches were then interviewed by the team of CSR experts about the relevance and clarity of items. Results were presented to the expert panel, which then revised and/or eliminated items as needed. Further revisions were made by iteratively examining field notes and videos collected by observers and receiving additional feedback from observers and other members of the design team.

Scoring Process

The nature of the identified behaviors informed the scoring system of the IVC-R; all teacher behavior items and most student behavior items are scored dichotomously, as observed (1)/not observed (0). Since collaboration and cooperative learning are key features of CSR and students work in groups of up to four as part of the learning process, taking note of the frequency of group participation was an important indicator of adherence to the CSR model. Therefore, for each of the general student behaviors (e.g., sharing gist statements with the group during reading), an additional set of items was designed to be scored as a percentage of students in a group engaged in the observable behavior. For example, if three out of the four students in the group "write their own main idea (gist) statements," the item is scored as .75. The scoring manual was adapted from the original IVC to reflect items from the IVC-R in providing operational definitions of the possible scores for every item on the IVC-R, a practice that has been used in other intervention studies examining the relationship between fidelity and outcomes using a rating scale (Childs et al., 2016; George & Childs, 2012). While conducting an observation, raters took field notes and used the manual to score items on the IVC-R by either checking a box if the item is observed or by entering a ratio of students engaged in the behavior.

Observing Student Groups

A process was needed for capturing a representative sample of student behaviors observed across the classroom at each phase of the CSR process. Since classes in the case study averaged more than 25 students each (with some class sizes as high as 35 students), the number of cooperative student groups per class (n = 4) typically ranged from six to eight. Because of the inherent variation in student participation in classroom activities, the difficultly of accurately capturing student behaviors across all student groups was a known weakness of the original IVC, however collecting data on every student's behavior in the classroom was impractical and inefficient. Devine, Rapp, Testa, Henrickson, & Schnerch (2011) suggest that accurately assessing students' behavior in a classroom setting requires that students be randomly selected for observation or that data be collected class-wide with the class considered as a single entity using recording methods such as time sampling. Examining discrete student behaviors in a classroom setting via a time-sampling method has been used in several observational studies (Flower, McKenna, Muething, Bryant, & Bryant, 2014; Guo, Connor, Yang, Roehrig, & Morrison, 2012; Tiger et al., 2013). For example, Cappella, Kim, Neal, & Jackson (2013) used time sampling in a series of systematic

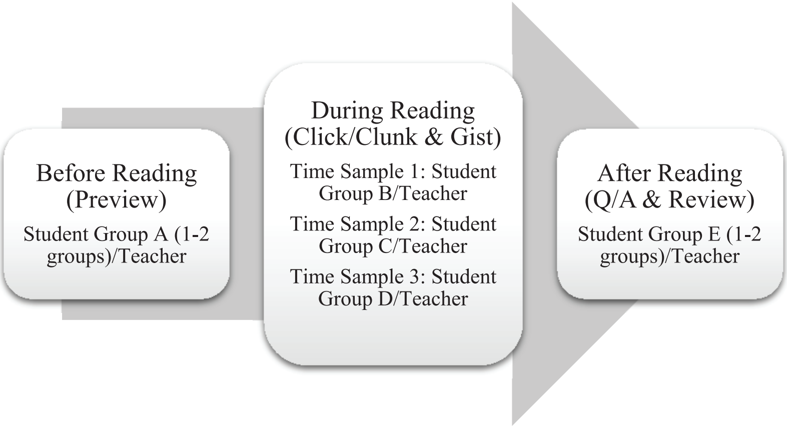

Before Reading (Preview)
Student Group A (1-2 groups)/Teacher
During Reading (Click/Clunk & Gist)
Time Sample 1: Student Group B/Teacher
Time Sample 2: Student Group C/Teacher
Time Sample 3: Student Group D/Teacher
After Reading (Q/A & Review)
Student Group E (1-2 groups)/Teacher

Figure 2: **Rotation schedule for observing cooperative/student groups using the IVC-R (for each rotation, a new student group is observed, for a total of up to 5 groups observed throughout a lesson). For classes with larger sizes, 1–2 groups (*n* = 4 per group) are observed for the "Before Reading" and "After Reading" portion of the CSR lesson.**

Note: IVC-R = Implementation Validity Checklist – Revised; CSR = Collaborative Strategic Reading; Q/A = Question and Answer.

classroom observations to assess behavioral engagement among second to fifth graders. Time sampling was also used by the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development to successfully capture discrete behaviors and classroom conditions related to the social and academic development of students in grades 1–5 (Pianta & Hambre, 2009). As described further below, a plan was thus developed for rotating observations of students so that most or all cooperative groups are observed for some portion of the lesson (Figure 2).

Identifying groups to observe was done randomly with the rater moving to an unobserved group during each rotation. For the first rotation, the observer watches and scores 1–2 student groups and the teacher during the "Preview" portion of the lesson. The second rotation occurs in the "During reading" part of the process and includes three 10-minute time samples with the observer watching and scoring an additional three student groups that were not observed during "Preview," one for each time sample, while simultaneously watching and recording teacher behaviors throughout all three time samples. For the Questions and Review components ("After reading"), the observer watches and scores 1–2 additional previously unobserved student groups, while continuing to record teacher behaviors.

Rater Response Processes

The investigation of the response processes employed by raters in recording observations made visible the ways in which the IVC tool captured key instructional and learning moves within a CSR session. To the extent possible, item responses were corroborated with evidence from field notes. Conflicting evidence was resolved by discussing the notes with the rater and revising wording so that each item was clear to observers.

Pilot study. The IVC-R was field tested using five raters consisting of four CSR coaches with up to 3 years of experience supporting CSR implementation, and a former administrator in the school district who was also serving as a CSR coach. These raters also had expertise in literacy, something Hill & Grossman (2013) say is necessary in utilizing instruments designed to measure teacher evaluation systems aimed at improving instructional practice. All teachers in the study agreed to participate in data collection. Classroom observations were scheduled in advance, and teachers were told that observers would be taking field notes. Raters were assigned to observe classrooms according to schedules and availability, resulting in each rater observing teachers across schools and content areas.

The pilot included two phases. In the first phase, which occurred in the first month of the school year, a lead rater from the expert panel led joint scoring sessions with all raters that included watching 50-minute video-recorded CSR lessons of three teachers and approximately 60 students from two middle schools and discussing the scoring of the observation. The raters' scores were compared with the lead rater's scores. The lead scorer thus served as the "gold standard of reliability," a training process that has been used in other studies that report FOI results (Wanzek et al., 2014). Discrepancies were discussed and this process continued with each of the three video-recorded observations, allowing raters practice in rating using the lead scorer's input and guidance. The expert panel then reviewed and discussed all forms of feedback until consensus was reached for wording and scoring of each item (Cavanagh & Koehler, 2013).

In the second phase, the initial IVC was field tested over a 3-month period with 12 veteran CSR teachers and approximately 300 students from two middle schools. These teachers and their students were not included in the validation study (i.e., step 3, Figure 1). A classroom observation consisted of an observation of one session of CSR. In this way, observation length varied from 44 to 90 minutes according to class schedules and the amount of time teachers engaged in CSR instruction. This second round of pilot observations was conducted by the same team of five raters who observed each lesson together and then discussed scores as a group. Discrepancies were addressed until reaching an interrater agreement of 90% on all items and problematic items (i.e., those with consistently low reliability) were revised by the expert panel. This process identified 77 observable IVC items that comprise teacher and student behaviors representing adherence to the CSR model (Table 1). To ease with interpretation, scored responses were summed across all 77 IVC items (19 items scored as a percent of student behaviors observed in cooperative groups and 58 student and teacher items scored dichotomously as observed/not observed) and then the raw scores were standardized to have a mean of 100 and standard deviation of 15.

Validation study. The validation study (step 3, Figure 1) occurred in the context of a RCT in which teachers were randomly assigned to either implement CSR (treatment) or continue to use their business-as-usual instructional strategies (control). The study included 79 social studies, science and language arts teachers (45 in treatment and 34 in control) from 11 middle schools across a large urban district in the Western United States. Students received CSR in two separate classes (i.e., social studies and science or language arts). In addition, some teachers taught multiple subject areas and/or grade levels. Therefore, to ensure that teachers and students maintained condition (treatment or control), randomization occurred at the teacher, grade, or school level, depending on each school's structure. Thus, the nature of the "coin flip" of randomization resulted in unequal numbers of teachers in each condition. Because the study was designed so that CSR was implemented 1 day a week throughout the school year in social studies and 1 day a week in either science or language arts classes, each social studies class roster was paired with either a science class roster or a language arts class roster within each school to create one social studies/science teacher pair or one social studies/language arts teacher pair. Thus, the experimental unit included a possible combination of 78 teacher pairs of social studies, science and/or language arts teachers and the initial sample included 3,375 students (2,133 in treatment and 1,602 in control). Raters conducted three rounds of IVC-R observations per teacher, starting in winter and re-occurring every 6 weeks until the end of the school year, resulting in a total of 210 classroom observations conducted in CSR and non-CSR classes using the IVC-R instrument.

Raters were assigned to observations based on the availability of their schedules and upon the days in which teachers responded that an observation could occur. Thus, raters observed across school sites within the district. Raters for the experimental study consisted of the same raters for the pilot study. Though having raters blind to condition is ideal in experimental design studies, such objectivity is generally not possible in fidelity observations since the many models, by design, have distinct features that are unmistakable to an observer (e.g., unique student materials, names of specific strategy components). To increase objectivity, raters received extensive training in how to become reliable using the IVC in CSR classrooms and in classrooms without CSR. In addition, as mentioned previously, all raters held

Table 1
Items on the Implementation Validity Checklist – Revised (IVC-R)

Preview: teacher behaviors	Item	Preview: student behaviors	Item
Presents the topic;	1	Write what they already know about the topic;	6
Presents a brainstorm prompt;	2	Share their brainstorm with their group or partner;	7
Presents a purpose for the lesson;	3	Are on task and attentive, e.g., listen, contribute when prompted, participate.	8
Prompts use of specific student roles;	4		
Reminds students to share brainstorms in groups.	5		
Clunks/fix-up strategies: teacher behaviors		**Clunks/fix-up strategies: student behaviors**[a]	
Attends to the quality of definitions by asking questions related to correctness, guiding students to redefine an unknown word/phrase that isn't quite right, or providing information when needed;	51	Identify unknown words/phrases;	9, 23, 37
		Use fix-up strategies to define unknown words/phrases;	10, 24, 38
		Check definitions by returning to the text;	11, 25, 39
		Use CSR resources such as flipbooks, role cards;	12, 26, 40
Encourages students to collaborate/work together.	52	Discuss unknown words and definitions.	13, 27, 41
Gist: teacher behaviors		**Gist: student behaviors**[a]	
Attends to quality of students' main idea statements by asking questions related to correctness, or guiding students to re-read or re-work a statement that is not quite right;	53	Identify the most important who/what and the most important information about the who/what;	14, 28, 42
		Write their own main idea statements;	15, 29, 43
		Share their main idea statements with their group;	16, 30, 44
Encourages students to discuss the most important "who/what" and the most important information about the "who/what";	54	Refer back to the text when discussing ideas;	17, 31, 45
		Revise their main ideas after discussing with their group or with the teacher;	18, 32, 46
Encourages students to collaborate/work together in discussing content and/or offering feedback related to main idea statements;	55	Use CSR resources such as flipbooks, role cards;	19, 33, 47
		Discuss ideas about the main idea;	20, 34, 48
		Discuss content of and/or offer feedback on each other's main idea statements;	21, 35, 49
Prompts students to work in a timely manner during reading: keeps students engaged and participating.	56	Are on task and attentive.	22, 36, 50

(continued)

Table 1 **(continued)**

Preview: teacher behaviors	Item	Preview: student behaviors	Item
Questions: teacher behaviors	**Item**	**Questions: student behaviors**	**Item**
Attends to quality of questions by discussing question types, asking questions related to correctness, or guiding students to re-read or re-work a question that isn't quite right;	57	Write own leveled questions;	60
Encourages students to collaborate/work together in discussing content and/or offering feedback related to leveled questions;	58	Answer their own questions;	61
Prompts students to work in a timely manner: Keeps students engaged and participating, e.g., students are not waiting for the teacher.	59	Share one/more of their questions with their group;	62
		Answer each other's questions in their groups;	63
		Return to/reference the text to answer questions;	64
		Use CSR resources such as flipbooks, role cards;	65
		Discuss the answers to each other's questions and/or offer feedback to each other in writing leveled questions;	66
		Are on task and attentive.	67
Review: teacher behaviors	**Item**	**Review: student behaviors**	
Attends to quality of summary statements by discussing content, asking questions related to correctness, or guiding students to re-read or re-work a statement that isn't quite right;	68	Write one or two of the most important ideas from the entire passage;	72
Brings whole class together for a brief review;	69	Share their summary with their group;	73
Encourages students to collaborate/work together in discussing content and/or offering feedback related to review;	70	Provide evidence from the text to support why the ideas they included in their summary statement are important;	74
Prompts students to work in a timely manner: Keeps students engaged and participating, e.g., students are not waiting for the teacher.	71	Discuss ideas about the summary;	75
		Discuss content of each other's summary statements and/or offer feedback on the ideas presented;	76
		Are on task and attentive.	77

[a] These items appear three times in the IVC-R instrument because during a CSR lesson, a teacher selects a content-focused text and then divides the text into sections (usually three for one class period).

deep knowledge of evidence-based reading comprehension practices and effective instruction. While the purpose of the IVC-R scoring protocol was to assess CSR usage, it was also designed to capture similar evidence-based comprehension practices in comparison classrooms.

Observations consisted of the full length of the class period and ranged from 50 minutes across six school sites (*n* = 40 teachers) to 90 minutes across five school sites (*n* = 39 teachers). The number of sections taught by each teacher varied from one to five classes a day. A different class was observed for each teacher during each round of IVC-R data collection to maximize the sampled breadth of CSR instruction. As such, at least one class was observed more than once for those teachers who taught less than three sections (*n* = 5). Likewise, teachers with more than three sections had some classes that were not observed during any of the three IVC-R rounds of data collection (*n* = 37). The remaining 32 teachers taught three sections, so each of their classes were observed during one round of IVC-R data collection. Observations were scheduled to ensure that a CSR lesson would be observed with treatment teachers, or a typical day of instruction with control teachers.

The global CSR adherence score (i.e., the sum of the 77 items), averaged across the three observations, was used for analysis. For each observation, items in the "during reading" portion of the instrument are averaged across each section completed. Thus, if a group observed during the time samples completed all three sections of the text, the denominator is 3; if the group completed only 2 sections, the denominator is 2, etc.

Of the 210 observations, approximately 20% were conducted by two raters observing the same classroom together and scoring the observation independently to test whether student and teacher behaviors embedded in the CSR model could be consistently assessed across raters, classrooms, and schools. These double-scored observations were used to determine interrater reliability (estimated using intra-class correlations). IRR analysis was based on scores assigned to observations before any discussion occurred among raters and the lead scorer about reconciling differences in scores. The remaining 80% of observations were conducted independently, of which approximately 10% were randomly sampled by the lead scorer to check for accuracy. The lead rater used field notes to analyze scores by CSR component (Preview, Click and Clunk, Get the Gist, Questions, Review). Discrepancies between the recorded rating and the lead scorer's rating were discussed and scores were revised as needed so as to ensure all double-scored observations were reconciled and that one score for each teacher across each observation round was used in the final analyses.

Instrument Analysis

In contrast to many traditional tests (for example, of cognitive ability, academic proficiency, or personality characteristics), FOI measures such as the IVC are not hypothesized to measure an attribute of persons (or "construct") that exists and possess a particular dimensional structure independent of the measuring instrument. For example, the IVC-R described in our case study can be used to summarize and communicate information about fidelity to the CSR instructional model according to overall fidelity, procedural and pedagogical fidelity, and fidelity on specific CSR components. Further, the choice to summarize information in these ways, and more specific choices regarding which of these levels of focus are of interest in any given application, are based on the theoretical definition of CSR and on human judgment concerning which forms of information are most valuable and meaningful, rather than on a priori theories concerning the "true" structure of fidelity to the CSR instructional model. In other words, it is not hypothesized that variation in the attribute of fidelity to CSR (or more specific subattributes) causes variation in the specific items on the IVC; rather, "fidelity to CSR" can be thought of as an inductive summary of the behaviors observed on the IVC. As such, reflective latent variable models (e.g., confirmatory factor analysis) may not be appropriate in validating FOI measures, as with the investigation presented as our case study (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). Nevertheless, basic item and test analysis can be pragmatically utilized in the service of quality control when measuring FOI; here we focus on three statistics used in Classical Test Theory (CTT): (1) the *p* value, or item difficulty, estimated based on the frequency with which a specific behavior was observed; (2) the (corrected) item-total (or point-biserial) correlation coefficient, which estimates the correlation between each item and scores on the total instrument after omitting that item; and (3) Cronbach's alpha, which estimates the proportion of total score variance that is not due to measurement error.

Relations With Other Variables

The IVC-R is intended to provide an assessment of fidelity to the CSR instructional model (that is, whether or not CSR was used as intended in a classroom), as opposed to an assessment of the overall quality of classroom teaching. It would therefore be problematic if the IVC-R could not differentiate between high-quality classes taught using CSR methods and high-quality classes not taught using CSR methods. Thus, given that treatment teachers had professional development in the CSR model and control teachers did not, it was hypothesized that the CSR teachers would receive higher scores on the IVC-R compared with teachers in the control condition. To test whether the IVC-R was indeed sensitive to the features unique to CSR, multilevel models were employed using the Hierarchical Linear Modeling 7.0 software program (Raudenbush, Bryk & Congdon, 2010), thereby accounting for the clustering of observations between schools within teachers. IVC-R scores, entered as grand mean centered, were regressed on a treatment indicator variable (level 1), in which the reference group was typical instruction. Random intercepts for school (level 2) were included, thus treating teachers as randomly sampled from hypothetical distributions of possible schools and allowing each school to have its own mean on the response variable. Applying the notation of Raudenbush and Bryk (2002), model specifications are provided below.

Level 1 (teacher-pair level):

$$y_{ij} = \beta_{0j} + \beta_{1j} * \left(Treatment_{ij}\right) + r_{ij}$$

where:

y_{ij} is the outcome of interest (i.e., overall fidelity) for teacher-pair i in school j;

$Treatment_{ij}$ is a dummy variable for treatment (1 if teacher-pair i in school j is in the treatment group, 0 otherwise);

β_{0j} is the estimated adjusted mean of the outcome of interest;

β_{1j} is the estimated treatment effect; and

r_{ij} is the residual associated with teacher-pair i in school j (assumed to be normally distributed with a mean of zero and variance of σ_e^2);

Level 2 (school level):

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

where:

γ_{OO} is the estimated adjusted mean of the outcome of interest in comparison schools;

u_{0j} is the residual associated with school j (assumed to be normally distributed with a mean of zero and variance of σ^2 and uncorrelated with the teacher-pair-level residual).

Additionally, given the empirical link between instruction—as measured by classroom observation tools—and achievement (e.g., Abbott & Fouts, 2003; Belsky et al., 2005; Hamre et al., 2007; Sawada et al., 2002) along with the theoretical basis for the CSR instructional model, higher fidelity to CSR should, in principle, be associated with higher levels of reading comprehension among students. In the case study, the reading comprehension subtest of the Gates-MacGinitie Reading Test (GMRT), Fourth Edition (MacGinitie, MacGinitie, Maria, & Dreyer, 2000) was used, which is the outcome measure in several experimental studies of CSR (e.g., Boardman et al., 2015, 2016b; Vaughn et al., 2011). Two parallel forms were employed to permit pre- and posttesting. The GMRT (a timed, paper-and-pencil test) was group-administered to students in treatment and comparison classrooms by trained researchers blind to condition at the beginning and end of the school year. Estimates of internal consistency for the GMRT range from .91 to .93 and estimates of alternate-forms reliability range from .80 to .87.

Multilevel models were again employed, this time to investigate differences in GMRT scores at posttest as a function of IVC-R scores, controlling for GMRT pretest scores. For this analysis, three-level models were estimated in which students were treated as nested in teacher pairs and teacher pairs were treated as nested in schools. Given the interest in a treated-on-the-treated model assessing the relationship between fidelity to the CSR model and reading outcomes, the IVC-R overall fidelity score of the teacher pair (as opposed to a dichotomous variable for condition) was the primary treatment variable. Since prior research has found CSR to be particularly effective for struggling readers (Boardman et al., 2016a, 2016b; Kim et al., 2006; Klingner et al., 2004), an additional model was fit with an interaction term between pretest GMRT scores and IVC-R scores to examine the hypothesis that higher adherence to the CSR model is associated with greater gains in reading for students with a lower initial level of reading ability compared to those with a higher initial level of ability. IVC-R scores varied at level 2 because that is where the random assignment took place, and GMRT pretest scores were entered into the model as grand mean centered. The model took the following form:

Level 1 (student level):

$$y_{ijk}=\pi_{0jk}+\pi_{1jk}*\left(Pretest_{ijk}\right)+e_{ijk}$$

where:

y_{ijk} is the outcome of interest (i.e., GMRT) for student *i* in teacher-pair *j* in school *k*;

$Pretest_{ijk}$ is the pretest score for student *i* in teacher-pair *j* in school *k*;

$\pi_{\ 0jk}$ is the estimated adjusted mean of the outcome of interest;

π_{1jk} is the total effect of pretest;

e_{ijk} is the residual associated with student *i* in teacher-pair *j* in school *k*;

Level 2 (teacher-pair level):

$$\pi_{0jk}=\beta_{00k}+\beta_{01k}*\left(Overall\,fidelity_{jk}\right)+r_{0jk}$$

$$\pi_{1jk}=\beta_{10k}+\beta_{11k}*\left(Overall\,fidelity_{jk}\right)$$

where:

$Overall\ fidelity_{jk}$ = Overall fidelity score of teacher-pair *j* in school *k*;

β_{00k} is the estimated adjusted mean of the outcome of interest in comparison teacher pairs;

β_{01k} is the estimated main effect of fidelity;

β_{10k} is the estimated main effect of the pretest;

β_{11k} is the estimated interaction between fidelity and the pretest;

r_{0jk} is the residual associated with teacher-pair *j* in school *k* (assumed to be normally distributed with a mean of zero and variance of σ_e^2);

Level 3 (school level):

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

where:

γ_{OOO} is the estimated adjusted mean of the outcome of interest in comparison schools;

u_{OOk} is the residual associated with school k (assumed to be normally distributed with a mean of zero and variance of σ_u^2 and uncorrelated with the student- and teacher-pair-level error terms).

Results

Evidence of Validity Based on Response Processes

Evidence of validity based on response processes was established by including assurances that raters were consistent in their understanding and use of the IVC-R instrument, and that raters' intended interpretation of the scores was accurate (as described in the methods section). An analysis of how raters collected and interpreted data during data collection showed high interrater reliability (ICC = .972) for the 20% of observations (n = 42) that were conducted by two raters independently scoring the IVC-R.

Evidence of Validity Based on Internal Structure

Results of the item analysis are displayed in Table 2. This table includes the average score of all respondents (across both treatment and control groups) on each item (i.e., the p value, or estimated difficulty level), and lists results by the 49 procedural fidelity and 28 pedagogical fidelity items.

Item Difficulty

Approximately 65% of the items had p values between .25 and .75. Of the 58 dichotomously scored items, only five (less than 10 percent) had p values over .50 of the time. In contrast, 14 items (about 25%) had p values below .10. Overall, among the dichotomous items, strategies in the Preview component were observed the most and Review strategies were observed the least. While there are more items about student behaviors than teacher behaviors, teacher behaviors were on average observed more often than student behaviors. Among the continuous items, student behaviors related to being on task (i.e., management) were observed more than student behaviors demonstrating the use of strategies (i.e., collaboration). These patterns are in relation to the items listed and categorized in Table 2.

Table 2
Item Analysis Used to Tap Each of the IVC-R Domains

Item	*M*[a]	*ITC*[b]	Domain	Subdomain	CSR component
1[d]	0.81	0.388	Procedural fidelity	Strategies	Preview
2[d]	0.68	0.452	Procedural fidelity	Strategies	Preview
3[d]	0.67	0.414	Procedural fidelity	Strategies	Preview
4[d]	0.38	0.521	Procedural fidelity	Strategies	Preview
6[c]	0.52	0.559	Procedural fidelity	Strategies	Preview
9	0.51	0.554	Procedural fidelity	Strategies	Clunk, fix-up (time sample 1)
10	0.25	0.469	Procedural fidelity	Strategies	Clunk, fix-up (time sample 1)
11	0.14	0.408	Procedural fidelity	Strategies	Clunk, fix-up (time sample 1)
12	0.27	0.585	Procedural fidelity	Strategies	Clunk, fix-up (time sample 1)
14	0.42	0.520	Procedural fidelity	Strategies	Gist (time sample 1)
15[c]	0.31	0.596	Procedural fidelity	Strategies	Gist (time sample 1)
16[c]	0.15	0.501	Procedural fidelity	Strategies	Gist (time sample 1)
17	0.22	0.358	Procedural fidelity	Strategies	Gist (time sample 1)
18	0.08	0.224	Procedural fidelity	Strategies	Gist (time sample 1)
19	0.17	0.378	Procedural fidelity	Strategies	Gist (time sample 1)
23	0.40	0.566	Procedural fidelity	Strategies	Clunk, fix-up (time sample 2)
24	0.19	0.521	Procedural fidelity	Strategies	Clunk, fix-up (time sample 2)
25	0.11	0.292	Procedural fidelity	Strategies	Clunk, fix-up (time sample 2)
26	0.15	0.447	Procedural fidelity	Strategies	Clunk, fix-up (time sample 2)
28	0.40	0.506	Procedural fidelity	Strategies	Gist (time sample 2)
29[c]	0.41	0.718	Procedural fidelity	Strategies	Gist (time sample 2)
30[c]	0.24	0.594	Procedural fidelity	Strategies	Gist (time sample 2)
31	0.14	0.308	Procedural fidelity	Strategies	Gist (time sample 2)
32	0.07	0.320	Procedural fidelity	Strategies	Gist (time sample 2)
33	0.11	0.429	Procedural fidelity	Strategies	Gist (time sample 2)
37	0.32	0.571	Procedural fidelity	Strategies	Clunk, fix-up (time sample 3)
38	0.11	0.485	Procedural fidelity	Strategies	Clunk, fix-up (time sample 3)
39	0.08	0.437	Procedural fidelity	Strategies	Clunk, fix-up (time sample 3)
40	0.09	0.390	Procedural fidelity	Strategies	Clunk, fix-up (time sample 3)
42	0.33	0.472	Procedural fidelity	Strategies	Gist (time sample 3)
43[c]	0.33	0.679	Procedural fidelity	Strategies	Gist (time sample 3)
44[c]	0.19	0.569	Procedural fidelity	Strategies	Gist (time sample 3)
45	0.13	0.334	Procedural fidelity	Strategies	Gist (time sample 3)
46	0.05	0.333	Procedural fidelity	Strategies	Gist (time sample 3)
47	0.11	0.370	Procedural fidelity	Strategies	Gist (time sample 3)
51[d]	0.33	0.463	Procedural fidelity	Strategies	Clunk, Fix-up
53[d]	0.44	0.626	Procedural fidelity	Strategies	Gist
57[d]	0.26	0.511	Procedural fidelity	Strategies	Questions
60[c]	0.30	0.775	Procedural fidelity	Strategies	Questions
61[c]	0.30	0.802	Procedural fidelity	Strategies	Questions
62[c]	0.12	0.620	Procedural fidelity	Strategies	Questions

(continued)

Table 2 **(continued)**

Item	M^a	ITC^b	Domain	Subdomain	CSR component
63	0.12	0.459	Procedural fidelity	Strategies	Questions
64	0.04	0.257	Procedural fidelity	Strategies	Questions
65	0.23	0.644	Procedural fidelity	Strategies	Questions
68[d]	0.10	0.345	Procedural fidelity	Strategies	Review
69[d]	0.24	0.334	Procedural fidelity	Strategies	Review
72[c]	0.23	0.590	Procedural fidelity	Strategies	Review
73[c]	0.05	0.349	Procedural fidelity	Strategies	Review
74	0.08	0.177	Procedural fidelity	Strategies	Review
5[d]	0.30	0.490	Pedagogical fidelity	Collaboration	Preview
7[c]	0.22	0.525	Pedagogical fidelity	Collaboration	Preview
13	0.37	0.566	Pedagogical fidelity	Collaboration	Clunk, fix-up (time sample 1)
20	0.28	0.483	Pedagogical fidelity	Collaboration	Gist (time sample 1)
21	0.07	0.297	Pedagogical fidelity	Collaboration	Gist (time sample 1)
27	0.24	0.607	Pedagogical fidelity	Collaboration	Clunk, fix-up (time sample 2)
34	0.29	0.529	Pedagogical fidelity	Collaboration	Gist (time sample 2)
35	0.09	0.428	Pedagogical fidelity	Collaboration	Gist (time sample 2)
41	0.19	0.603	Pedagogical fidelity	Collaboration	Clunk, fix-up (time sample 3)
48	0.23	0.481	Pedagogical fidelity	Collaboration	Gist (time sample 3)
49	0.08	0.420	Pedagogical fidelity	Collaboration	Gist (time sample 3)
52[d]	0.24	0.360	Pedagogical fidelity	Collaboration	Clunk, fix-up
54[d]	0.39	0.329	Pedagogical fidelity	Collaboration	Gist
55[d]	0.28	0.477	Pedagogical fidelity	Collaboration	Gist
58[d]	0.09	0.189	Pedagogical fidelity	Collaboration	Questions
66	0.13	0.409	Pedagogical fidelity	Collaboration	Questions
70[d]	0.02	0.023	Pedagogical fidelity	Collaboration	Review
8[c]	0.86	0.241	Pedagogical fidelity	Management	Preview
22[c]	0.88	0.265	Pedagogical fidelity	Management	Gist (time sample 1)
36[c]	0.81	0.365	Pedagogical fidelity	Management	Gist (time sample 2)
50[c]	0.67	0.236	Pedagogical fidelity	Management	Gist (time sample 3)
56[d]	0.68	0.386	Pedagogical fidelity	Management	Gist
59[d]	0.39	0.662	Pedagogical fidelity	Management	Questions
67[c]	0.45	0.670	Pedagogical fidelity	Management	Questions
71[d]	0.34	0.428	Pedagogical fidelity	Management	Review
75	0.06	0.226	Pedagogical fidelity	Management	Review
76	0.02	0.105	Pedagogical fidelity	Management	Review
77[c]	0.43	0.327	Pedagogical fidelity	Management	Review

Note. IVC-R = Implementation Validity Checklist – Revised; [a]*M* = mean value (average score); [b]Item-to-Total Correlation (ITC) = Point-biserial correlation coefficient. [c]Items scored as a percentage demonstrating the behavior (the remaining items are scored dichotomously, with 1 = observed). [d]Item measuring a teacher behavior.

Table 3
Fixed and Random Effects of Condition on Fidelity to CSR, as Measured by the IVC-R

	Fixed Effects			
Predictor	Estimate *(SE)*	*T* Ratio[a]	*p* value	Hedges' *g*
Intercept, $\beta0$				
Intercept, $\gamma000$	85.75 (1.59)	54.10	<.001	
Intercept, $\beta10$				
Condition[b], $\gamma100$	22.83 (2.11)	10.84	<.001	2.61
	Random Effects			
	Variance (SD)	*T* ratio[a]	*p* value	% of total variation
Level 1 (teacher)	74.80 (8.65)	8.65		97.1%
Level 2 (school)	2.28 (1.51)	1.51	.264	2.9%

Note. CSR = Collaborative Strategic Reading (CSR); IVC-R = Implementation Validity Checklist – Revised. IVC-R raw scores were transformed into z-scores and then standardized with a mean of 100 and standard deviation of 15. [a]The T ratio for fixed effects was determined by dividing the estimate by its standard error; for random effects, the T ratio was determined by dividing the variance component by its standard deviation. [b]Reference group is business-as-usual instruction (control group). Condition (CSR = 1, Control = 0).

Item-Total Correlations

For the majority of items, item-total correlations were in line with expectations; that is, high total scores on the IVC-R were associated with high scores on individual items, and vice versa. No items had an item-total correlation at or below zero. As presented in Table 3, however, 12 of the 77 items had marginal to low item-total correlations ($r < .30$). Six of these 12 items were targeted at strategies related to Questions and Review, components of the CSR model that were scored as "not observed" for the majority of lessons. Observation data indicate that teachers frequently did not get to these components, which are included as part of the "after reading" portion of the CSR lesson. Factors such as pacing, short class periods, and emphasizing other CSR strategies contributed to the low observation rate of these items.

Reliability

Cronbach's alpha was estimated at .951, suggesting an acceptable level of internal consistency.

Evidence of Validity Based on Relations to Other Variables

As described previously, we investigated the extent to which the IVC-R was sensitive to differences between classrooms taught by CSR-trained teachers and control classrooms. Descriptive statistics show the mean IVC-R score for the treatment group was 109.66 (*SD* = 14.49) for round 1, 107.12 (*SD* = 12.90) for round 2, and 109.92 (*SD* = 13.08) for round 3, and the mean IVC-R score for CSR teachers averaged across all three rounds was 109.73 (*SD* = 13.03). In contrast, the mean IVC-R score for the control group was 85.37 (*SD* = 5.90) for round 1, 85.46 (*SD* = 4.47) for round 2 and 87.75 (*SD* = 4.38) for round 3 with a mean averaged across all three rounds of 87.67 (*SD* = 4.77). The large standard deviation in the treatment group (compared with control), however, suggests varying levels of CSR implementation, a finding that was consistent across all three rounds of data collection. Results, presented in Table 3, show that the average IVC-R score was higher for teachers in treatment than for teachers in control ($\beta = 22.83$, *SE* = 2.11, *p* < .001). Thus, on average, CSR was implemented to a greater extent in the treatment group compared to the control group. This difference is equivalent to a bias-adjusted Hedges *g* effect size of 2.61. The interpretation of Hedge's *g* (like most effect size metrics), however, assumes that the standard deviations of the two groups are equal and are normally distributed. Since there is a floor effect in the control group (that is, most of the items on the IVC-R were marked as "not observed"), these assumptions are violated; therefore, we corroborated our finding with the observer field notes.

Adherence to the CSR model, as measured by the IVC-R, was positively associated with Gates-MacGinitie Reading Test scores at posttest, while controlling for pretest scores ($\beta = 0.05$, *SE* = 0.02, *p* < .05; Table 4). Contrary to expectations, there was a significant positive interaction effect between fidelity and pretest scores ($\beta = 1.5^{-3}$, $SE = 5.0^{-4}$, *p* < .05), indicating that higher IVC-R scores are associated with greater growth in reading for students with initially-higher scores. However, the magnitude of this difference in effects was very small.

Discussion

Assessing FOI is important for understanding treatment effectiveness (Nelson et al., 2012), offering insight into whether or not a program was implemented as planned (Dane & Schneider, 1998; O'Donnell, 2008), enhancing the validity of claims derived from studies of treatment effectiveness, and optimizing how findings can be used to inform practice (Cordray & Pion, 2006). Findings produced from valid measures of FOI can improve upon researchers' ability to attribute outcomes to the intervention and help practitioners feel more confident in implementing the chosen intervention as

Table 4

Fixed and Random Effects of Fidelity to CSR (as Measured by the IVC-R) and Pretest Scores on Posttest Scores

	Fixed Effects		
Predictor	Estimate *(SE)*	*T* ratio[a]	*p* value
Intercept, β0			
Intercept, γ000	96.59 (0.59)	162.68	<.001
Fidelity[b], γ010	0.05 (0.02)	2.10	.040
GMRT pretest, β10			
Intercept, γ100	0.70 (0.01)	60.54	<.001
Fidelity[b], γ110	0.002 (0.0001)	2.59	.010

	Random Effects			
	Variance (SD)	*T* ratio[a]	*p* value	% of total variation
Level 1 (individual)	71.24 (8.44)	8.44		90.15%
Level 2 (teacher)	5.97 (2.44)	2.44	<.001	7.55%
Level 3 (school)	1.82 (1.35)	1.35	.006	2.31%

Note. Pretest scores were entered into the model as grand mean centered. CSR = Collaborative Strategic Reading (CSR). IVC-R = Implementation Validity Checklist – Revised. GMRT = Gates MacGinitie Reading Test.

[a] The T-ratio for fixed effects was determined by dividing the estimate by its standard error; for random effects, the T-ratio was determined by dividing the variance component by its standard deviation.

[b] Fidelity = The sum of the 77 items on the IVC-R. IVC-R raw scores were transformed into z-scores and then standardized with a mean of 100 and standard deviation of 15.

it was intended. Yet, rigorous assessment of FOI observational tools is a complex endeavor that many intervention research studies do not undertake (Mowbray et al., 2003; Nelson, 2013; O'Donnell, 2008). Following Kane (2006), who argued that researchers should construct an evidence-based argument for defending the appropriateness of a test for its intended uses, and the guidelines of the Standards for Educational and Psychological Testing that provide types of evidence relevant to the evaluation of an argument for the validity of an assessment procedure (*Standards:* AERA, APA, & NCME, 2014), this study describes a process for the validation of FOI observational measures and applies this process to a case study involving the IVC-R. The IVC-R is a classroom observation tool designed to measure the extent to which teacher and student enactment adhere in practice to the CSR model. Findings from the multistep process of item writing, scoring calibration, piloting, and revisions suggests that the IVC-R is a valid instrument for measuring adherence to CSR.

Results of this study reveal important considerations for the field regarding how to measure and analyze FOI using observational tools in an effort to better understand the translation of evidence-based intervention results into actual practice. Lessons learned therefore fall into one of two categories related (1) specifically to the case study presented (i.e., CSR) from data collected via the IVC-R instrument and (2) to the field in general learned from the process of developing a valid classroom observational measure of fidelity. In terms of the former, based on the data constructed through the IVC-R observational tool, results reveal that CSR may best fit a class schedule that extends beyond a 50-minute period; as such, efforts to clarify methods for shortening sessions without compromising quality may be an important next step in the program's professional development process. Further, greater focus on fostering in-depth discussions and providing feedback through additional professional learning sessions may provide the needed support for teachers who struggle with such aspects of CSR. Finally, while the 77-item IVC-R can be used to assess fidelity in future CSR evaluations, the instrument also provides examples of items that have been shown to validly measure specific, observable teacher and student behaviors or actions that can be adapted and tested in assessing fidelity to other instructional models aimed at improving achievement among middle school populations.

We also learned several lessons about what it takes to develop a valid instrument for observing teachers in action. One most evident point was the need for following an iterative approach, as contended by Messick (1988) and Kane (2006). While this article discusses the measurement properties of a specific instrument, results highlight the rigorous and ongoing process necessary for measuring fidelity to a treatment in education intervention studies. As such, this article presents a model that addresses a gap in the literature on methods for supporting claims about the validity of a fidelity observational tool through various sources of evidence. While there is general consensus among researchers that FOI is important, limited time and financial resources is likely the driving force behind this dearth of literature. Our study is the first we are aware of that provides a comprehensive example of how researchers who investigate education interventions can provide the systematic decision making and transparency needed for better understanding the effects of a treatment. In adapting models on creating high-quality educational measures (e.g., Kane, 2006; Wilson, 2004) to the domain of fidelity, and reporting on the multistep process that required frequent revisions and empirical assessment of the IVC-R used to assess FOI to CSR, this article makes more visible what is neglected in efficacy studies. It also provides a model that researchers can use to develop their own observational fidelity tools without having to start from scratch, thereby streamlining the process for validating FOI measures.

Specific to our case study, future investigations should continue to involve multiple aspects of fidelity to determine the strength of CSR and the implications of variation across classrooms. Now that there is a process for validating the IVC-R, the most important items can be identified to develop a shortened, more practical version of the instrument for regular use by school district personnel. An interesting area for research might then be to examine the costs versus benefits of the IVC-R compared with a pared-down version of the same instrument. On a more global level, however, we argue that all such intervention research should support and elevate issues related to validity and instrument design of fidelity instruments, thus holding the field more accountable to higher standards. A stronger focus on the validity of fidelity observational assessments will add nuance and clarity to such investigations.

FOI measures that demonstrate validity through various sources of evidence, and that have practical application, will help the field in making data-based decisions at the local, district, state, and federal levels. Finding ways to expand uses of FOI data could help justify the time and effort required to develop such instruments as well as the ongoing FOI data collection and analysis. For example, information can be used to provide feedback and information about implementation to teachers and to inform improvements in professional development as well as teacher and student resources. Intervention delivery must therefore be evaluated for fidelity to content and process so that one can explain whether failure to replicate designed outcomes is a function of the intervention or of its application. This distinction between intervention and application variation is not unique to education or any one field. As demand increases for evidence-based programs and policies, so does the expectation that service providers and organizations be held accountable for their outcomes (Schoenwald et al., 2011). However, fidelity is just one aspect of a complex system and should be used as a starting point to explore the reasons that a particular intervention may be easier to implement well, encouraging the field to explore competing demands on teachers and schools, limitations in time and resources, and more personal factors such as teaching style and philosophy.

Limitations and Future Research

Measuring treatment adherence and quality of delivery is challenging for any intervention when some items are inconsistently or rarely observed. In the case study, few teachers achieved a very high level of adherence to the CSR model, which is a limitation of this research. For instance, while most teachers taught Preview components (which are implemented at the beginning of CSR lessons and are also features of overall quality literacy instruction emphasized in the district in which the study was conducted), the majority failed to consistently implement the Questions and Review components that occur toward the end of the model. Other items were observed infrequently because they were more difficult (e.g., discussing and providing feedback). Additional research will be needed to understand how rarely observed items related to a particular intervention influence learning (in the case of CSR) or outcomes in general.

A second limitation is that FOI provides information about adherence to the intended model. As we have argued, accurate measurement of FOI is needed to draw conclusions about a program's outcomes. Still, we are cognizant of Guttiérez and Penuel's (2014) caution that

> "Scientifically rigorous research on what works in education requires sustained, direct, and systematic documentation of what takes place inside programs to document not only 'what happens'... but also how students and teachers change and adapt interventions in interactions with each other in relation to their dynamic local contexts" (p. 19).

The IVC-R was designed to assess features of the CSR model and was therefore not sensitive to adaptations. Future research should continue to expand on rigorous methods for measuring fidelity in ways that also capture localized adaptations.

Finally, we recognize that significant shifts in research planning and allocation of resources may be needed to meet the standards for validation reported here. Yet, we argue that attention to validation of FOI instruments allows for more complex ways of understanding how new practices are taken up in classrooms, and may thus allow us to learn more from the studies we are able to undertake. For example, the research presented in the case study included significant collaboration between measurement experts, researchers, program developers, and practitioners to plan how best to develop and validate the IVC-R. Such collaborations can only benefit the field as those with different expertise build on each other's knowledge and experiences. The approach outlined in this article serves as an example of a way in which FOI can be better understood. Future research should continue to expand the uses of fidelity data to inform design, implementation, and evaluation of classroom-based instructional programs.

Notes

The authors would like to acknowledge the substantial intellectual contribution of **Dr. Janette Klingner** to the development and study of CSR. Dr. Klingner's legacy in special education, reading comprehension strategy instruction for culturally and linguistically diverse students, and enhancing the sustainability of evidence-based and culturally responsive instructional practices will inform teachers and researchers for years to come. Dr. Janette Klingner passed away in March, 2014. We would also like to acknowledge **Dr. Sharon Vaughn** and **Dr. Greg Roberts** and their team at the University of Texas at Austin for their contribution to the study of CSR fidelity and the development of the original CSR IVC.

This research was supported by grant U396B100143, from Investing in Innovation, U.S. Department of Education. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Department of Education.

Underlying research materials related to this article can be accessed by contacting Dr. Pamela Buckley (Pamela.Buckley@colorado.edu).

References

Abbott, M. L., & Fouts, J. T. (2003). Constructivist teaching and student achievement: The results of a school-level classroom observation study in Washington. Lynnwood, WA: Washington School Research Center. Retrieved from: http://eric.ed.gov/?id=ED481694

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Belsky, J., Booth-LaForce, C. L., Bradley, R., Brownell, C. A., Burchinal, M., Campbell, S. B., & Weinraub, M. (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal*, *105*(3), 305–323. doi:10.1086/428746.

Boardman, A. G., Buckley, P., Vaughn, S., Roberts, G., Scornavacco, K., & Klingner, J. (2016a). The relationship between implementation of Collaborative Strategic Reading and student outcomes for adolescents with disabilities. *Journal of Learning Disabilities*, *49*(6), 644–657. doi:10.1177/0022219416640784.

Boardman, A. G., Klingner, J. K., Buckley, P., Annamma, S., & Lasser, C. J. (2015). Collaborative Strategic Reading in content classes: Results from year 1 of a randomized control trial. *Reading and Writing: An Interdisciplinary Journal*, *28*(9), 1257–1283. doi:10.1007/s11145-015-9570-3.

Boardman, A. G., Vaughn, S., Buckley, P., Reutebuch, C., Roberts, G., & Klingner, J. (2016b). Efficacy of Collaborative Strategic Reading with upper elementary school students: Results of a randomized control trial. *Exceptional Children*, *82*(4), 409–427. doi:10.1177/0014402915625067.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314. doi:10.1037/0033-2909.110.2.305.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavioral Interventions*, *12*, 161–179. doi:10.1177/1098300709334798.

Cappella, E., Kim, H.Y., Neal, J. W., & Jackson, D. R. (2013). Classroom peer relationships and behavioral engagement in elementary school: The role of social network equity. *American Journal of Community Psychology*, *52*, 367–379. doi:10.1007/s10464-013-9603-5.

Carroll, C., Patterson, M., Wood, S., Booth, A., Risk, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*, 40. doi:10.1186/1748-5908-2–40

Cavanagh, R.F. & Koehler, M.J. (2013) A turn toward specifying validity criteria in the measurement of Technological Pedagogical Content Knowledge (TPACK). *Journal of Research on Technology in Education*, *46*(2), 129–148. doi: 10.1080/15391523.2013.10782616.

Childs, K. E., Kincaid, D., George, H. P., & Gage, N. A. (2016). The relationship between school-wide implementation of positive behavior intervention and supports and student discipline outcomes. *Journal of Positive Behavior Interventions*, *18*(2), 89–99. doi:10.1177/1098300715590398.

Cohen, R., Kincaid, D., & Childs, K. E. (2007). Measuring School-wide Positive Behavior Support implementation: Development and validation of the

Benchmarks of Quality. *Journal of Positive Behavior Interventions*, *9*(4), 203–213. doi:10.1177/10983007070090040301.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103–124). Washington, DC: American Psychological Association.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45. doi:10.1016/S0272-7358(97)00043-3.

Devine, S. L., Rapp, J. T., Testa, J. R., Henrickson, M. L., & Schnerch, G. (2011). Detecting changes in simulated events using partial-interval recording and momentary time sampling III: Evaluating sensitivity as a function of session length. *Behavioral Interventions*, *26*, 103–124. doi:10.1002/bin.328.

Drake, R. E., & Resnick, S. G. (1998). Models of community care for severe mental illness: A review of research on case management. *Schizophrenia Bulletin*, *24*, 37–43. doi:https://doi.org/10.1093/oxfordjournals.schbul.a033314.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174. doi: 10.1037/1082-989X.5.2.155.

Fish, M. C., & Dane, E. (2000). The classroom systems observation scale: Development of an instrument to assess classrooms using a systems perspective. *Learning Environments Research*, *3*, 67–92. doi:10.1023/A:1009979122896.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906–911. doi:10.1037/0003-066X.34.10.906.

Flower, A., McKenna, J., Muething, C. S., Bryant, D. P., & Bryant, B. R. (2014). Effects of the good behavior game on classwide off-task behavior in a high school basic algebra resource classroom. *Behavior Modification*, *38*(1), 45–68. doi:10.1177/0145445513507574.

Freeman, J., Simonsen, B., McCoach, D. B., Sugai, G., Lombardi, A., & Horner, R. (2016). Relationship between School-Wide Positive Behavior Interventions and Supports and academic, attendance, and behavior outcomes in high schools. *Journal of Positive Behavior Interventions*, *18*(1), 41–51. doi:10.1177/1098300715580992.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945. doi: 10.3102/00028312038004915.

George, H. P., & Childs, K. E. (2012). Evaluating implementation of schoolwide behavior support: Are we doing it well? *Preventing School Failure*, *56*, 197–206. doi:10.1080/1045988X.2011.645909.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: systematic review and recommendations. *Milbank Quarterly*, *82*(4), 581–629. doi:10.1111/j.0887-378X.2004.00325.x.

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, *43*, 19–23. doi:10.3102/0013189X13520289.

Guo, Y., Connor, C. M., Yang, Y, Roehrig, A. D., & Morrison, F. J. (2012). The effects of teacher qualification, teacher self-efficacy, and classroom practices on fifth graders' literacy outcomes, *The Elementary School Journal*, *113*(1), 3–24. doi:10.1086/665816.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. New York, NY: Foundation for Child Development. Retrieved from http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf.

Heck, S., Steigelbauer, S. M., Hall, G. E., & Loucks, S. F. (1981). *Measuring innovation configurations: Procedures and applications*. Austin, TX. Research and Development Center for Teacher Education, University of Texas.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56–64. doi:10.3102/0013189X12437203.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371–384. doi:10.17763/haer.83.2.d11511403715u376.

Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The School-wide Evaluation Tool (SET): A research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions*, *6*(1), 3–12. doi:10.1177/10983007040060010201.

Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for school-wide positive behavior support. *Focus on Exceptionality*, *42*(8), 1–14.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research and Educational Effectiveness*, *2*, 88–110. doi:10.1080/19345740802539325.

Johnson, D. W., & Johnson, R. T. (1999). Making cooperative learning work. *Theory into Practice*, *38*(2), 67–73. doi:10.1080/00405849909543834.

Kagan, S. (1986). Cooperative learning and sociocultural factors in schooling. In Bilingual Education Office, California State Department of Education (Ed.), *Beyond language: Social and cultural factors in schooling language minority students*. Los Angeles, CA: Evaluation, Dissemination and Assessment Center, California State University.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Santa Barbara, CA: Greenwood Publishing Group.

Kazdin, A. E., & Kendall, P. C. (1998). Current progress and future plans for developing effective treatments: Comment and perspectives. *Journal of Clinical Child Psychology*, *60*, 733–747. doi:10.1207/s15374424jccp2702_8.

Kim, A. H., Vaughn, S., Klingner, J. K., Woodruff, A. L., Klein, C., & Kouzekanani, K. (2006). Improving the reading comprehension of middle school students with disabilities through Computer-Assisted Collaborative Strategic Reading (CACSR). *Remedial and Special Education*, *27*, 235–248. doi:10.1177/07419325060270040401.

Klingner, J. K., Vaughn, S., Arguelles, M. E., Hughes, M. T., & Leftwich, S.A. (2004). Collaborative Strategic Reading: "Real world" lessons from classroom teachers. *Remedial and Special Education*, *25*, 291–302. doi:10.1177/07419325040250050301.

Klingner, J. K., Vaughn, S., Boardman, A. G., & Swanson, E. (2012). *Now we get it! Boosting comprehension with Collaborative Strategic Reading*. San Francisco, CA: Jossey Bass.

Klingner, J. K., Vaughn, S., & Schumm, J. S. (1998). Collaborative Strategic Reading during social studies in heterogeneous fourth-grade classrooms. *Elementary School Journal*, *99*, 3–22. doi:http://dx.doi.org/10.1086/461914.

Lawrence, J. F., & Snow, C. (2010). Oral discourse and reading. In M. Kamil, P. D. Pearson, E. B. Moje, & P. Afflerbach (Eds.), *Handbook of reading research* (Vol. *4*). Mahwah, NJ: Erlbaum.

Lewis, T. J., & Sugai, G. (1999). Effective behavior support: A systems approach to proactive schoolwide management. *Focus on Exceptional Children*, *31*(6), 1.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates–MacGinitie Reading Tests* (4th ed.). Itasca, IL: Riverside.

McKennel, A. C. (1974). Surveying attitude structures. Amsterdam, the Netherlands: Elsevier

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–48). Hillsdale, NJ: Erlbaum.

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

Munter, C., Wilhelm, A. G., Cobb, P., & Cordray, D. S. (2014). Assessing fidelity of implementation of an unprescribed, diagnostic mathematics intervention, *Journal of Research on Educational Effectiveness*, *7*(1), 83–113. doi:10.1080/19345747.2013.809177.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research*, *39*(4), 374–396. doi:10.1007/s11414-012-9295-x.

Nelson, M. C. (2013). *New tools for intervention fidelity assessment* (Doctoral dissertation, Vanderbilt University). Retrieved from http://etd.library.vanderbilt.edu/available/etd-04032013-015440/unrestricted/mnelson.pdf.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relation to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*, 33–84. doi:10.3102/0034654307313793.

Palincsar, A. S., & Brown, A. L. (1984). The reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*, 117–175. doi:10.1207/s1532690xci0102_1.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. doi:10.3102/0013189X09332374.

Piburn, M., & Sawada, D. (2000). Reformed Teaching Observation Protocol (RTOP) reference manual. Retrieved from: http://files.eric.ed.gov/fulltext/ED447205.pdf.

Raudenbush, S. W., Bryk, T., & Congdon, R. (2010). HLM 7 hierarchical linear and nonlinear modeling. [Computer software]. Skokie, IL: Scientific Software International.

Raudenbush, S. W., & Bryk, T (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rogers, E. (2003). *Diffusion of innovations*. New York, NY: Free Press.

Rosenbaum, P. R. (1995). *Observational Studies*. New York, NY: Springer-Verlag.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *Measuring Reform Practices*, *10*(6), 245–253. doi:10.1111/j.1949-8594.2002.tb17883.x.

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). Reading interventions for adolescent struggling readers: A meta-analysis with implications for practice. Portsmouth, NH: RMC Research

Corporation, Center on Instruction. Retrieved from http://files.eric.ed.gov/fulltext/ED521837.pdf.

Schechter, S., Blair, J., & Hey, J. V. (1996). Conducting cognitive interviews to test self-administered and telephone surveys: Which methods should we use? In 1996 *Proceedings of the Section on Survey Research Methods* (pp. 10–17). Alexandria, VA: American Statistical Association.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. (2011). Toward the effect and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1): 32–43. doi:10.1007/s10488-010-0321-0.

Shulman, L. (1990). Foreword. In M. Ben-Peretz, *The teacher–curriculum encounter: Freeing teachers from the tyranny of texts*. Albany, NY: State University of New York Press.

Tiger, J. H., Miller, S. J., Mevers, J. L., Mintz, J., Scheithauer, M. C., & Alvarez, J. (2013). On the representativeness of behavior samples in classrooms. *Journal of Applied Behavior Analysis*, *46*, 424–435. doi:10.1002/jaba.39

Vaughn, S., Klingner, J., Swanson, E. A., Boardman, A. G., Roberts, G., Mohammed, S., & Stillman-Spisak, S. J. (2011). Efficacy of collaborative strategic reading with middle school students. *American Educational Research Journal*, *48*, 938–964. doi:10.3102/0002831211410305.

Vaughn, S., Roberts, G., Klingner, J., Swanson, E., Boardman, A., Stillman-Spisak, S.J., Mohammed, S., & Leroux, A. (2013). Collaborative strategic reading: Findings from experienced implementers. *Journal of Research on Educational Effectiveness*, *6*(2), 137–163. doi:10.1080/19345747.2012.741661.

Wanzek, J., Vaughn, S., Kent, S., Swanson, E. A., Roberts, G, Haynes, M., . . . Solis, M. (2014). The effects of team-based learning on social studies knowledge acquisition in high school. *Journal of Research on Educational Effectiveness*, *7*, 183–204. doi:doi:10.1080/19345747.2013.836765.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. New York, NY: Routledge.

Manuscript received January 28, 2016
Final revision received June 22, 2017
Accepted June 24, 2017