Whole-part-whole: Construct validity, measurement, and analytical issues for fidelity assessment in education ....

Chapter · January 2013

CITATIONS
6

READS
139

3 authors, including:

Chris Hulleman
University of Virginia
48 PUBLICATIONS **2,297** CITATIONS

SEE PROFILE

Sara Rimm-Kaufman
University of Virginia
85 PUBLICATIONS **3,954** CITATIONS

SEE PROFILE

FINAL PRE-COPYEDITED VERSION

October 3, 2012

This chapter appears in:
Halle, T. G., Metz, A. J., & Martinez-Beck, I. (2013).
Applying implementation science in childhood settings.
Baltimore, MD: Brookes.

Chapter 3: Innovative Methodologies to Explore Implementation

Whole-Part-Whole: Construct Validity, Measurement, and Analytical Issues for Intervention Fidelity Assessment in Education Research

Chris S. Hulleman

James Madison University

Sara E. Rimm-Kaufman and Tashia Abry

University of Virginia

Author Note:
Correspondence should be addressed to Chris S. Hulleman, Center for the Advanced Study of Teaching and Learning, 2200 Old Ivy Road, Charlottesville, Virginia, 22903; Email: chris.hulleman@virginia.edu. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A070063 to the University of Virginia (PI: Sara Rimm-Kaufman). The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. Chris Hulleman is now at the University of Virginia.

Abstract

This chapter addresses the conceptualization of intervention fidelity and its implications for measurement, design, and analysis of educational interventions, including community-based early care and education and school-based programs. The authors will highlight a five-step model for assessing intervention fidelity and apply it to early care and education interventions. In particular, this chapter will focus on the *Responsive Classroom*® approach to elementary education. The *RC* approach focuses on the socio-emotional development of the learner as equal in importance to academic development. Various challenges in the field related to fidelity measurement, research design, and analysis will be highlighted. Examples include treatment strength (i.e., achieved relative strength), zone of tolerable adaptation, weighting of components of implementation within analyses, and sequencing of components. Key terminology will be introduced and explained.

I. Introduction

In the early 20th century, structuralism was a popular approach in psychology. Structuralists sought to understand human behavior by deconstructing the conscious experience into progressively smaller pieces. Their key methodology was introspection, which required the researcher to report on discrete aspects of their own thoughts and feelings. Although the structuralist approach was defined by methodological rigor, a group of German psychologists (i.e., the Gestaltists) took issue with its fundamental principle of breaking the whole into parts. Their viewpoint, led by Wertheimer (e.g., see Weiner, 1972 for a review), was that the whole was more than the sum of its parts. These Gestaltists believed that consciousness was best studied in its entirety, rather than by breaking it into finite pieces as advocated by the structuralists. By breaking consciousness into smaller pieces, the Gestaltists argued, the essence was lost. This fundamental disagreement between structuralists and Gestaltists is more than an historical footnote – it resurfaces as a contemporary measurement problem: How do we develop fine-grained measures that capture the breadth and depth of a construct, often by breaking it into pieces, without losing its essence? For example, by defining school quality in terms of domain-specific aspects (e.g., curriculum, environment, assessment, diversity; Tout, Starr, Soli, Moodie, Kirby, & Boller, 2010),and developing measures that focus on these aspects, do we lose the essence of what it means to have a quality learning environment?

This measurement challenge, which is most often associated with the assessment of constructs such as personality traits (e.g., extraversion) or affective states (e.g., test anxiety), is also relevant to implementation science in general, and assessment of intervention fidelity in particular. That is, measuring the extent to which an early care and education (ECE) program is implemented as intended (i.e., intervention fidelity) requires the same construct and

psychometric validity as do measures of personality constructs. For fidelity assessment, establishing construct validity means breaking down an intervention model into its core components (Whole-Part) and ensuring that the fidelity indices, when administered individually, sum to represent full-scale implementation (Part-Whole) without losing the essence of the intervention. This chapter presents a model of intervention fidelity assessment in the context of a school-based intervention, with the aim of demonstrating to researchers and evaluators that this measurement challenge can be overcome by developing valid measures of intervention fidelity. To achieve this aim, we selected an intervention, the *Responsive Classroom®(RC)* approach, focused on the improvement of elementary school classrooms that exemplifies some distinct challenges associated with measuring Whole-Part and Part-Whole construct validity.

The same principles and challenges also apply to all types of interventions, including those focusing on early care and education, such as community-based home healthcare and school-based behavioral health programs (Lang & Franks, 2011; Lowell et al., 2010), and to the measurement of key implementation supports (i.e., implementation drivers). Although this chapter will use an elementary education example to examine issues of construct validity in intervention fidelity assessment, linkages to early care and education environments and implementation science in general will be explicated throughout.

II. Are Schools Ready for Kids? The *Responsive Classroom®(RC)* Approach

Developed by the Northeast Foundation for Children, the *RC* approach is designed for K-6 classrooms and is widely disseminated in the United States with over 10,000 teachers trained annually (Zubrzycki, 2012). The approach focuses on integrating children's social and academic learning in elementary school through 7 core principles and 10 practices (see Table 1). Examples of principles include: "the social curriculum is as important as the academic

curriculum; how children learn is as important as what children learn; and knowing the children we teach—individually, culturally, and developmentally—is as important as knowing the content we teach (Northeast Foundation For Children [NEFC], 2007)." *RC* practices provide a set of strategies designed to help teachers enact the developmentally-oriented recommendations embodied in the *RC* principles.

INSERT TABLE 1 ABOUT HERE

For example, the Morning Meeting consists of a daily gathering as a class when the teachers and children greet each other, share personal news in a structured way, engage in a lively and playful activity together, and prepare for the day ahead. Rule Creation involves a set of practices that engage students in the process of creating the classroom rules so that all children can meet their classroom learning goals. Academic Choice is an approach to instruction where the teacher determines the learning objectives and then provides students structured choices as to how to meet those objectives. In an Academic Choice lesson, students choose from several approaches to conduct their work, describe their plan to their teacher, engage in their work, and then reflect (in writing or to the group) upon their new learning. Guided Discovery is an approach to introducing materials in an organized manner that fosters creative and careful use of those materials.

In essence, the *RC* approach is designed to offer teachers a set of principles and strategies to build teacher capacity and support the quality of their interactions with children. Two hallmark characteristics stand out in describing teachers who fully implement the *RC* approach: 1) teachers foster the development of a caring classroom community, and 2) teachers emphasize proactive rather than reactive approaches to classroom management and discipline. A core premise of the *RC* approach is that an environment where children's socio-emotional needs are

recognized and met sets the stage for academic learning. This framework aligns well with early childhood theory and resembles programmatic efforts, such as Head Start, that emphasize emotional and cognitive development and focus broadly on the needs of the whole child.

The *RC* approach offers some unique challenges to measuring intervention fidelity with validity. First, many of the practices have features that resemble those used in typical classrooms. For example, many early childhood teachers use a class meeting format in the beginning of the day. Although elements of such class meetings may resemble the Morning Meeting, there are important differences that differentiate high from low intervention fidelity of the *RC* approach. As an example, many morning meetings offer routine approaches to discussing the date, weather, and plan for the day. In contrast, the *RC* Morning Meeting involves a greeting in which each child is greeted by name and a group activity designed to be fun and engaging and help children feel a connection toward their peers and toward school.

Second, the *RC* approach, at its core, is comprised of a set of practices. However, simply implementing those practices without adhering to the core beliefs about children represented in the *RC* principles may not represent the intended effect of the *RC* approach. For instance, a Morning Meeting that is rushed and mechanical as opposed to a Morning Meeting that draws teachers and children together as a community represent two different levels of intervention fidelity. This illustrates the challenge of obtaining construct validity in fidelity assessment. In addition, by measuring *RC* practices with very exact language and precise measures, we may not adequately reflect comparable practices that are present in control groups. If we break apart assessment of the *RC* approach into key pieces, then we may lose the essence of creating a warm, safe learning environment to maximize learning for all children. For these, as well as other

reasons, the *RC* approach offers a useful exemplar for understanding issues of construct validity when measuring intervention fidelity.

III. Fidelity Definitions and Construct Validity

In intervention research, the primary research question is usually whether the program had its intended effects. In a randomized field trial, this is the intent-to-treat analysis: Did the program work for those who were assigned to the treatment compared to those who were not assigned to the treatment? Policy decisions are often made at this level of analysis. What this evaluation misses, however, is whether any effects (or lack thereof) were due to the program as intended or the program as *implemented*. To the extent that a program is implemented with less than complete fidelity, the causal inference about program effectiveness is in regards to the program as implemented, particularly as contrasted with the comparison condition. Attributing a lack of program impacts to a program that was not well implemented is to commit a Type III error (Dobson & Cook, 1980), and risks discarding a potentially effective intervention. Thus, a critical question to be answered when evaluating program effectiveness is to what extent was the program implemented as intended, and did variations in implementation impact the outcomes?

Intervention fidelity is the extent to which the program, as designed, was actually implemented. Crucial to fidelity assessment is the identification of the *intervention core components* that comprise the intervention model (also called essential functions or active ingredients). These are the key conceptual aspects of the intervention that must be in place in order for the intervention to be implemented. Intervention components are defined through the use of conceptual logic models (see Section IV). For example, in the *RC* approach, the 10 practices outlined in Table 1 are the core intervention components. Our definition of intervention fidelity is different from *implementation* fidelity, which involves the contextual factors that

support the implementation of the intervention core components, such as staff selection, administrative training, and the provision of resources. These *implementation core components* (Fixsen et al., 2005) are not part of the intervention model per se, but rather help support the implementation of the intervention core components. In the case of the *RC* approach, creating a school-wide time for Morning Meeting is not one of the 10 *RC* practices outlined in Table 1, but could facilitate the implementation of the these practices. Thus, planning time facilitates implementation of *RC* practices, but is not an intervention core component because planning time is not explicitly part of the program model.

The distinction between intervention fidelity and implementation fidelity nearly parallels the differentiation between core components that have direct versus indirect effects on children. For example, teachers play the primary role in the implementation of the *RC* approach as defined by the core components of using *RC* practices in classroom. As such, we are focusing on the *direct* effects of *RC* practices (implemented by the teacher) on children. At a broader, contextually-based level, are implementation core components that create essential conditions that support teacher implementation of *RC* practices, but are not core components of the theorized model. These implementation core components (Fixsen et al., 2005), which include administrative support and training, selection of staff, and teacher coaching, have an *indirect* effect on children as mediated by the intervention core components. That is, administrative support can provide a facilitative condition for the implementation of *RC* practices, but it is the *RC* practices themselves that are theorized to directly enhance student outcomes. Thus, the implementation core components can be referred to as *implementation drivers* (Fixsen, Blase, Naoom, & Wallace, 2009). The caveat to this distinction is the case in which the contextual level is specifically a part of the intervention model; then, those aspects become intervention core

components (see Chapter 7 in this volume for an example of coaching as a core intervention component of ECE models).

Drivers of effective implementation are important to understanding whether interventions work in a variety of settings, and are the focus of other work (e.g., Metz, Halle, and Bartley, this volume; Wanless, Patton, Rimm-Kaufman, & Deutsch, 2012). In contrast, we focus on the assessment of intervention core components. Intervention fidelity assessment helps us understand how much *actual* difference in causal elements exists post-implementation. The actual difference in core intervention components, or *achieved relative strength* of the intervention (Cordray& Pion, 2006; Hulleman &Cordray, 2009), determines intervention outcomes and is what needs to be measured in both treatment and control conditions. In the case of the *RC* approach, one treatment teacher may implement six of the ten practices as intended by the developer, whereas in the control condition, three of the ten practices may be present in control classrooms. This would mean that the achieved relative strength of the intervention for this *RC* teacher was only 3 components, instead of a possible 10. Due to our focus on intervention fidelity and intervention core components, in the remainder of the chapter any reference to core components will refer to intervention core components.

Although definitions of intervention fidelity vary throughout the literature (for a review see O'Donnell, 2008), fidelity is generally composed of five dimensions: exposure (to what extent are participants exposed to the treatment), adherence (were treatment components delivered as designed), quality (how well was the treatment implemented), responsiveness (to what extent are participants engaged and involved in the treatment), and differentiation (whether critical program components that differentiate treatment from control are present). Responsibility for treatment fidelity within the five dimensions can be due to implementers of the intervention (such as early

care providers or teachers), participants (such as children, in an intervention with teachers as implementers), or both (Hulleman & Cordray, 2009). Importantly, these five dimensions are meant to provide a framework for fidelity assessment that enables the evaluator to capture the whole of the intervention and not simply disjointed pieces. Other fidelity frameworks that use slightly different language can be equally as helpful; the important point is to measure fidelity across different dimensions of implementation.

There is ample evidence within health efficacy and effectiveness studies that variation in treatment fidelity accounts for variability in outcomes (e.g., Dane & Schneider., 1998; Mitchel et al., 1984; Tortu & Botvin, 1989). Increasingly, evidence within early care and education research demonstrates that intervention fidelity can explain some of the variation in treatment effectiveness (e.g., Durlak & DuPre, 2008; Durlak, Weissberg, et al, 2011; Franks & Schroeder, this volume; McIntyre, Gresham, DiGennaro, & Reed, 2007; Nunnery et al., 1997).

In order to be able to assess whether a program was implemented, the conceptual foundation of the program must be elucidated, along with the essential components required for proper implementation. This involves breaking down an intervention model into its core components and ensuring that the fidelity indices, when administered individually, can be combined to represent full-scale classroom implementation without losing the essence of the intervention. Thus, the measurement of fidelity is, at its core, an exercise in construct validity. In essence, this is the Whole-Part-Whole tension present between the structuralists and Gestaltists. The aim of this chapter is to demonstrate that this measurement challenge can be overcome by developing valid measures of intervention fidelity within the context of a school-based intervention. We apply the five-step model of fidelity assessment developed by Nelson, Cordray, Hulleman, Darrow, and Sommer (2012) to the *Responsive Classroom® (RC)* approach. The first step of this

process is to develop systematic and detailed representations of the program (i.e., logic models) to guide fidelity assessment.

IV. The First Step in Fidelity Assessment: Logic Models

Logic models are a common method for explicating program elements. As defined by the Kellogg Foundation (2004), a logic model visually represents the relationships between the resources required, activities to be put in place, and outcomes that occur as a result of program implementation. Knowlton and Phillips (2009) further differentiate logic models into two distinct types that pertain to fidelity assessment: a conceptual model and an operational model. The conceptual logic model, or theory of change, represents program elements in conceptual terms by describing the constructs that underlie activities and resources. The operational logic model represents program elements in practical terms by describing the activities and resources involved in implementing the intervention.

First, the core intervention components that form the foundation of the theory of change are specified (see Figure 1).The theory of change lays out the *sequence* of intervention components and subsequent outcomes expected under ideal conditions. Each intervention component is described in terms of constructs, and the sequence of how constructs are related to each other, and to proximal and distal outcomes (i.e., outputs and outcomes, respectively; Lugo-Gil, Sattar, Ross, Boller, Kirby, & Tout, 2011), corresponds to the theoretical intervention model. Figure 1 presents the conceptual logic model for the *RC* approach. The initial step in creating change is the *Training* teachers receive in how to use *RC* practices in the classroom. Next, teachers need to implement the *Practices* in the classroom. These two intervention components, *Training and Practices*, set into motion two classroom-level Mediators: improved classroom processes, and

enhanced student engagement in school. These Mediators results in student outcomes such as improved social skills and achievement.

INSERT FIGURE 1 ABOUT HERE

Next, the operational logic model specifies program implementation (i.e., who, what, where, how, and when) and describes how program activities are mapped onto the conceptual intervention core components from the theory of change. These operational core components are the specific teacher behaviors that represent the theorized core intervention components. As presented in Figure 2, *RC* implementation begins with the planned work, including Resources and Inputs, Activities, and School and Classroom Outputs. The presence of Resources and Inputs occurs first and includes elements of the NEFC training infrastructure, including consulting teachers and coaches, manuals and books, as well as other multi-media resources. Next, Activities include training in the *RC* approach, involving weeklong training sessions and follow-up coaching as key intervention components. Training sessions involve certified *RC* consulting teachers who provide week-long training workshops in the first (*RC* 1) and second (*RC2*) summer, along with in-school, follow-up coaching. The final part of implementation, School and Classroom Outputs, refer to the use of *RC* practices in classrooms. Specifically, use of *RC* practices describes teachers' frequency of use and adherence to the ten *RC* practices including Morning Meeting, Rule Creation, Interactive Modeling, Positive Teacher Language, among others.

INSERT FIGURE 2 ABOUT HERE

The Planned Work should ideally lead to Intended Results, including School and Classroom Outcomes, Child Outcomes, and Impact. School and Classroom Outcomes are processes induced by the use of *RC* practices, such as enhanced sense of community in the school, improved

emotional support (i.e., more warmth, caring and responsiveness among teachers and children), improved quality of classroom management, and higher quality instruction. These processes should then increase children's engagement in learning, self-regulatory skills, and school bonding. Finally, under ideal conditions, program Impact will be evidenced by improved student social skills and achievement.

When specifying the model, it is important to consider the level at which the intervention is enacted. Many school-based interventions can be enacted district-wide, school-wide, or by an individual teacher (Greenberg, 2010). For example, the *RC* approach is often implemented school-wide or by an individual teacher at a school. School-wide implementation implies that virtually all administrators, teachers, and other school staff receive some training in the *RC* approach. Subsequently, the *RC* approach is leveraged to build community and improve climate within the whole school through school-wide meetings, and consistency in discipline across all of the adults who interact with children (e.g., music teachers, art teachers, cafeteria staff). Thus, school-wide change results from synergistic effects that emerge when *RC* administrators and *RC* teachers come together to create changes in school culture. In fact, school level change can be viewed as an outcome unto itself for schools using *RC* practices to achieve improvement to the school climate, guided by coaching and materials aimed toward school administrators (Casto & Audley, 2008; Northeast Foundation for Children, 2006). Similar to measuring implementation of an evidenced-base practice on a large scale, intervention fidelity would be measured in relation to efforts to change the broader adult community, such as interactions between implementers, or implementers and administrators. However, the difference between implementation assessment and fidelity assessment is that the former is focused on the drivers of

implementation at scale whereas the latter is focused on measuring core components of the intervention model.

In contrast, many school-based interventions are enacted almost exclusively by individual teachers. As a result, the processes and behaviors most pertinent to fidelity assessment shifts to the classroom. Several options are available for measuring intervention fidelity when teachers are implementers (e.g., individual teachers at the school become *RC* teachers). On one hand, intervention fidelity can be measured by considering the quality of the weeklong workshops and the frequency and quality of the follow-up coaching (i.e., Activities that are a part of the Planned Work, see Figure 2). On the other hand, intervention fidelity can be measured by assessing the teachers' use of practices in the classroom (i.e., School and Classroom Outputs, see Figure 2).

This discussion of level of implementation highlights an important first objective in considering measurement of intervention fidelity with validity: defining the implementers and the outcome(s) of interest. One of the core premises of the *RC* approach is that student change occurs through the experience of *RC* practices. Intervention fidelity at this level assesses *RC* teaching practices or student responsiveness to classroom practices. This localizes the main thrust of intervention fidelity assessment at the classroom level. However, other interventions that are enacted at the school or district level will therefore focus fidelity assessment accordingly. However, this does not mean that the *RC* approach does not have school or district level components. Rather, this means fidelity measurement will focus on teachers and the practices they use to interact with children. For brevity, we focus on teachers' implementation of *RC* practices in this chapter. Other work (e.g., Wanless et al., 2012) considers school-level predictors of the use of the *RC* approach.

V. A 5-Step Model of Intervention Fidelity Assessment

Specifying the core conceptual and operational program components is the first step in a five-step process that we advocate for assessing intervention fidelity with construct validity (Cordray & Hulleman, 2009; Nelson et al., 2012). The other four steps are: 2) Identify appropriate fidelity indicators, 3) Determine index validity (including reliability), 4) Combine indices where appropriate, and 5) Link fidelity to outcomes where possible.

In Step 2, the conceptual and operational logic models focus assessment on the specific actions required to implement the intervention core components. Common methods used to measure fidelity include surveys, observations, and interviews. To capture a comprehensive and rich picture of implementation , we recommend that measures of each core component include as many dimensions of fidelity as possible (adherence, exposure, quality, responsiveness, and differentiation). If some core components are measured more with adherence measures, whereas others include primarily measures of quality, then any differences in their relationship with outcomes may be due to the content of the items, the type of fidelity, or both. In the *RC* example below, the fidelity measures were equivalently focused on exposure and adherence across core intervention components.

Before being incorporated within the analysis of treatment variation, reliability and validity evidence should be gathered for the different fidelity measures (Step 3). Reliability involves the extent to which the fidelity measures inter-correlate. For instance, we would expect that teacher self-reports of *RC* practices would correlate with classroom observations. Validity involves the extent to which the fidelity instruments measure what is intended (i.e., construct validity). For instance, assessing the frequency of the Morning Meeting (adherence), but not the quality of delivery (quality), may fail to cover the breadth of the construct (Shadish, Cook, & Campbell, 2002).If several fidelity measures are highly interrelated, they can be combined into composite

fidelity indices (Step 4). Composite fidelity indices are useful, as they can be more easily utilized to determine achieved relative strength of the intervention (Abry, Rimm-Kaufman, Hulleman, Thomas, & Ko, 2012; Hulleman & Cordray, 2009). That is, composite indices more succinctly quantify overall fidelity, which can then be used to link fidelity to outcomes in a simpler fashion. In addition, creating indices that represent individual core components are helpful in assessing the extent to which individual components have been implemented. Once the appropriate fidelity measures have been combined, fidelity can be linked to outcome measures using both descriptive and inferential statistics (Step 5). It is possible that fidelity for some core components is more strongly linked to outcomes than others. This can help identify core (i.e., vs. ancillary) intervention components, and provide clues as to how best adjust the intervention to be more effective and adjust the theoretical model. For example, if fidelity measures of 2 of the 10 *RC* practices were strongly related to outcomes, whereas the others were unrelated, this might lead to changes in the *RC* model and how heavily some components are emphasized compared to others.

VI. Application of the 5-Step Model of Fidelity Assessment: The *Responsive Classroom* Efficacy Study

In this section, we use results from the Responsive Classroom Efficacy Study (Rimm-Kaufman, Fan, Berry, Justice, 2011) to highlight various issues involving model specification (Step 1), fidelity measurement (Step 2), determining indices of intervention fidelity (Steps 3-4), and linking fidelity to intervention outcomes (Step 5). For brevity, we have limited our scope to examining intervention fidelity of teachers' use of *RC* practices in the classroom, thus considering the direct effect of the intervention on students.

Step 1: Model Specification

The Responsive Classroom Efficacy Study (RCES) research team began to develop intervention fidelity measures meeting construct validity and data collection feasibility objectives. RCES was the second study of the *RC* approach conducted by Rimm-Kaufman et al., thus the research group leveraged lessons learned in the early efficacy study to hone measures (Rimm-Kaufman & Sawyer, 2004; Rimm-Kaufman et al., 2007). The research team is comprised of independent evaluators of the *RC* approach, and thus, were able to take an objective stance toward assessing the intervention. However, being independent evaluators posed a challenge in understanding intervention fidelity, requiring the research team to become immersed in the approach and establish a collaborative relationship with NEFC while also maintaining its objective stance.

Several steps were necessary to articulate the logic model and develop our measures. First, researchers on the measurement development team attended *RC* workshops. Second, the researchers engaged in thorough review of *RC* practices through the *RC* books and materials available. Third, we engaged several people from the NEFC to provide feedback to us on measurement development. Fourth, realizing that measurement development would be an iterative process, we held weekly conversations with the NEFC developers to be sure that we were measuring the presence of *RC* practices without losing the essence of the *RC* approach. Finally, we coordinated our efforts with existing efforts at NEFC. Specifically, the NEFC team was developing a measure designed to be used in school so that school administrators could measure the presence or absence of *RC* practices (Wilson, Freeman-Loftis, Sawyer, &Denton, 2009), and thus, our dialogue was effective in both directions. Throughout the process, the research team considered the feasibility of the proposed measurement process in relation to resources available through the study.

Step 2: Developing Fidelity Indicators

Our research team developed three distinct measures of intervention fidelity, focusing mainly on exposure and adherence. The *Classroom Practices Teacher Survey* (CPTS) is a 46-item teacher-report measure ($\alpha = 0.91$) measuring teachers' perception of their use of *RC* practices. The CPTS was designed to measure exposure and adherence. Teachers rate their use of specific classroom practice on a five-point Likert scale, ranging from 1 (*not at all characteristic*) to 5(*extremely characteristic*). Each item asks about use of *RC* practices, however, the items are phrased carefully to avoid use of *RC* language. Further, seven of the items were reversed scored to further reduce biased responding. For example, to measure Morning Meeting, teachers were asked, "In the morning, we have a class meeting where we sit in a circle facing one another." To measure Academic Choice, teachers were asked, "When my students are working on activities of their own choosing, I have structures in place that assist them in planning their activity."

The *Classroom Practice Frequency Survey* (CPFS) is an 11-item teacher-report measure ($\alpha = .89$) assessing the frequency of teachers' use of *RC* practices, thus assessing students' exposure to the intervention. Teachers are asked to recall their use of specific practices. Again, the practices are described to avoid use of *RC* language. For example, to measure Academic Choice, teachers were asked, "I provide opportunities for student to choose how to do work, what kind of work to do, or both (e.g., in studying marine biology, students may choose the animal they want to study and/or students can demonstrate knowledge about this animal through writing a report, drawing a picture book, crafting a clay model)." Teachers report their responses on an 8-point scale ranging from 1 (*almost never*) to 8 (*more than once per day*).

The *Classroom Practices Observation Measure (CPOM)*is a 16-item observational measure ($\alpha = .88$) of exposure and adherence to the intervention core components. The measure described

RC practices without using *RC* terminology to ensure that the measure could be used in intervention and control classrooms and because the classroom observers had not been trained in the *RC* approach. For example, to measure Rule Creation one item stated, "Three to five general, positively worded rules are posted in the classroom." To measure Positive Teacher Language, one item stated, "Teacher asks questions or makes statements that invite students to remember expected behaviors." Each item was coded on a three-point Likert scale ranging from not at all characteristic to very characteristic. A 16-item version was administered during morning observations and an abbreviated 10-item version that excluded Morning Meeting items were used during observations conducted during mathematics instruction. Two items were reverse scored. An extensive training process was developed to establish and maintain coder reliability, as described in Abry, Rimm-Kaufman, Larsen & Brewer (2012).

Table 2 lists the 10 *RC* intervention core components and total number of indicators across the three types of measures developed by Rimm-Kaufman and colleagues (Rimm-Kaufman, Berry, Fan, McCracken, & Walkowiak, 2008).

INSERT TABLE 2 HERE

Once the measures have been developed, it is essential to work backwards from the indicators to the construct (i.e., the intervention model) to determine whether these constructs fully capture the treatment. Crucially, the essence of the intervention needs to be represented. If not, it is possible that by breaking the intervention model into pieces that the big picture, or Gestalt, of the intervention will be lost. One way of ensuring that this validity check occurs is for the instrument developers to check back with program developers (or implementers) after the fidelity instrument has been developed. Not only can program developers consider whether any important components are missing, or if there extraneous components that could be eliminated,

but they could also consider two different scenarios. First, developers or implementers could consider whether a teacher who scores highly on these measures would be considered an *RC* teacher whereas a teacher who scores low on these measures would not be an *RC* teacher. Conversely, they could also envision the practices of the ideal *RC* teacher and consider the extent to which the measure adequately taps these practices. The answers to these questions will help instrument developers understand whether their fidelity measures adequately map back not only to the individual core components, but also to the entirety of the intervention they are intending to measure.

Steps 3 & 4: Determining Index Validity and Reliability

Items and indices were combined to measure fidelity to core intervention components and to quantify achieved relative strength of the intervention. Indices were created to correspond to several core components. First, we determined how to combine items across different types of measures. There are several possible methods to combine indicators into indices. The most basic combination is to create composite fidelity indices by taking a variety of different measures (teacher report, self-report) and combining them into one overall indicator of intervention fidelity. A second approach would be to create indices based on the measure. That would involve creating one fidelity index for the teacher self-report measure and a second for the classroom observation measure. A third approach would be to combine items in meaningful ways across the different types of measures. This would mean taking all items pertaining to a specific core component and combining them into an index that represents fidelity to that core component. We recommend this third approach for several reasons (c.f., Abry, Rimm-Kaufman, Hulleman, Thomas, &Ko, 2012). Creating indices by core components allows the researcher to examine fidelity to each core component, as well as overall fidelity (if the separate indices are combined).

Core component indices support efforts to diagnose features of the intervention model that are easier or more difficult to implement. The core component approach allows the researcher to examine the construct validity of the intervention model and its core components rather than only providing a diffuse measure of overall fidelity produced by the first two approaches.

In the *RC* data, there is one observational measure (CPOM) and two teacher self-report measures (CPFS, CPTS). Because the measures were on different scales, items were first standardized across the sample, and then relevant items from across the measurement instruments were combined to form the index for each core component. For this chapter, we focused on three core *RC* practices: Morning Meeting (10 items), Academic Choice (8 items), and Interactive Modeling (5items). Importantly, these measures focused on the levels of adherence and exposure of the intervention, as opposed to quality, responsiveness, and differentiation, which constrains our inferences about implementation to these dimensions of fidelity.

Table 3 presents the number of items from each type of measure that contributed to the fidelity indices for each core component. Each *RC* practice (i.e., intervention component) was broken down further into sub-components. For Interactive Modeling, key sub-components of this practice include the teacher demonstrating the skill or activity, students observing teacher demonstrations, and students practicing the skill or activity. Importantly, not every sub-component was measured with each type of measure, nor was every core component measured with an equal number of items. This resulted in some fidelity indices containing more of one type of measure than another, and with some core components containing significantly more items than others (e.g., 25 items for Morning Meeting compared to 7 for Interactive Modeling).

INSERT TABLE 3 ABOUT HERE

Achieved Relative Strength Indices(ARSI). Once indices for each core component have been created, they can be used to compare the relative strength in intervention components between treatment and counterfactual conditions. As mentioned earlier, achieved relative strength refers to the difference between the actual implementation of the intervention in the treatment group and the actual implementation of the intervention in the control group. Using achieved relative strength means that "the estimates of effects on the outcome are the result of the achieved relative strength of the contrast between treatment and control, not the theoretically expected difference," (Hulleman & Cordray, 2009, p. 91). The achieved relative strength values correspond to effect size values to aid interpretation. One of the challenges of ensuring construct validity is that many *RC* practices resemble practices used in typical classrooms. For instance, many teachers provide instructional choice to their students and may be creating the same advantages for children as *RC* teachers. Further, within *RC* classrooms, some teachers are not fully implementing the *RC* approach as intended. In this section we outline two types of achieved relative strength indices (ARSI): the average index and the binary complier index.

The average ARSI is computed by standardizing the average difference between fidelity indices from each condition (subtracting the mean scores and dividing by the pooled standard deviation; see Hulleman & Cordray, 2009 for details). Table 4 presents the average achieved relative strength index values for intervention and control teachers for 84 fourth-grade teachers at 24 schools (13 intervention, 11 control). Because the items were standardized across the entire sample, positive values are above the mean of the sample and negative values are below the mean of the sample. The ARSI column reveals that although intervention teachers are implementing all four components to a higher degree than control teachers on average, the achieved relative strength values differed across core components.

INSERT TABLE 4 ABOUT HERE

Morning Meeting shows the largest ARSI of over 2 standard deviations. It is comprised of four subcomponents including a greeting, group activity, sharing, and morning message. Aside from the sharing subcomponent, the other practices are very characteristic of *RC* classrooms and are seldom apparent in control classrooms.

The Academic Choice practice was substantially different between *RC* and control conditions, corresponding to roughly 4/5 of a standard deviation. The Academic Choice practice involves the teacher creating the learning objective but the student choosing how to engage in the work. Academic Choice has subcomponents including planning, working, and reflecting. Although these subcomponents are carefully articulated in the NEFC manual(2011), the subcomponents have similar characteristics to practices following from the theory of multiple intelligence and differentiated instruction (Tomlinson, 2001).

Psychometrically, the variation in achieved relative strength indices across the three core components can be interpreted in relation to the approach to measurement. Teacher-reported measures, though designed to tap exposure and adherence, also detect aspects of teachers' underlying belief system, as well as their day-to-day practices. In contrast, observed measures of *RC* reflect observed evidence of exposure and adherence to the intervention. In thinking through the process of teacher change, it may be easier to produce change in adherence to classroom practices than it is to change underlying beliefs (Rimm-Kaufman, Storm, Sawyer, Piant, & La Paro, 2006). For example, it may be easier for a teacher to adjust his daily routine to include the Morning Meeting four days a week than to change his philosophy about how and why students should be disciplined. Consider the following quotes from two *RC* coaches:

> "Although the teachers at XX School are very excited about Morning Meeting, they have many doubts about whether the rest of the *RC* approach, especially discipline and teacher language, will work for the students at their school. They seem somewhat discouraged that they still have behavior issues with which to deal. Many struggle with basic management issues and attribute much of their difficulties to the students."
>
> "I met with the three *RC* teachers about logical consequences and problem-solving conferences. They are still having some difficulty separating proactive approaches to discipline from reactive. Some also continue to question whether logical consequences are tough enough perhaps because they still have a basically punitive philosophy."

This may be more true for some *RC* practices than others, particularly for the observational measures. Teachers are observed only five times per year, and thus, the observed behaviors represent a sampling of what occurs during the year. On any given day, only a subset of *RC* practices can be observed within the classroom. For instance, an observer can notice and code the teachers' use of Morning Meeting components, the presence of posted rules and the extent to which they follow *RC* recommendations, and use of student choice in the classroom. However, observers may not be privy to classroom practices that occur only occasionally. Further, an observer in the classroom five days a year is unlikely to observe disciplinary practices used only several times per year to address very disruptive behavior. Also, without frequent observations conducted toward the beginning of the year, observers may be less able to report the full extent that teachers use the Interactive Modeling practices when introducing new materials to the class. Perhaps as a result of the approach to sampling for the observational measure, teachers are observed when they are engaged in the-day-to-day practices that best differentiate between intervention and control conditions. However, the actual practices that make the *RC* approach the most effective are not as easily observed or present frequently enough to be observed by sampling only five times per year.

In terms of best practices of fidelity assessment, these issues are not easily rectified *post hoc*. Rather, they are most effectively addressed by creating measurement instruments that reflect the breadth of core intervention components, and calibrated to the types of information that are best measured using observational versus self-reported measures. This will ensure that the measures of the intervention core components are being captured with construct validity. When this is accomplished, the average achieved relative strength helps intervention evaluators ensure construct validity by indexing the relative difference in intervention core components between treatment and control conditions. Because the average ARSI can be calculated on a familiar metric (i.e., Cohen's *d*), it provides a useful metric to communicate the strength of the intervention as implemented.

A second type of achieved relative strength index, the Binary Complier Index (Hulleman &Cordray, 2009),refers to a dichotomous index, or threshold, that represents whether or not students are receiving sufficient treatment to create change. In essence, the binary complier index can be determined by choosing practices, and setting a level of those practices regarded as an adequate dose of the intervention to create change. Cut-off values for the binary complier index can be derived theoretically or empirically based upon their ability to discriminate between *RC* teachers and non-*RC* teachers. In this section, we focus on the binary complier index for the Morning Meeting and Academic Choice practices.

In relation to Morning Meeting, two thresholds were set by the NEFC based upon theoretical grounds: one to capture exposure and frequency of the practice, and a second that also included aspects of adherence. The initial threshold was to hold a Morning Meeting at least four times a week. To capture adherence, there were several questions on the teacher survey (CPTS) that reflected subcomponents of Morning Meeting, including having the students greet each other,

sharing thoughts and feelings, having a community-building activity, and displaying a message of the day. Thus, the second Morning Meeting threshold was set such that a teacher needed to indicate that they did each of the additional aspects of Morning Meeting four or more times per week in order to be considered an *RC* teacher. In relation to Academic Choice, the threshold was set based on conceptual grounds. Specifically, we set the threshold for being a high fidelity *RC* teacher at once per week on each of three indicators from teacher self-reports. This decision reflected the notion that using the Academic Choice practice once a week is the minimum use to be considered an *RC* teacher..

For Morning Meeting, the exposure criteria indicated that nearly all (85%) of treatment teachers would be classified as *RC* teachers, whereas the adherence criteria indicated that just over half (58%) of treatment teachers were *RC* teachers. The percentage of teachers who were classified as *RC* teachers in the control condition can be thought of as the extent to which *RC* practices represent commonly used practices in the school. Interestingly, for Morning Meeting, this also depended on whether exposure (29%) or adherence (6%) was used. In this case, the difference reflects the construct breadth that the measure captures – the single-item measure captured whether Morning Meeting occurred at all (exposure), whereas the additional items helped capture adherence to specific sub-components of the Morning Meeting. For Academic Choice, 67% of teachers in the treatment group were classified as high fidelity *RC* teachers compared to 49% of teachers in the control group.

As outlined in the conceptual model, changes in *RC* practices are designed to shift classroom social processes and enhance children's engagement in learning. Such changes require repeated exercise and practice at sufficient frequency for teachers and children to develop new habits. In theory, holding a Morning Meeting four or more times per week gives opportunities for teachers

to teach and children to practice specific social skills (e.g., cooperation, turn-taking, empathy toward peers). Further, the Morning Meeting offers each child an opportunity to be a part of a positive social activity at school, thus promoting children's engagement in learning and supporting their positive feelings about school. The complier index provides a very useful tool for understanding construct validity in intervention fidelity. By setting a benchmark for intervention implementation, program developers are operationally defining what it takes to implement the intervention. Thus, implementers have a specified target to aim for when implementing the program that allows for some imperfection in implementation. Additionally, this assessment indicates aspects of the intervention where implementation drivers could facilitate higher levels of *RC* practices (c.f., Metz et al., this volume).

Measuring achieved relative strength in both intervention and control conditions raises questions about how much intervention is *enough* to create the intended impact of the intervention, as specified by the logic models (i.e., improved school, classroom, and child outcomes; see Figure 2). One way to determine this threshold is to empirically examine the relationship between intervention fidelity and outcomes. There are a hierarchy of approaches, from descriptive to inferential, that can be used to incorporate measures of intervention fidelity and achieved relative strength when analyzing program impacts.

Step 5: Linking fidelity measures to outcome measures.

There are numerous approaches that could be used to incorporate fidelity assessment into impact analyses, from descriptively comparing levels of implementation in treatment and control (e.g., Snyder et al., 2010), computing indices of achieved relative strength (Hulleman & Cordray, 2009), and correlational analyses between fidelity indices and outcomes in both treatment and control conditions, to more sophisticated approaches including replacing the

treatment indicator with the fidelity indicator or complier index (Schochet & Burghardt, 2007; Peck, 2003) and instrumental variables (Bloom, 2005). Various analytical frameworks can be employed including OLS multiple regression (Unlu, Bozzi, Layzer, Smith, Price, & Hurtig, 2011), multilevel modeling (e.g., Justice, Mashburn, Pence, & Wiggins, 2008), and structural equation modeling (e.g., Abry, Rimm-Kaufman, Hulleman, Thomas, & Ko, 2012; Kopp, Hulleman, Rozek, & Harackiewicz, 2012). Analyses can also be conducted to understand the sources of variation in intervention fidelity (e.g., Hulleman & Cordray, 2009). Importantly, all of these analytic approaches deviate, to a greater or lesser extent, from the causal inference framework produced by random assignment (Holland, 1986; Rubin, 1974). That is, even if construct validity of our fidelity measures has been secured, we are sacrificing some of our ability to infer cause and effect with these approaches because fidelity has been measured and not manipulated (cf. Imai, Keele, Tingley, & Yamamoto, 2011; Shadish, Cook, & Campbell, 2002). It is within this more limited causal inference context that the results of these analytic approaches can be understood. Given these limitations, we turn to the utility of linking these fidelity indices to outcomes.

First, descriptive and correlational analyses of fidelity measures within both treatment and control conditions can be used to examine presence and relationships among intervention core components. This approach allows us to examine, as directly as possible, whether the construct we theorized to impact outcomes was actually implemented. In essence, this is our test of construct validity: Did we successfully actualize our theorized model within the educational context? These initial questions are best examined through descriptive statistics of fidelity measures within both treatment and control conditions. Calculating achieved relative strength indices then provides an index of how strong the intervention was in comparison to the control

(or counterfactual) condition. In addition, zero-order correlations among core components provide evidence regarding the consistency of implementation core components. The higher the correlations among core components, then the more uniformly the intervention was implemented across core components. Furthermore, zero-order correlations between core components and outcomes provide us with two important pieces of information. First, significant (statistically, practically, or both) correlations provide information regarding the correlational aspect of causality: Does the supposed cause correlate with the outcome? If the cause (the intervention and associated core components) does not correlate with the outcome, then our ability to infer that the intervention causes changes in the outcome is seriously undermined. Second, correlations between core components and outcomes provide information about the impact of intervention strength: Does a higher amount of implementation lead to better outcomes?

A recommended practice is to table fidelity indices of core components in both treatment and control conditions. Included within these tables, or in a separate table, could be the achieved relative strength indices that are based on these values (e.g., Table 4). Formulas for calculating the achieved relative strength indices can be found in Hulleman and Cordray (2009), and are derivations of Cohen's *d* and Hedges' *g* adjusted for the type of index. In addition, presenting correlation matrices containing fidelity indices of intervention core components, mediational variables, and outcomes needs to be standard. Preferably, these correlation matrices would have the treatment condition values on one side of the diagonal and control condition values on the other side. This provides the opportunity to examine how relationships with intervention core components vary by condition. For example, Table 5 presents the correlation matrices among *RC* core components, classroom outcomes, and student outcomes. Correlations for the control group are above the diagonal, and correlations for the treatment group are below the diagonal.

INSERT TABLE 5 ABOUT HERE

Next, we can turn attention toward further analyses to examine the relationships between fidelity indices and outcomes. At the most basic level, comparing differences in program outcomes using the binary complier index is straightforward. Once the complier index is calculated, this analysis can be as simple as computing t-tests between the groups using the outcomes as dependent variables. Hulleman and Cordray (2009) presented the results of a motivation intervention in both the laboratory and the classroom using the binary complier index (see Figure 2, p. 99). Using the RCES data, we conducted similar analyses. As shown in Table 6, we computed binary complier indices for the *RC* core components of Morning Meeting and Academic Choice. Note that the Morning Meeting core component related more to classroom outcomes, whereas the Academic Choice component related more to student outcomes. Also note that the index created from the Morning Meeting adherence threshold showed larger contrasts on classroom outcomes than the index based on the exposure only threshold.

INSERT FIGURE 3 ABOUT HERE

Another approach is to utilize the framework described by Schochet & Burghardt (2007), which involves several variations of the complier index within a broader analytic framework. Within the context of clustered data, which are quite common in early care and education research, this means using a multilevel modeling or random effects framework. The treatment indicator can then be replaced with the binary complier index, or even the fidelity indicator itself (c.f., Justice, Mashburn, Pence, & Wiggins, 2008; Unlu, Bozzi, Layzer, Smith, & Price, 2011).

If the goal is to understand the role that intervention fidelity plays in explaining the treatment effect, it is possible to conduct path modeling to estimate the mediated or indirect effects of the intervention through the fidelity indices. This approach involves adding intervention fidelity

indices into the statistical model used to predict program outcomes (e.g., Abry, Rimm-Kaufman, Larson, & Brewer, 2012; Kopp et al., 2012). For example, within the context of the *RC* approach, Abry et al. (2011) computed multi-item indices of four core *RC* practices: Morning Meeting, Classroom Organization, Interactive Modeling, and Academic Choice. Utilizing a multi-level modeling framework, they found that Academic Choice accounted for unique variance in student math and reading test scores, whereas the other practices did not.

Although we focus on analysis of fidelity within randomized experiments, fidelity can also be incorporated within quasi-experimental designs. Although quasi-experimental designs have weaker claims to causal inference than experimental designs, valuable information about levels of program fidelity can still be made, particularly if fidelity was assessed within the comparison condition. In those cases, the analyses would be quite similar: overall descriptive statistics that capture the degree of fidelity in treatment and control conditions, indices of achieved relative strength, correlations between level of fidelity and outcomes, and the treatment indicator could be adjusted for the level of fidelity in both treatment and control conditions. Measuring fidelity in quasi-experiments can even strengthen internal validity claims because more is known about the level of fidelity in the non-treated group.

VII. Applications of Fidelity Indices to Address Important Questions in Intervention Fidelity Assessment

In this section, we outline several issues within the measurement of intervention fidelity and its application to understanding early childhood interventions, including how much adaption is too much (i.e., the zone of tolerable adaptation), weighting and combining fidelity indices, sequencing of core intervention components, the interaction between components (and between

components and program implementers), matching measures to core components, and the development of implementation skills.

Zone of Tolerable Adaptation

How much can program implementers change an intervention before treatment effects disappear? Is it more important for implementers to follow a very specific protocol established by the developers or are interventions more effective when adapted to local conditions (c.f., Dusenbury, Brannigan, Falco & Hansen, 2003)? These questions are raised repeatedly as early care and education settings adopt new interventions and strive to maintain their presence in a district (Dusenbury et al., 2003; Greenberg, 2010; Stringfield, Reynolds, & Schaffer, 2008). In essence, these questions speak to the issue of *zone of tolerable adaptation,* defined as the extent to which an intervention can be modified and tailored before the accompanying treatment effects disappear. On one hand, modifications to an intervention may help make the intervention well-suited for local conditions (Datnow & Stringfield, 2000). For example, teachers may adopt the intervention to be more culturally sensitive or tailor the intervention activities to meet the developmental level of the students in the classroom. In fact, Stringfield et al. (2008) report that school-wide reform interventions are most effective only when such mutual adaptation, or co-construction, of the intervention occurs. On the other hand, adaptation of program components may result in dropping components of the intervention demonstrated to be critically important to its effectiveness.

The need to adapt interventions to local conditions has become an accepted condition of scale-up efforts (Dusenbury, et al., 2003), particularly so that the intervention is sustained in early childhood programs or initiatives such as home-based or center-based early childhood programs, home visitation programs, local ECE professional development programs, or

statewide ECE professional development systems (Greenberg, 2010; Stringfield et al., 2008). As a result, intervention developers have increased the flexibility of their programs so that teachers can use, integrate, and take ownership of the programs and their implementation. Adaptations can only be conducted, however, when implementers adhere to core principles of the intervention, knowing which elements can and cannot be adapted (Dusenbury et al., 2003). Crucially, applying the key supports of successful implementation, or implementation drivers, when adapting interventions can facilitate successful adaptation (Fixsen et al., 2005).

Knowledge of systematic research that links intervention outcomes to fidelity of core intervention components, grounded within the conceptual framework of the intervention, is necessary in order to ascertain which components are crucial and which are adaptable. The conceptual framework identifies the scope of practices that are, and are not, consistent with core components that combine to construct the intervention. Research support for the relationship between implemented core components and outcomes guides an understanding of which components are crucial for intervention impact. In this way, the gestalt of the intervention is retained (thus maintaining construct validity), yet guidance in terms of intervention adaptation is provided.

Within the *RC* approach, tolerable adaptation of Morning Meeting might require students moving chairs and desks out of the way for a Morning Meeting, or moving the Morning Meeting later in the day to create time for art or music. In contrast, adaptation that is not tolerable might include a regular Morning Meeting conducted in rows so that children cannot see one another, a Morning Meeting that engages in straightforward rehearsal and memorization tasks instead of an activity, or a Morning Meeting in which children are being unkind to one another but the teacher does not intervene. In relation to Academic Choice, tolerable adaptation might include having

children engage in a working phase that is tailored to students' individual ability levels or having students reflect upon their learning through conversation or in writing. However, intolerable adaptation might include limiting choice within activities to letting children decide where they will sit in the room during reading or dropping the planning and reflection *RC* core subcomponents from the academic choice activity. The analytic methods described earlier could assist in determining which adaptations were or were not tolerable in a planned variation model (Yeh, 2000), where variations of the intervention are implemented and connected to variations in outcomes.

Weighting and Combining Components

Once fidelity indicators have been developed, an important issue during the analytic phase is to consider how best to weight each indicator when creating indices, either to specific intervention components or to the overall model. The first step in the process is to link the measurement of fidelity back to the logic models, and not simply create indices based upon the method of data collection. In the *RC* Efficacy Study, three measures of fidelity were utilized. One option for creating fidelity indices would be to sum the scores on each measure to create three distinct fidelity measures: the CPOM, CPTS, and CFTS. However, each of these instruments contains items from some or all of the core components, and often not equally weighted. For example, 64% of the items on the CPTS measure the Morning Meeting, compared to only 43% of the items on the CPOM. The approach we recommend, in contrast, is to create indices based on core components. Not only does this approach enable the researcher to evaluate fidelity to intervention core components separately, but it also has more power empirically. As demonstrated by Abry et al. (in preparation), fidelity indices created by core components accounted for more variance in child outcomes than those created by method of measurement.

In addition, some components might be weighted more heavily than others when creating indices and conducting analyses. This emphasis might be because, at the conceptual level, some components are thought to be more crucial, or based on prior research demonstrating their primacy, or other considerations such as measurement precision or frequency. The program logic models should be our first point of reference in this process. In the *RC* model, not all *RC* practices are equivalently likely to produce changes in children's achievement outcomes. For example, the Morning Meeting is the most common entry point to implementing the Responsive Classroom approach. In contrast, Academic Choice is expected to be a more prominent driver of academic outcomes and is thus weighted more heavily. Table 6 assigns approximate weights to the various components in relation to *RC* program outcomes. Because some of the *RC* practices were not measured adequately, we only include the practices for which we have strong measures. This weighting will then have implications for the development of indices and incorporation into data analyses. For example, in the creation of an overall fidelity index, the measures of each of the core components could be multiplied by their respective weights when computing a total fidelity score.

INSERT TABLE 6 ABOUT HERE

Sequencing of Components

Does the theory of change specify that some components should be implemented before others, either as part of a developmental sequence or setting the foundation for later instruction, and what does this mean for measurement and analyses? For example, the *RC* practice of Morning Meeting occurs early in the day, and early in the year, which provides the facilitative socio-emotional context for optimal learning later in the day and later in the year. Thus, measuring its presence in the beginning of the year may have a different meaning, and thus be of

more importance in terms of construct validity, than later in the year. A second *RC* example is that Guided discovery offers a structure for introducing students to the use and care of new materials. Teachers use this more in the beginning of the year and/or the beginning of new lessons. Thus, observing its presence in the beginning of the year or as teachers introduce new lessons would reflect the highest construct validity.

Matching Measures and Core Components

From a measurement standpoint, researchers and evaluators need to consider the extent to which the type of measure might be better suited for some core components than others. In general, program implementers (health care providers, coaches, teachers)are probably better at reporting whether they did certain aspects of the program, and in what order, than they are at reporting how engaged children were during the learning activity, or on how many off-task behaviors occurred during a learning session (suggesting that children may not have been exposed to the intervention). The latter two are probably better suited for observational measures. In the *RC* approach, teachers are probably better at reporting their use of a Morning Meeting and the Morning Meeting components. However, they may be less effective at assessing the quality of the social interactions during that Morning Meeting because they are participants in the meeting.

In contrast, there are other instances when self-report is likely to be more valid measure of the intervention than observation. This may be particularly true for behaviors that occur infrequently during the year. *RC* teachers may set up school-family activities where families articulate and represent what they hope for their children during the year (i.e., "Hopes and Dreams"). Another example is the rare but salient negative situation that can swamp months of teacher effort toward high implementation. Imagine the high implementing *RC* teacher who

"loses it" with her students and gets angry. This kind of very negative behavior may have long-lasting consequences but may be very hard to measure via observation. The most valid way to get at these low frequency events is probably through teacher report. A different example pertains to teacher beliefs about schooling, which are an important part of the theory of change in most educational interventions, including the *RC* approach. It is relatively straightforward to observe whether a teacher has a Morning Meeting, it is not as plausible that an observational measure can directly capture a teacher's beliefs about how children learn. Measures of beliefs are probably better measured through teacher self-report (Rimm-Kaufman et al., 2006).

The above discussion highlights the fact that there are no perfect measures. One way to address this challenge is by triangulating among several different types of measures. For example, on-site coaches could report on teacher level and quality of implementation, as could independent observers. When combined with teacher reports, this variety of measures strengthens the breadth and depth of the assessment of fidelity to intervention core components.

Timing and the Implementation Dip

Becoming a high implementing practitioner reflects a process of human change (Evans, 2001). For example, when practitioners initiate their training, the intervention likely produces a disruption in normal practice and may actually decrease the effectiveness of new practices, not increase them. This process, or implementation dip (Fullan & Miles, 1992), is important to recognize and consider analytically (Borman, Gamoran, & Bowdon, 2008).As noted earlier in this volume (Metz et al., this volume), reaching full implementation may take a period of two to four years, depending on the complexity of the intervention model, the availability of implementation supports and resources, and characteristics of the individual practitioners and contexts in which they are intervening. Within the *RC* approach, for example, setting aside time

for a Morning Meeting may be relatively easy; however, adhering to the multiple aspects of a Morning Meeting, and doing so with quality, may take time. Initially, it may take longer to run a Morning Meeting than intended, and thus reduce the amount of instructional time. However, once the Morning Meeting routine has been established, then increases in efficiency and quality may occur, thus boosting learning in other areas. In contrast, the effects of Academic Choice on student motivation and engagement to learn particular content may have more immediate effects. Thus, *RC* teachers may need more time to develop the socio-emotional foundation of the classroom, whereas some aspects of the instructional environment may be quicker to become fully implemented. These principles hold true regardless of the intervention.

These examples demonstrate that the development of practitioners' skills in implementing a particular intervention's core components needs to be considered when evaluating the effectiveness of an intervention. This bigger picture perspective on the intervention, which should be incorporated within the logic models, reinforces the whole-part-whole challenge evaluators and researchers face when assessing intervention fidelity. The process is not as simple as developing indicators of each core component and averaging them together to form an index of intervention fidelity. Issues such as sequencing, timing, development of implementers' skills, and tolerable adaptation can all influence the construct validity of intervention fidelity measures.

VIII. Conclusions and Implications for Early Childhood Education Program Evaluation and Research

One of the goals of measuring intervention fidelity with careful attention to construct validity is to contribute needed information at various points in the intervention development cycle. Furthermore, the methods have different utility depending on the various stages of

implementation (i.e., exploration, installation, initial implementation and full implementation; see Metz et al., this volume) and stages of evaluation (e.g., efficacy trials, impact evaluation). For instance, achieved relative strength values may be necessary in examining the feasibility of an intervention and whether the training actually produces changes in implementation. Further, measuring the sequencing of components and interactions between components and implementers may play an important role in building a feasible intervention that can be used by a variety of implementers in a broad range of settings. The Binary Complier Index may be particularly relevant in efficacy trials as a way of conducting treatment-on-the treated analyses. Given the unevenness in intervention uptake, thresholds that establish sufficient and insufficient use of the intervention may direct intervention developers and researchers to return to an earlier stage of implementation or to replication and scale-up. Understanding the Zone of Tolerable Adaptation may be particularly relevant to addressing questions that rise from scale-up studies (Yeh, 2000). Specifically, mixed methods techniques can be used to assess the adaptations to the approach that are present, and subsequently evaluate the point at which those adaptations no longer fit with the definition or intention of the intervention.

Breaking down intervention fidelity into its essential elements helps to identify practices that are core to the effectiveness of the intervention compared to those practices that may be less essential. As a consequence, new, more streamlined interventions can be developed that focus more attention on training in the more important practices compared to the less important practices. By linking fidelity indices to core components, implementation drivers can be added for core intervention components that were not implemented well. However, in conducting these analyses, it is important to remember that the whole is not merely a sum of its parts, and analyzing the relationship between individual core components and outcomes might not fully

represent the intervention model or its effectiveness. Thus, researchers and evaluators would be wise to take a step back from their analytic framework to ensure they are capturing the whole, and not merely the sum of the parts. Indeed, early childhood interventions discussed later in this volume should be considered in terms of the time it takes practitioners to acquire new skills, and the alignment of the evaluation with the stage of implementation. The various analytic methods introduced in this chapter (e.g., creating overall fidelity indices as well as those representing individual core components) can assist researchers and evaluators in this endeavor.

References

Abry, T., Rimm-Kaufman, S. E., Hulleman, C. S., Thomas, J. B., &Ko, M. (2012, March). *The how and for whom of program effectiveness: Dissecting the Responsive Classroom® approach in relation to academic achievement*. Paper presented at the Spring conference of the Society for Research on Educational Effectiveness, Washington, DC.

Abry, T., Rimm-Kaufman, S. E., Larsen, R. A. & Brewer, A. J. (2011, September).*Applying new methods to the measurement of fidelity of implementation: Examining the critical ingredients of the* Responsive Classroom® *approach in relation to mathematics achievement*. Poster presented at the Society for Research on Educational Effectiveness, Washington, DC.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-Year achievement effects. *Journal of Research on Educational Effectiveness, 1*, 237-264.

Casto, K.,& Audley, J. (2008). In Our School: Building Community in Elementary Schools. Turner Falls, MA: Northeast Foundation for Children.

Cordray, D. S., &Hulleman, C. S. (2009, June).*Assessing intervention fidelity in RCTs: Models, methods and modes of analysis*. Invited panel session at the 2009 Institute of Education Sciences Research Conference, Washington, D.C.

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(1),* 23-45.

Datnow, A., & Stringfield, S. (2000).Working together for reliable school reform. *Journal of Education for Students Placed at Risk, 5(1&2),* 183-204.

Dobson, L., & Cook, T. (1980). Avoiding Type III error in program evaluation: results from a field experiment. *Evaluation and Program Planning, 3, 269 - 276.*

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41,* 327-350.

Evans, R. (2001) *The Human Side of School Change.* San Franciso, CA: Jossey-Bass, Inc.

Fullan, M., & Miles, M. (1992, September).Getting reform right: What works and what doesn't. *Phi Delta Kappan,* 745–752.

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 on Intervention Effectiveness, 2(1),* 88-110.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011).Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *American Political Science Review, 105 (4),* 765-789.

Justice, L. M., Mashburn, A., Pence, K. L.,& Wiggins, A. (2008). Experimental evaluation of a preschool language curriculum: Influence on children's expressive language skills. *Journal of Speech, Language, and Hearing Research, 51,* 983-1001.

Knowlton, L. W., & Phillips, C. C. (2009).*The logic model guidebook: Better strategies for great results.* Washington, D. C.: Sage.

Lugo-Gil, J., Sattar, S., Ross, C., Boller, K., Kirby, G., & Tout, K. (2011).*The Quality Rating and Improvement System (QRIS) evaluation toolkit* (OPRE Report 2011-31). Office of Planning, Research and Evaluation: Washington, D.C.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007).Treatment integrity of school-based interventions with children in the Journal of Applied Behavior Analysis, 1991-2005.*Journal of Applied Behavior Analysis, 40,* 659-672.

Mitchel, M. E., Hu, T. W., McDonnell, N. S., & Swisher, J. D.(1984). Cost-effectiveness analysis of an educational drug abuse prevention program. *Journal of Drug Education, 14(3),* 271-292.

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. *Journal of Behavioral Health Services and Research,* 1-22. DOI: 10.1007/s11414-012-9295-x

Northeast Foundation for Children (2006). *Creating a Safe and Friendly School*. Turner Falls, MA: Northeast Foundation for Children.

Nunnery, J. A., Slavin, R. E., Madden, N. A., Ross, S. M., Smith, L. J., Hunter, P., et al. (1997, March). *Effects of full and partial implementations of* Success for All *on student reading achievement in English and Spanish*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78,* 33-84.

Rimm-Kaufman, S. E., Fan, X, Berry, R. & Justice, L. (March, 2007-February, 2011).*The efficacy of the Responsive Classroom approach for improving teacher quality and*

children's academic performance. Institute for Education Sciences, U.S. Department of Education, Teacher Quality-Mathematics.

Rimm-Kaufman, S. E., Berry, R., Fan, X., McCracken, E. &Walkowiak, T. (2008, June).*The Efficacy of the Responsive Classroom Approach for Improving Teacher Quality and Children's Academic Performance*. Paper presented at the Institute of Education Sciences Research Conference, Washington, D.C.

Rimm-Kaufman, S. E., & Sawyer, B. E. (2004). Primary-grade teachers' self-efficacy beliefs, attitudes toward teaching, and discipline and teaching practice priorities in relation to the *Responsive Classroom* approach. *Elementary School Journal, 104*(4), 321-341.

Rimm-Kaufman, S. E., Fan, X., Chiu, Y. I., & You, W. (2007). The contribution of the *Responsive Classroom* approach on children's academic achievement: Results from a three year longitudinal study. *Journal of School Psychology, 45,* 401-421.

Rimm-Kaufman, S. E., Storm, M., Sawyer, B., Pianta, R. C., and La Paro, K. 2006.The Teacher Belief Q-Sort: A measure of teachers' priorities and beliefs in relation to disciplinary practices, teaching practices, and beliefs about children. *Journal of School Psychology*, 44, 141-165.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002).*Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Snyder, F., Flay, B., Vuchinich, S., Acack, A., Washburn, I., Beets, M., & Li, K-K. (2010). Impact of a Social-Emotional and Character Development Program on School-Level Indicators of Academic Achievement, Absenteeism, and Disciplinary Outcomes: A Matched-Pair, Cluster-Randomized, Controlled Trial. *Journal of Research on Educational Effectiveness, 3*, 26-55.

Stringfield, S., Reynolds, D., & Schaffer, E. C. (2008).*Improving secondary students' academic achievement through a focus on reform reliability: Four- and nine-year findings from the High Reliability Schools project*. CfBT Education Trust. Accessed online: http://www.cfbt.com/evidenceforeducation/pdf/high%20reliability_v5%20final.pdf.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2010). *Compendium of quality rating systems and evaluations*. Office of Planning, Research and Evaluation: Washington, D.C.

Tortu, S., & Botvin, G. J. (1989). School-based smoking prevention: The teacher training process. *Preventive Medicine,* 18, 280-289.

Unlu, F., Bozzi, L., Layzer, C., Smith, A., Price, C., & Hurtig, R. (2011). *Linking implementation fidelity to impacts in an RCT: A matching approach*. Abt Associates.

Wanless, S.B., Patton, C.S., Rimm-Kaufman, S.E., Deutsch, N.L. (2012). Setting-level influences on implementation of the *Responsive Classroom* approach. *Manuscript in press at Prevention Science.*

Weiner, B. (1972). *Theories of Motivation: From Mechanism to Cognition*. Chicago: Markham Publishing Company.

Wilson, Freeman-Loftis, Sawyer, & Denton, 2009.*The* Responsive Classroom *Assessment Guide.* Turner Falls, MA: Northeast Foundation for Children.

Yeh, S. S. (2000).Improving educational and social programs: A planned variation cross-validation model. *American Journal of Evaluation, 21(2)*, 171-184.

Zubrzycki, J. (2012). Research links 'responsive' teaching to academic gains. *Education Week, 32(4)*. Accessed online: http://www.edweek.org/ew/articles/2012/09/13/04responsive.h32.html.

Table 1. *Responsive Classroom*® Core Principles and Practices.

Core Principles	Core Practices
1. The social curriculum is as important as the academic curriculum. 2. How children learn is as important as what children learn. 3. The greatest cognitive growth occurs through social interaction. 4. Cooperation, assertion, responsibility, empathy, and self-control are needed to be successful socially and academically. 5. Knowing the children individually, culturally, and developmentally is as important as knowing the content. 6. Knowing parents is as important as knowing the content. 7. How the adults at school work together to accomplish their mission is as important as the content.	The following 10 practices emanate from the 7 principles: 1. Morning Meeting 2. Academic Choice 3. Interactive Modeling 4. Rule Creation 5. Positive Teacher Language 6. Logical Consequences 7. Guided Discovery 8. Classroom Organization 9. Working with Families 10. Collaborative Problem-Solving

Note: For more information, see http://www.responsiveclassroom.org/.

Table 2. *Responsive Classroom* Fidelity Indicators.

Core Components	Sub-components	Total number of indicators	Indicators per component
Morning Meeting	General	5	25
	Greeting	3	
	Sharing	4	
	Group activity	6	
	Morning message	7	
Academic Choice	Plan	4	11
	Work	4	
	Reflect	3	
Interactive Modeling	Teacher demonstration	2	7
	Student observations	2	
	Student practice	3	
Rule Creation	Students generate hopes and dreams	2	11
	Students brainstorm rules	4	
	Rules consolidated	3	
	Rules posted	2	
Teacher Language	Reinforcing	1	4
	Reminding	2	
	Redirecting	1	
Logical Consequences	Respectful	0	7
	Relevant	3	
	Realistic	0	
	Time-out	4	
Guided Discovery	Introduce material	0	3
	Generate and model ideas for use and care	2	
	Explore and experiment with material	1	
Classroom Organization	Arrangement	1	4
	Materials	2	
	Displays	1	
Working with Families	General communication	1	2
	Involve parents in goal setting	1	
	Involve parents in classroom and school activities	0	
Collaborative Problem Solving	Conferencing	0	0
	Role play	0	

Table 3.*Responsive Classroom* Intervention Fidelity Indicators by Intervention Core Component, Sub-component, and Type of Measure.

Core *RC* Component	Sub-component	# of indicators			Total # of indicators
		CPOM	CPTS	CPFS	
Morning Meeting	General	1	3	1	5
	Greeting	1	1	1	3
	Sharing	1	2	1	4
	Group activity	1	4	1	6
	Morning message	2	4	1	7
Academic Choice	Plan	2*	1	1	4
	Work	2*	1	1	4
	Reflect	0	2	1	3
Interactive Modeling	Teacher demonstration	1*	0	1	2
	Student observations	1*	0	1	2
	Student practice	1*	1	1	3

Note: *RC* = Responsive Classroom; CPOM = Classroom Practices Observation Measure; CPTS = Classroom Practices Teacher Survey, CPFS = Classroom Practices Frequency Survey
* Indicates a single item that addresses multiple sub-components.

Table 4.Achieved relative strength values for 4th-grade teachers in the *Responsive Classroom* Efficacy Study.

RC Practice	*N*	Overall Mean (*SD*)		Treatment Mean (*SD*)		Control Mean (*SD*)		Min	Max	ARSI
MM	84	-0.13	(1.00)	0.65	(0.58)	-0.95	(0.61)	-1.49	1.10	2.19
IM	84	0.00	(0.74)	0.19	(0.76)	-0.20	(0.65)	-2.09	1.46	0.47
AC	84	-0.04	(0.86)	0.33	(0.84)	-0.42	(0.71)	-1.64	2.23	0.80

Note: RC = Responsive Classroom; MM = Morning Meeting; IM = Interactive Modeling; AC = Academic Choice; ARSI = Achieved Relative Strength Index. N = 43 (treatment) and 41 (control).

Table 5.*Correlations for Core Intervention Components and Selected Outcomes by Experimental and Control Groups.*

	1	2	3	4	5	6	7	8
RC Practice								
1. MM	--	.43**	.19	.16	.23	.10	-.10	-.02
2. IM	.39**	--	-.06	.39*	.15	.22	-.03	.02
3. AC	.16	.49**	--	-.03	-.11	-.20	.36*	.36*
Classroom Outcome								
4. ES	.32*	.16	.29	--	.54***	.69***	.26	-.16
5. CO	.23	.18	.47***	.67***	--	.59***	-.16	-.05
6. IS	.22	.18	.28	.72***	.67***	--	.03	-.26
Student Outcome								
7. Mathematics	-.26	.08	.24	.10	.32	.07	--	.10
8. Reading	-.24	-.16	.25	-.05	.08	-.10	.69***	--

Note: MM = Morning Meeting; IM = Interactive Modeling; AC = Academic Choice; ES = Emotional Support; CO = Classroom Organization; IS = Instructional Support. ES, CO, and IS are domains of the Classroom Assessment Scoring System N = 43 (treatment) and 41 (control). Mathematics and Reading achievement are aggregated at the classroom level. Correlations for the control group are presented above the diagonal. Correlations for the experimental group are presented below the diagonal.

Table 6.Example Weighting of *RC* Practices for Child Social and Academic Outcomes.

	Social Outcomes		Achievement Outcomes	
	Weight assuming equal importance	Weight based upon logic models	Weight assuming equal importance	Weight based upon logic models
Morning Meeting	.2	.3	.2	.15
Creating Rules with Children, Modeling Rules, and Approaching Discipline	.2	.3	.2	.15
Teacher Language	.2	.2	.2	.3
Interactive Modeling and Guided Discovery	.2	.1	.2	.1
Academic Choice	.2	.1	.2	.3

Figure 1.The conceptual logic model for the Responsive Classroom approach.

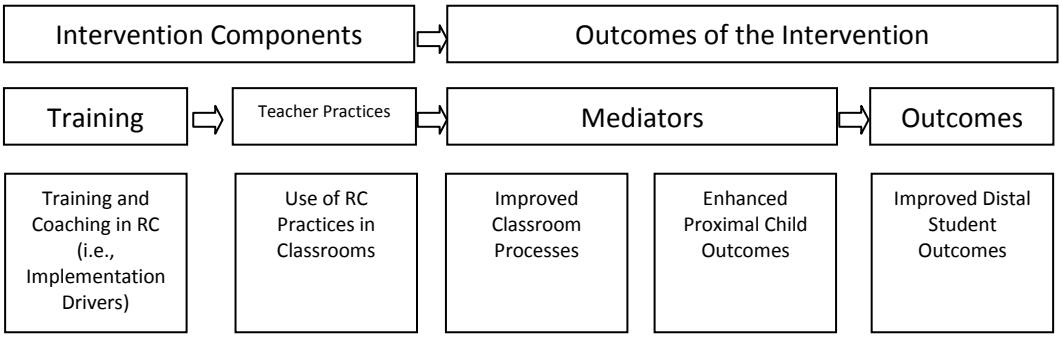


Figure 2.The operational logic model for the Responsive Classroom approach.

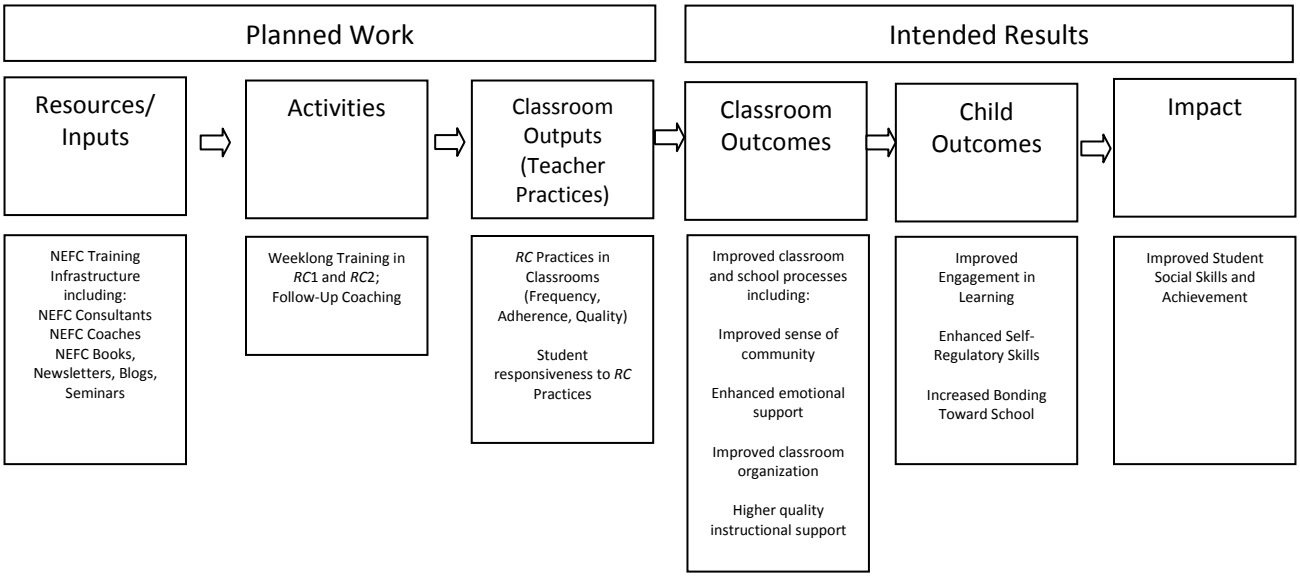
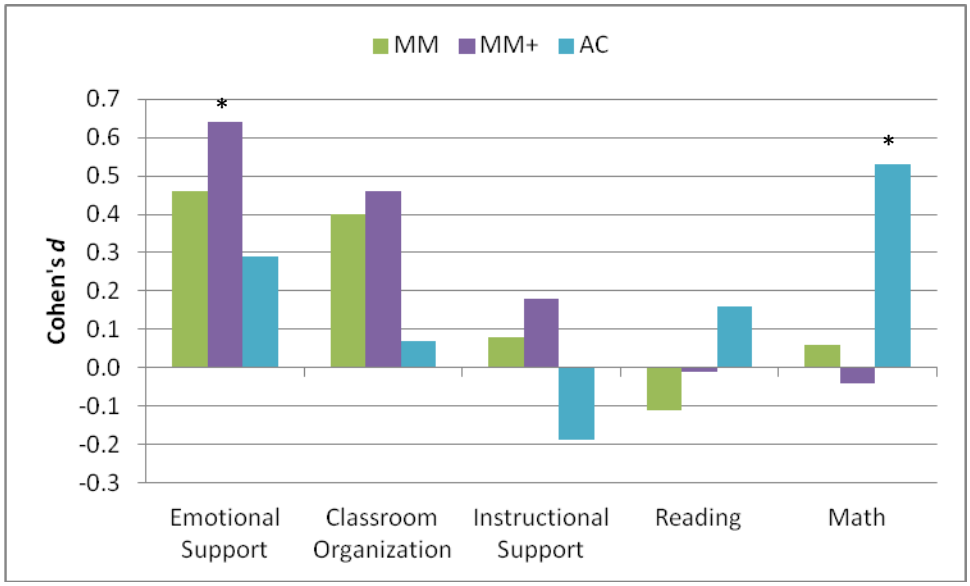


Figure 3.The Relationship between Intervention Fidelity (i.e., the Binary Complier Index) and Program Outcomes in the *Responsive Classroom* Efficacy Study.



Note: The Binary Complier Index was computed such that teachers in the study were categorized as either meeting the threshold for implementing that specific program core component (compliers) or not (non-compliers). The *RC* practices presented here are: MM = Morning Meeting measured with a single-item (exposure); MM+ = Morning Meeting measured with multiple items (adherence); AC = Academic Choice. The outcome variables are listed along the X-axis and include emotional support, classroom organization, and instructional support (classroom outcomes), and reading and math (intervention impacts; see Figure 2).The bars represent differences between compliers and non-compliers in standard deviation units of the outcomes on the Y-axis.

* Denotes that compliers were significantly different from non-compliers at $p < .05$.