Check for updates

Article

American Journal of Evaluation
2015, Vol. 36(3) 320-338
© The Author(s) 2014
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1098214014557009
aje.sagepub.com
SAGE

Using Indices of Fidelity to Intervention Core Components to Identify Program Active Ingredients

Tashia Abry[1], Chris S. Hulleman[2], and Sara E. Rimm-Kaufman[2]

Abstract

Identifying the active ingredients of an intervention—intervention-specific components serving as key levers of change—is crucial for unpacking the intervention black box. Measures of intervention fidelity can be used to identify specific active ingredients, yet such applications are rare. We illustrate how fidelity measures can be used to identify program active ingredients in the context of a social-emotional learning intervention (*Responsive Classroom*). We applied one customary and two novel approaches to create indices of fidelity. In the customary approach, we averaged fidelity ratings across all core components. In the novel approaches, we computed fidelity indices for specific components by (a) averaging responses from like items and (b) deriving factor scores from a multitrait, multimethod factor analysis. We then tested indices in relation to achievement gains ($N = 1{,}442$). Indices derived using novel approaches explained more outcome variance than indices from the customary approach. Further, novel approaches revealed one component as a potential active ingredient. Discussion highlights strengths and limitations of the indices and implications for identifying program active ingredients.

Keywords

intervention fidelity, implementation, core components, active ingredients, social and emotional learning, *Responsive Classroom*

Identifying the active ingredients of an intervention—intervention-specific components serving as key levers of change—is a crucial part of unpacking the intervention black box. Knowledge of active ingredients can be used to identify specific practices that promote desired change in participants, optimize interventions, and create highly effective integrated interventions that combine active ingredients.

[1] Arizona State University, Tempe, AZ, USA
[2] University of Virginia, Charlottesville, VA, USA

Corresponding Author:
Tashia Abry, Arizona State University, T. Denny Sanford School of Social and Family Dynamics, 951 S. Cady Mall, Tempe, AZ 85287, USA.
Email: tabry@asu.edu

Identifying active ingredients requires an understanding of the extent to which component parts of an intervention relate to targeted outcomes. Measures of implementers' fidelity to intervention core components—intervention elements hypothesized to promote desired change—can be used for this purpose, but to do so requires more nuanced indices of fidelity than are typically utilized. Using a social–emotional learning intervention, we illustrate the use of customary and novel approaches to creating indices of fidelity to intervention core components. We then exemplify how the indices of fidelity to intervention core components can be used to identify program active ingredients by testing fidelity indices in relation to outcomes targeted by the intervention.

Intervention Fidelity

Intervention fidelity can be defined as the extent to which the core components of a program, differentiated from "business as usual," are carried out as intended upon program enactment (Century, Rudnick, & Freeman, 2010; Dusenbury, Brannigan, Falco, & Hansen, 2003; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). Intervention fidelity includes domains of adherence (compliance to core components), dosage (frequency of use of core components or amount of exposure), and quality (how closely the use of core components resembles the theoretical ideal; Century, Cassata, Rudnick, & Freeman, 2012; Dane & Schneider, 1998). Measures of intervention fidelity are crucial in program evaluation studies. Namely, fidelity assessments help increase the internal validity of conclusions made regarding program ineffectiveness by allowing researchers to distinguish between intervention failure and implementation failure (Dobson & Cook, 1980). High levels of intervention fidelity increase the likelihood of detecting program effects and are consistently associated with better outcomes in curricular, health, and preventive interventions (Dane & Schneider, 1998; Durlak & Dupre, 2008; Dusenbury et al., 2003).

Intervention Core Components and Active Ingredients

Various frameworks for conceptualizing intervention fidelity have emerged in the last decade (O'Donnell, 2008) primarily in response to the call to better understand the implementation process and intervention effectiveness (Greenberg, 2010). An important strength of these frameworks is that they provide procedures and templates for defining, measuring, and evaluating fidelity criteria that can be applied across educational, behavioral, prevention, and health promotion programs. A common thread among these frameworks is the importance of measuring implementers' fidelity to the intervention-specific practices hypothesized to effect change. For example, Mowbray, Holter, Teague, and Bybee (2003) propose measurement of the *structural and procedural fidelity criteria* of health and education interventions. Century, Rudnick, and Freeman (2010) suggest measuring the *critical components* of instructional materials. Recently, Nelson, Cordray, Hulleman, Darrow, and Sommer (2012) recommended measuring *intervention components* articulated in the intervention theory of change. Differences in terminology aside, all three frameworks advocate identifying and measuring implementers' fidelity to the intervention-specific components hypothesized to promote desired outcomes, which we term *intervention core components*. Program developers and evaluators are responsible for articulating the intervention core components specific to their program. For example, the intervention core components of a school-based prevention program may consist of weekly lessons and activities, while the core components of a community-based health promotion program may include community partnerships, public service announcements, and the provision of health-related services.

Intervention core components are the elements of the intervention *hypothesized* to transmit effects and therefore are the target of fidelity assessments. Fidelity assessments are inherently unique to each intervention and thus rely primarily on guidance from developers. In other words, fidelity measures are created to assess adherence to the practices, procedures, and use of materials that make up the

intervention, created and packaged by program developers because their use is expected to have a desirable effect. Yet in practice, it is unlikely that the core components of an intervention carry equal weight in terms of their importance. The use of some core components may relate to outcomes more strongly than others; in other cases, core components thought to be important may be entirely superfluous. As such, in order to know how an intervention is actually working, core components must be not only identified but also isolated and examined in relation to measured outcomes. Those intervention core components whose use can be empirically linked to targeted results serve as catalysts for change and comprise the *active ingredients* of the intervention (Sidani & Sechrest, 1999). If a core component does not relate to outcomes as anticipated, it may be deemed less or nonessential.

Why Identify Active Ingredients?

Active ingredients describe program mechanism (Shadish, Cook, & Campbell, 2002) and distinguish essential from nonessential components (Collins, Murphy, Nair, & Strecher, 2005; O'Donnell, 2008). This knowledge has important applications. First, it can inform refinements to an intervention to help strengthen its effectiveness (Collins et al., 2005) and provide guidance to practitioners and support staff (e.g., coaches) on what to prioritize to get the most leverage from the program. In some cases, costly and complex interventions may not be necessary—active ingredients may reflect relatively simple methods of behavior that can be used to target specific problems or populations in lieu of a comprehensive intervention, thus saving time, energy, and resources (Embry & Biglan, 2008). When accumulated over time and across interventions, knowledge of active ingredients can contribute to overarching theories of best practice. For example, in their review of interventions designed to serve at-risk children and youth, Li and Julian (2012) deduced that program effects were diminished in the absence of a focus on developmental relationships. They posited that the fostering of warm, reciprocal relationships is an element key to the success of these kinds of interventions.

Second, cumulative knowledge of active ingredients can lead to the formulation of integrated interventions that combine active ingredients from multiple interventions with shared goals. An example of a school-based integrated intervention is the Promoting Alternative Thinking Strategies (PATHS) to Pax model. PATHS to Pax combined elements of two social–emotional learning programs, PATHS and the Pax-Good Behavior Game (Domitrovich et al., 2010). Program components selected for inclusion in the integrated model had been previously studied and were conceptually complimentary. Integrated interventions like PATHS to Pax represent a promising direction because they may maximize impact through the synergistic effects of packaged best practices. As an added benefit, program sustainability may be enhanced by reducing the number of interventions needed, and therefore, lessening the burden on implementers (e.g., teachers) (Domitrovich et al., 2010).

Third, practitioners, evaluators, and policy makers are increasingly interested in answering the question "how much is enough?" (Bruns, Suter, & Leverentz-Brady, 2008; Salyers et al., 2003). An understanding of a program's active ingredients is crucial to the creation of meaningful thresholds of fidelity to a program. Otherwise, practitioners may be advised to aim for a certain level of fidelity to intervention components even if that intervention component is nonessential.

Deriving and Testing Indices of Fidelity to Intervention Core Components

Despite the utility of knowing which intervention core components constitute active ingredients (Irwin & Supplee, 2012), there is surprisingly little empirical work examining fidelity to individual intervention core components and their relation to program outcomes (Collins et al., 2005; Mowbray, Holter, Teague, & Bybee, 2003). Implementation research across education, prevention, and behavioral health fields continues to focus primarily on personal and contextual factors that promote or hinder fidelity to an intervention model (Century et al., 2012; Lunn et al., 2011) and relating measures of (overall) fidelity to outcomes (Pas & Bradshaw, 2012). The customary approach to

evaluating intervention fidelity involves aggregating fidelity ratings across core components, resulting in a composite index of fidelity to the intervention as a package (Durlak & Dupre, 2008; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). This approach efficiently combines a large amount of data and produces a composite that is easy to understand. Indeed, it has led to crucial advancements in implementation science, as it is now well established that fidelity to a treatment model is a requisite for success (Durlak & Dupre, 2008; Dusenbury et al., 2003). However, there are notable shortcomings to the customary approach. Namely, because composite indices are diffuse measures of fidelity to the overall program, they cannot be used to test associations between specific core components and outcomes. As a result, such measures cannot be used to distinguish active ingredients from those that are less essential.

Creating indices of fidelity to core components can present challenges because most fidelity instruments are designed to quantify fidelity across intervention core components. Thus, computing an index of fidelity to a single core component likely necessitates the isolation and aggregation of a subset of items. In some cases, the combination of items across measures (capturing different domains of fidelity or administered to different informants) with different scaling may be required. To date, there are few guidelines for creating and analyzing indices of fidelity to individual intervention core components. Therefore, in the present study, we used a multifaceted strategy in which we employed three distinct approaches (one customary and two novel) to creating indices of fidelity toward our aims to (a) explicate ways in which fidelity measures can be used to create indices of fidelity to specific core components and (b) illustrate how indices of fidelity to core components can be used to identify the active ingredients of a program. Although our primary goal is not to examine the effectiveness of a specific program or core component, fidelity data collected in the context of a program evaluation provide an ideal context for our illustrative purposes. Specifically, in the context of a larger evaluation of a social–emotional learning intervention, *Responsive Classroom* (*RC*), we compared the predictive utility of the three types of fidelity indices by testing them in relation to achievement gains, an outcome targeted by *RC*, to see if some intervention core components related to outcomes more strongly than others, implicating them as potential active ingredients.

The first, more customary, approach was to create a single composite fidelity index aggregated across all items within a fidelity measure (*intervention composite approach*), which provided a global index of fidelity to the intervention overall. In contrast, we employed two novel approaches that isolated fidelity to individual core components. The second approach involved averaging fidelity ratings from items that capture the same core component across measurement instruments (*core component averaged approach*). The third approach involved deriving factor scores from a multitrait, multimethod factor analytic model (*core component factor score approach*), resulting in fidelity indices that account for shared variance attributable to a core component factor and measure factor. The two novel approaches have advantages and disadvantages. The averaging approach is simpler, thus more accessible to evaluators, but presumes that all items pertaining to an intervention core component should be weighted equally. The factor approach requires more advanced statistical modeling but empirically weights items and yields factor scores that minimize error. Acknowledging the likelihood that intervention components may be present in the control group, we followed recommendations to assess intervention fidelity in both treatment and control groups (e.g., Abry, Rimm-Kaufman, Larsen, & Brewer, 2013; Hulleman & Cordray, 2009; Hulleman, Rimm-Kaufman, & Abry, 2013) and include fidelity indices for both groups in our models.

The Intervention Context

Developed by the Northeast Foundation for Children (NEFC, 2007, 2009), *RC* provides elementary school teachers with a set of principles and practices designed to optimize classroom conditions for academic and social adjustment. Ten *RC* intervention core components focus on building

relationships and classroom community, promoting student accountability and self-regulation, and supporting developmentally appropriate levels of student autonomy.

Previous studies have linked *RC* to student achievement gains directly and indirectly (Rimm-Kaufman, Fan, Chiu & You, 2007; Rimm-Kaufman et al., 2014). These findings show the potential of *RC* to promote achievement but fall short of revealing specific *RC* core components that relate most strongly to academic performance (i.e., the active ingredients), thus providing a perfect case example in which to examine core components in relation to targeted program outcomes. In this article, we focus on the following four *RC* core components: Morning Meeting, Rule Creation, Interactive Modeling, and Academic Choice. These core components were chosen because they characterize fundamental tenets of *RC* and were measured using both observational and teacher-reported assessments. Moreover, each has theoretical and empirical grounding described subsequently in brief.

Morning Meeting is a daily circle time in which teachers participate in and facilitate a student greeting, fun group activity, students' sharing of personal news, and the processing of an interactive message written by the teacher. Morning Meeting has empirical roots in studies linking student–teacher relationships and emotionally supportive classroom climates to academic growth (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008). *Rule Creation* is a collaborative process in which students work with the teacher to create a set of classroom rules that support students in reaching self-identified social and academic goals. *Interactive Modeling* is a multistep process of modeling expectations for routine behaviors that includes a demonstration by the teacher and opportunities for students to practice the behavior and receive feedback. These two core components reflect bodies of work demonstrating links among effective management and achievement (Ponitz, Rimm-Kaufman, Brock, & Nathanson, 2009). *Academic Choice* is a teaching practice through which students are afforded developmentally appropriate levels of autonomy in an effort to promote active engagement in learning. Through Academic Choice, teachers provide students with opportunities to plan, enact, and reflect on the process and content of academic work they have chosen. Such autonomy promoting classroom structures have been associated with student motivation and achievement (Stipek & Weisz, 1981).

Method

Participants

Twenty-four demographically diverse schools from a mid-Atlantic school district were randomized into treatment ($n = 13$) and wait-list control ($n = 11$) conditions after stratifying on the percentage of students eligible for free/reduced price lunch. Teacher participants included 78 fourth-grade teachers from the 24 schools enrolled in their second year of a 3-year longitudinal randomized controlled trial of *RC*. The majority of teachers were female, held a master's degree, and identified themselves as European American. The student sample included 1,442 fourth-grade students attending treatment and control schools. Students were included if they were eligible for the standard versions of the Virginia Standards of Learning (SOL) tests in reading and math. Approximately half of the students were female and identified as European American. About a quarter were eligible for free/reduced price lunch and a third were receiving English language learner (ELL) related services or monitoring. Detailed sample characteristics are provided in Table 1.

Procedures

As part of a larger study of the efficacy of *RC*, data for the present inquiry were collected using classroom observations of teachers' fidelity to *RC* core components, questionnaires administered to teachers regarding their fidelity to *RC* core components, and standardized tests of students' academic achievement, used to test the predictive utility of fidelity indices created using both traditional

Table 1. Teacher and Student Sample Characteristics.

	N	%	*M*	*SD*
Teacher sample ($N = 78$)				
Female	69	91		
Has master's degree	51	73		
European American	67	87		
Hispanic American	3	4		
Other racial/ethnic minority	5	6		
Assigned to intervention condition	39	50		
Age			43	13
Years of teaching experience			11	9
Student sample ($N = 1,442$)				
Female	736	51		
European American	658	46		
Asian American	270	19		
Hispanic American	259	18		
African American	124	9		
Other racial/ethnic minority	125	9		
Eligible for free/reduced lunch	396	27		
Receiving ELL services/monitoring	498	34		
Age			10	.38

Note. ELL = English language learner. Three teachers did not report race/ethnicity.

and novel approaches. All study procedures were approved by the institutional review boards (or equivalent) at the university and in the collaborating district.

Fourth-grade teachers were recruited in the fall of 2007. Response rates were over 95%, and teachers were compensated US$100 for participation in observational and survey data collection. Teachers in the experimental group attended *RC* training institutes during two consecutive summers in 2008 and 2009. In addition, these teachers received in-person coaching support from *RC* personnel throughout the 2008–2009 and 2009–2010 school years and access to *RC* books and materials. Counterparts in the control group received no sanctioned exposure to *RC* training, coaching, or materials and continued "business as usual" instruction.

During the 2009–2010 school year, treatment and control teachers were assessed on their fidelity to *RC* core components during five separate 60-min classroom observations spaced throughout the year. Eight research assistants, not trained in *RC*, conducted the observations after becoming reliable in the coding procedure following the process described subsequently. Observations corresponded to three windows: fall (late September to late November), winter (late November to mid-February), and spring (mid-February to late April). Teachers were observed teaching math instruction in each window and during morning instruction in two of the three windows (not systematically chosen).

Teachers completed online self-report questionnaires on their fidelity to *RC* core components, demographics, and classroom characteristics during a 3-week period at the end of the school year spanning from late April into May.

Baseline measures of achievement were collected at the end of students' third grade year (2008–2009) using the Virginia SOL tests in reading and math.

Posttest measures of achievement were collected at the end of students' fourth grade year (2009–2010) using the fourth grade version of the reading and math SOL tests.

Measures

Observed fidelity. Teachers' observed fidelity to *RC* core components was assessed using the *Classroom Practices Observation Measure* (CPOM). The CPOM (Abry, Brewer, Nathanson,

Sawyer, & Rimm-Kaufman, 2010) is a16-item observational measure assessing adherence and quality of teachers' implementation of *RC* core components. Items were rated on a 3-point Likert-type scale (*not at all characteristic* to *very characteristic*). The measure described *RC* core components in general terms to minimize observer bias. Example items included, "Teacher facilitates students sharing brief, personal news or stories with the rest of the class" and "Students make individualized choices related to an academic lesson or goal." Research assistants scored the 16-item version during morning observations and an abbreviated 10-item version, excluding items pertaining to the Morning Meeting, during math observations (Cronbach's αs in the analytic sample $\geq$.90).

Prior to conducting CPOM observations, research assistants completed a 2-day training upon which they established initial reliability on eight videos master coded by CPOM authors. Exact agreement with master codes exceeded 80% for all coders. Coders conducted between 5 and 15 observations per month throughout the academic year. Ongoing interrater reliability was evaluated via monthly meetings in which coders independently scored a 60-min video observation. Intraclass correlations derived from these scores were greater than .92.

Teacher-reported fidelity. Teachers reported their perceived fidelity to *RC* core components using two instruments. The *Classroom Practices Teacher Survey* (CPTS; Nathanson, Sawyer, & Rimm-Kaufman, 2007a) consisted of 46 items assessing adherence and quality of implementation (e.g., "In the morning we have a class meeting where we sit in a circle facing one another" and "When a rule is introduced, I ask students to model what following the rule looks like."). Teachers were prompted to reflect over the course of the school year and respond to each item on a 5-point Likert-type scale ranging from *not at all characteristic* to *extremely characteristic*. CPTS items were phrased without the use of specific *RC* vocabulary to minimize bias and were administered to teachers in both intervention and control groups (Cronbach's α in the analytic sample = .92).

The *Classroom Practices Frequency Survey* (CPFS; Nathanson, Sawyer, & Rimm-Kaufman, 2007b) was administered concurrently with the CPTS. The CPFS comprised 11 items that assessed the frequency of teachers' use of *RC* core components using an 8-point scale ranging from *almost never* to *more than once per day*. Example items included "When a rule or procedure is introduced, I demonstrate to students how to correctly follow the rule or procedure" and "I provide opportunities for students to choose how to do work, what kind of work to do, or both." CPFS items were administered to intervention and control teachers and used generic phrasing to help minimize response bias (Cronbach's α in the analytic sample = .92).

Creating fidelity indices

Intervention composite indices. Customary indices of intervention fidelity were created by combining fidelity ratings across *RC* core components within each of the three fidelity measures. For the observation measure, ratings from a single observation were averaged to create an observed fidelity score. The five observed fidelity scores were then averaged to create a single indicator of teachers' observed fidelity to *RC* core components. For the teacher-reported measures, responses from the 46 and 11 items, respectively, were averaged. This approach resulted in three different composite indices derived from the three fidelity measures, each indicative of a teachers' overall fidelity to *RC* core components. We refer to these scores as *intervention composite indices*.

Core component indices. Teachers' fidelity to individual *RC* core components was calculated using the two novel approaches (one basic and one more advanced). Twenty-nine items were selected a priori from the three fidelity measures because of their correspondence to four hallmark *RC* core components. Specifically, selected items assessed the major subcomponents of each core component presented as key in *RC* manuals (NEFC, 2007, 2009). For example, selected Morning Meeting items

assessed the presence of a greeting, sharing, activity, and interactive message, and Academic Choice items assessed teachers' implementation of planning, enacting, and reflecting subcomponents.

Averages. In the averaging approach, the 29 items were standardized using *z*-score transformations to account for differences in scaling across the three measures. Scores on items pertaining to an individual core component, taken from all three measures, were then averaged. The resulting Morning Meeting composite consisted of 10 items ($\alpha = .94$), Rule Creation comprised 7 items ($\alpha = .77$), Interactive Modeling comprised 5 items ($\alpha = .81$), and Academic Choice consisted of 7 items ($\alpha = .84$). We refer to these scores as *core component averaged indices*.

Factor scores. In the factor approach, factor scores for the four core components were derived from a multitrait, multimethod confirmatory factor analysis conducted on the 29 *z*-scored items. This method separated variance attributable to a trait from variance attributable to measurement methods (Campbell & Fiske, 1959), thus providing a more rigorous alternative to the core component averaged indices. Figure 1 presents a graphic depiction of the model. Each of the 29 items was allowed to load on one method (i.e., measure) factor and one trait (i.e., core component) factor. For example, the item, "Teacher facilitates students sharing brief, personal news or stories with the class." was loaded on both a measure (i.e., CPOM) factor and a Morning Meeting factor. Measure factors were allowed to correlate, as were core components factors. Measure factors were held orthogonal to core component factors. Residual variances of 2 items from the observational measure were allowed to correlate because a rating of 1 on the first item automatically indicated a rating of 1 on the second item, resulting in a lack of independence.

This model showed adequate fit as indicated by the comparative fit index (.92), the Tucker–Lewis index (.91), the root mean square error of approximation (.06), and the standardized root mean square residual (.08; Bentler, 1990; Hu & Bentler, 1999). Standardized factor loadings are displayed in Figure 1. All core component loadings were significant, and 20 of the 29 loadings were above .60. In contrast, 22 of the 29 measure loadings were significant and only 3 of the loadings were above .60. This model yielded factor scores for the four core component factors, capturing teachers' fidelity to core components while accounting for shared measure variance. These factor scores were retained as indices of teachers' fidelity to individual *RC* core components relative to other teachers in the sample. We refer to these scores as *core component factor score indices*.

Teacher, Student, and Classroom Characteristics

Teachers reported their years of teaching experience and class size on items administered during the spring survey. Student characteristics including sex, eligibility for free and reduced lunch, and ELL status were gathered from students' fourth grade district records. These variables were included as model controls.

Student Achievement Outcomes

Standardized tests of academic achievement were examined as dependent variables in a series of models comparing the three types of fidelity indices created using the methods described previously. Baseline and posttest reading and math achievement were measured using the standard versions of the Virginia SOL in reading and math for Grades 3 and 4, respectively. District testing professionals and classroom teachers administered tests in accordance with district protocols. The reading test comprised 35 multiple-choice items assessing comprehension, word analysis, and use of information resources. The math test comprised 40 multiple-choice items (third grade) or 50 multiple-choice items (fourth grade) assessing number sense; computation and estimation; measurement and geometry; and probability, statistics, patterns, functions, and algebra. Students received a scale score ranging from 200 to 600 for each subject.

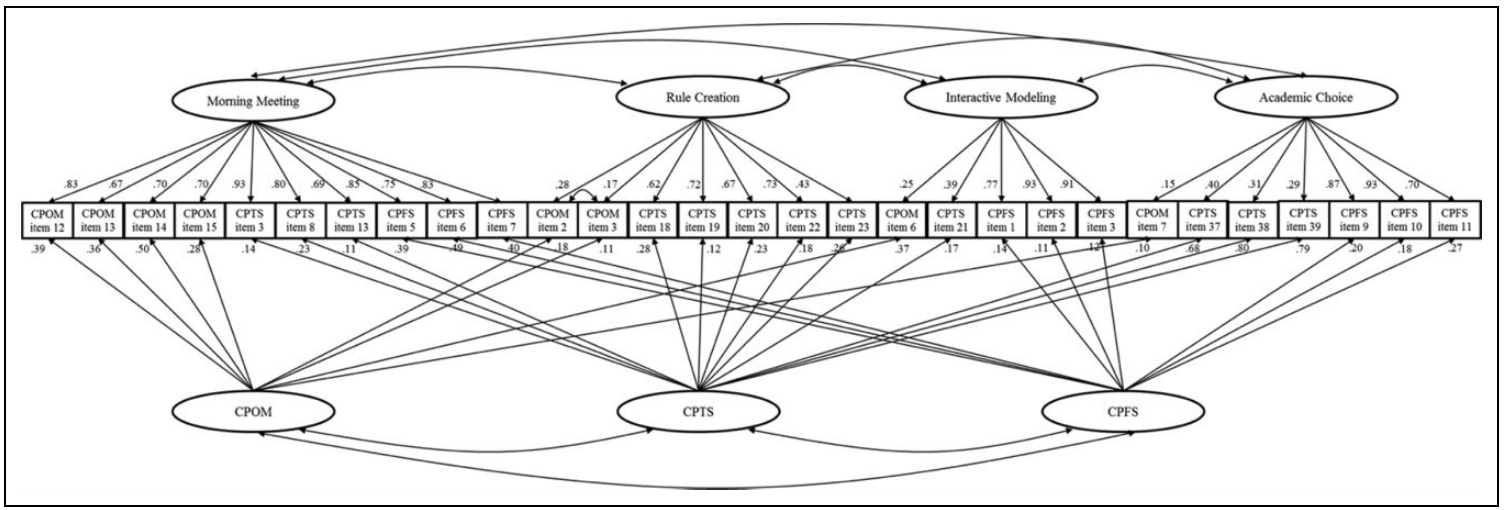

Figure 1. Representation of the multitrait, multimethod confirmatory factor analysis. Standardized factor loadings presented. Duplicate item numbers represent items taken from different measures. CPOM = Classroom Practices Observation Measure; CPTS = Classroom Practices Teacher Survey; CPFS = Classroom Practices Frequency Survey.

Analytic Approach and Data Screening

Three sets of hierarchical regressions were conducted using fidelity indices to predict achievement scores. Separate analyses were conducted for reading and math outcomes to account for different patterns of student nesting within reading and math teachers. Model 1 used the intervention composite indices, Model 2 used the core component averaged indices, and Model 3 used the core component factor score indices. All models were analyzed in Mplus 7.1 (Muthén & Muthén, 1998–2012) using TYPE=COMPLEX TWOLEVEL, which adjusts for the nonindependence among cases at three levels (children nested in classrooms within schools) while maintaining two levels of analysis (i.e., child and teacher). We did not employ a three-level model because the small number of schools in combination with the large number of parameters estimated led to difficulties with model convergence. Intraclass correlations indicated that 22% and 36% of the variability in reading and math outcomes was attributable to the classroom, empirically justifying the need to account for classroom-level nesting. Level 1 intercepts for reading and math scores were treated as random.

Within each model, we tested the relative association of each fidelity index with achievement gains. For example, in Model 1, composite indices derived from the three fidelity measures were simultaneously examined as independent variables. In Models 2 and 3, indices of fidelity (core component averaged indices and core component factor score indices, respectively) to Morning Meeting, Rule Creation, Interactive Modeling, and Academic Choice were simultaneously examined as the independent variables. Covariates were held constant to facilitate comparison across models. At the child level, models controlled for baseline achievement, free and reduced price lunch eligibility, ELL status, and child sex. At the classroom level, models controlled for intervention status, years of teaching experience, and class size. Intervention status was included as a covariate because the sample included teachers in both treatment and control groups. As such, the results are interpreted as the extent to which the use of *RC* core components relates to student achievement gains, independent of study group membership. *RC* treatment by fidelity interactions were not tested because they were outside the scope of the present inquiry and an existing study using an overlapping sample revealed no such interaction effect (Curby, Rimm-Kaufman, & Abry, 2013). Model fit was assessed using classroom-level R^2 (i.e., R^2 between) and was compared across models to determine which type of index had the most explanatory power.

Prior to analysis, data were screened for outliers and for assumptions of linearity, homoscedasticity, and normality. Assumptions of linearity and homoscedasticity were supported and no outliers were detected. However, reading and math scores were not normally distributed; ceiling effects were evident for each variable (12% and 14% of students received the maximum score of 600 for reading and math, respectively). To address the ceiling effect, we employed the Mplus (Muthén & Muthén, 1998–2012) CENSORED command in all analyses, which applies an algorithm that treats a score at the ceiling as a lower bound estimate of the individual's true score. No students were missing data on achievement outcomes. Full information maximum likelihood estimation was used to handle missing data for all independent variables (at most, 8.6% missing for teachers' years of experience and class size), which minimizes bias in parameter estimates and retains the original sample size (Enders, 2001).

Results

Descriptive Statistics and Correlations

Means for the intervention composite indices, derived from the observational and teacher-reported measures, showed that overall fidelity ratings varied across the three measures: means for the observation measure and teacher-reported frequency measure were near the scale midpoint, while the mean from the teacher-reported adherence measure was higher than the scale midpoint. Standard

deviations for all fidelity indices indicated that teachers varied considerably in their fidelity to *RC* core components. The three intervention composite indices showed moderate correlations with one another ($r = .65$ to .80). In contrast, correlations within core component averaged indices ($r = .25$ to .49) and core component factor score indices ($r = .02$ to .41) were lower. The low to moderate correlations between core components demonstrated that teachers implementing one core component with high levels of fidelity did not necessarily implement other core components with equally high fidelity. Table 2 provides descriptive statistics and bivariate correlations for student achievement scores and all indices of fidelity.

Linking Fidelity Indices to Outcomes

Model 1: Intervention composite indices. Table 3 displays results for the three sets of regression analyses. In Model 1-Reading, none of the intervention composite indices predicted reading achievement. In Model 1-Math, the teacher-reported adherence measure (i.e., CPTS) emerged as a significant predictor of math achievement ($\gamma = 15.36$, $p = .04$, $\beta = .40$). Thus, a 1 standard deviation increase in teachers' CPTS scores was associated with four tenths of a standard deviation gain on math scores, equivalent to a 29-point increase. This model explained 20% of the classroom-level variance in reading scores and 21% of the variance in classroom-level math scores.

Model 2: Core component averaged indices. In Model 2, Academic Choice emerged as a significant predictor of gains in both reading ($\gamma = 6.84$, $p < .001$, $\beta = .36$,) and math ($\gamma = 11.34$, $p < .01$, $\beta = .32$). A 1 standard deviation increase on the Academic Choice averaged index was associated with a one-third standard deviation increase on reading and math scores, a 25- and 23-point gain, respectively. The model explained 58% of the variance in classroom-level reading scores and 39% of the variance in classroom-level math scores.

Model 3: Core component factor score indices. In Model 3, Academic Choice again emerged as a significant predictor of gains in both reading ($\gamma = 5.36$, $p = .01$, $\beta = .36$) and math achievement ($\gamma = 10.07$, $p = .02$, $\beta = .36$). A 1 standard deviation increase on the Academic Choice factor score index was associated with a 25- and 26-point gain in reading and math test scores, respectively. The model explained 46% of the classroom-level variance in reading achievement and 36% of the classroom-level variance in math achievement.

Comparing Fidelity Indices and Post Hoc Analyses

Classroom-level R^2 increased substantially from Model 1 to Model 2. Model 2-Reading R^2 (58%) represented a 38% increase over Model 1-Reading (20%); Model 2-Math R^2 (39%) represented an 18% increase over Model 1-Math (21%). Explained variance in Model 3-Reading (46%) and Math (36%) represented a respective 26% and 15% increase over Model 1, but a respective 12% and 3% decrease from Model 2. Change in R^2 between models could not be tested for significance because models were not nested; however, given all other specifications were held constant, we attribute variation in R^2 to the different fidelity indices used.

Two sets of post hoc analyses were conducted. First, given moderate correlations among the three fidelity assessments and the potential for multicollinearity, three follow-up analyses tested each intervention composite index individually in relation to achievement gains. The pattern of findings was consistent with the results of Model 1-Reading and Math.

Second, we tested the robustness of the improved prediction in Models 2 and 3 in light of the increased amount of variance explained in comparison to Model 1. In the creation of the core component averaged and factor score indices, 29 items were chosen to represent the four core components, out of a possible 73. Although these items were selected a priori and independent of their correlations

Table 2. Bivariate Correlations and Descriptive Statistics for Achievement Outcomes and Indices of Fidelity to Core Components.

	Reading	Math	1	2	3	4	5	6	7	8	9	10	11
Intervention composite indices													
1. CPOM	−.13	.08	—										
2. CPTS	−.02	.04	.72*	—									
3. CPFS	−.02	.10	.65*	.80*	—								
Core component averaged indices													
4. Morning Meeting	−.11	−.03	.79*	.87*	.89*	—							
5. Rule Creation	−.16	.10	.47*	.47*	.23*	.26*	—						
6. Interactive Modeling	−.09	.06	.56*	.55*	.71*	.49*	.34*	—					
7. Academic Choice	.32*	.17	.39*	.64*	.62*	.45*	.25*	.31*	—				
Core component factor score indices													
8. Morning Meeting	−.12	−.06	.74*	.82*	.83*	.97*	.18	.45*	.34*	—			
9. Rule Creation	−.08	.08	.21	.25*	.04	.05	.85*	.22*	.11	.02	—		
10. Interactive Modeling	.04	.08	.38*	.42*	.63*	.40*	.20	.89*	.24*	.41*	.17	—	
11. Academic Choice	.28*	.27*	.31*	.50*	.59*	.43*	.23*	.33*	.77*	.39*	.21	.36*	—
N	1401	1422	77	75	75	78	78	78	78	78	78	78	78
M	502.55	512.50	1.62	3.63	3.61	−0.09	−0.04	0.04	0.05	−0.17	−0.14	0.00	−0.06
SD	69.83	71.77	0.33	0.59	1.62	0.89	0.64	0.72	0.66	1.02	0.84	0.93	0.82
Minimum	261	281	1.02	2.41	0.45	−1.32	−1.90	−1.77	−1.57	−1.67	−2.84	−2.11	−1.68
Maximum	600	600	2.30	4.74	6.55	0.98	1.20	1.48	1.67	1.27	1.24	1.43	2.08

Note. CPOM = Classroom Practices Observation Measure; CPTS = Classroom Practices Teacher Survey; CPFS = Classroom Practices Frequency Survey.
*$p \leq .05$

Table 3. Model Results for Three Types of Indices of Fidelity to Core Components Predicting Reading and Math Achievement.

	Reading		Math	
	γ	*SE*	γ	*SE*
Model 1—Intervention composite indices				
CPOM	2.23	11.16	7.74	12.90
CPTS	.64	7.56	15.36*	7.35
CPFS	−.35	2.30	−2.22	2.47
R^2 within	.40		.42	
R^2 between	.20		.21	
Model 2—Core component averaged indices				
Morning Meeting	−4.94	3.38	−11.19	7.44
Rule Creation	−3.62	3.40	5.78	3.95
Interactive Modeling	2.30	3.10	5.93	5.62
Academic Choice	6.84*	1.67	11.34*	3.76
R^2 within	.40		.43	
R^2 between	.58		.39	
Model 3—Core component factor score indices				
Morning Meeting	−3.40	3.28	−9.41	5.02
Rule Creation	−1.39	2.29	.37	3.45
Interactive Modeling	2.07	2.27	2.79	4.65
Academic Choice	5.36*	1.92	10.07*	4.33
R^2 within	.40		.42	
R^2 between	.46		.36	

Note. *SE* = standard error; CPOM = Classroom Practices Observation Measure; CPTS = Classroom Practices Teacher Survey; CPFS = Classroom Practices Frequency Survey.
*$p < .05$.

with achievement scores, it was possible that we had inadvertently selected 29 items that were among those most highly correlated with our outcomes. To test this possibility, we created intervention composite indices, similar to those used in Model 1 but comprised only of the 29 items selected for use in the core component indices. A pattern of results identical to those of Model 1-Reading and Math emerged (i.e., CPTS alone predicted gains in mathematics achievement), strengthening the interpretation that the core component indices outperformed the composite indices.

Discussion

In this study, we demonstrated alternative approaches to the creation of indices of fidelity to specific intervention core components, in contrast to the traditional approach that does not isolate fidelity to individual intervention core components. We then tested the three resulting types of fidelity indices in relation to student achievement gains using hierarchical regression models as a way to compare the predictive utility of each type of fidelity index and demonstrate how fidelity indices can be employed to identify program active ingredients. To illustrate, we used data collected as part of a large-scale evaluation of a social–emotional learning intervention.

Our findings highlight two important points. First, fidelity ratings can be successfully combined across measures—in statistically simple ways, accessible to evaluators—to create reliable indices of implementers' fidelity to individual intervention core components. Second, variability in implementers' fidelity to intervention core components (captured by fidelity indices) can be exploited to identify the components of a program that serve as catalyst for change—the active ingredients. In the

present example, the core component indices derived from the novel approaches (Models 2 and 3) outperformed composite indices derived from the customary approach (Model 1) in the amount of variance explained in achievement outcomes. Further, the pattern of relations between fidelity to core components and achievement gains implicated one *RC* component, Academic Choice (i.e., developmentally appropriate, teacher-structured opportunities for student autonomy in the classroom), as a potential active ingredient. Taken together, this study contributes to the growing field of implementation science by demonstrating that analyses of fidelity to individual intervention core components can be used by evaluators and practitioners alike to better understand underlying program mechanisms, revealing links between intervention components and outcomes that may go unnoticed when relying on customary composite indices of intervention fidelity.

Using Individual Core Components to Identify Active Ingredients

Three findings underscore the value of utilizing core component indices over the customary composite indices. First, the composite fidelity indices used in Model 1 concealed associations between intervention core components and gains in reading achievement apparent in the models that utilized indices of fidelity to individual intervention core components (Models 2 and 3). On the surface, the composite indices suggested that fidelity to core components did not contribute to reading achievement. However, results from Models 2 and 3 showed that the implementation of a specific core component, Academic Choice, was positively and significantly associated with both reading and math gains. The determination of an intervention as ineffective overall when its component parts have not been independently is a variant of what Dobson and Cook (1980) described as a Type III error or making conclusions about a program's overall effectiveness when it has not been properly implemented. Indeed, this finding serves as an important warning to evaluators that when using composite measures of intervention fidelity, a package of core components (i.e., an intervention) could be dismissed as ineffective when, in fact, specific intervention core components affect outcomes in desired ways.

Second, the core component indices explained more variance in student achievement than the composite indices, even though they included fewer than 40% of the total fidelity items. The imprecision of the composite approach may lead to a reduction in explanatory power. Composite measures of fidelity may contain superfluous items that either correlate highly with other items and thus do not explain unique variance in the outcome or correlate weakly with outcome measures. Our results demonstrate that the creation of core component indices of fidelity could aid in the refinement of existing fidelity measures, which could ultimately increase explanatory power and decrease assessment burden by reducing the number of items assessed.

Third, correlations among fidelity to individual *RC* core components were small to moderate, signifying that teachers who implemented one core component with high fidelity do not necessarily implement other components with high fidelity. Such variability in intervention fidelity is obscured when using composite indices that do not differentiate fidelity across individual program core components. Variability in fidelity to core components illustrates the improbability of even uptake of program core components among implementers and lends support to the growing body of work indicating the need for implementation supports in the form of ongoing coaching and formative evaluation (Fixsen, Blase, Naoom, & Wallace, 2009).

Core component averaged indices (Model 2) and core component factor score indices (Model 3) each revealed relations between teachers' use of Academic Choice and achievement gains. A natural question arises: Which approach is better? In the core component averaged indices, the contribution of each item to its respective core component was assumed to be equal, whereas in the factor score indices, the contribution of each item to its core component was empirically estimated. Further, the factor score indices extracted variance shared as a function of belonging to the same source measure. The result was a distilled index of teachers' fidelity to *RC* core components, independent of the measure from which

the item came, and thereby minimizing error in the indices. Despite the psychometric advantages associated with the core component factor score indices, the core component averaged indices used in Model 2 explained slightly more variance in achievement gains. Moreover, the averaged indices are substantially easier to compute and interpret. Thus, we contend that the averaged indices may be more desirable, given their accessibility to practitioners and evaluators. Further research is needed to determine conditions in which the factor score approach may be more appropriate.

Implications and Applications for the Identification of Active Ingredients

Findings from this study and others seeking to identify program active ingredients represent an important first step by pointing to specific practices that warrant further investigation. For example, in the context of this study, a valuable next step would be to conduct further examination of Academic Choice under more rigorous designs (Collins et al., 2005). If a potential active ingredient continues to predict outcomes in the anticipated way in quasi-experimental and experimental designs, then the component warrants special consideration by program developers, practitioners, and policy makers as an effective practice. On the other hand, if associations between a core component and targeted outcomes are not observed, it may indicate an inactive ingredient and the need to reevaluate its emphasis or inclusion in the treatment model. Armed with knowledge of active (and inactive) ingredients, program developers can refine an intervention to optimize its impact. For instance, cumulative evidence on Academic Choice could prompt program developers to increase its emphasis in teacher training and bolster implementation supports to ensure teachers' use of Academic Choice in their classrooms. Inclusion and emphasis of other core components would be adjusted based on evidence of their relations across an array of targeted outcomes. Distinguishing active and inactive ingredients is of particular use to evaluators seeking to identify thresholds for intervention fidelity. To the extent that active ingredients have been identified, thresholds can be based on fidelity to essential, rather than nonessential, intervention components.

The value of identifying program active ingredients extends beyond informing intervention-specific decisions. Knowledge of active ingredients can help practitioners and developers to be more efficient and resourceful in program selection and development processes. With a solid understanding of evidence-based active ingredients relevant to a given focus (e.g., social and emotional learning, health promotion, drop-out prevention), developers would be better equipped to create integrated interventions that combine active ingredients into a single optimized package (e.g., Domitrovich et al., 2010). Integrated interventions are hypothesized to be more sustainable, potent, and effective at promoting a host of positive effects in contrast to programs implemented in isolation that may target very specific behavioral or academic outcomes (Domitrovich et al., 2010). Likewise, practitioners would be better positioned to select existing programs that integrate essential elements, which vary in size and scope. For example, Embry and Biglan (2008) identified and described 52 evidence-based "kernels" or "fundamental units of behavioral influence" (p. 75) demonstrated, over time, through experimental trials to reliably affect behavior. Li and Julian (2012) describe developmental relationships as an active ingredient of interventions focused on at-risk youth. Kernels represent fundamental units that cannot be further reduced while retaining their impact, while developmental relationships represent a coarser example of an active ingredient. As evidence of active ingredients accumulates, practitioners will be better able to choose wisely among the myriad of available programs and practices across intervention types and foci. In essence, knowledge of active ingredients potentiates more efficient and effective intervention. Ultimately, as evidence on active ingredients accumulates, we will make progress toward an important long term goal: Knowledge of active ingredients will contribute to the development of unified theories of change (Embry & Biglan, 2008). The unification process will impact practice, resulting in a shift away from the burden of implementing many different programs and toward the uptake of a coherent set of theoretically and empirically grounded practices.

Limitations and Future Directions

Three limitations require mention. First, this study focused on intervention core components, which represent the behaviors, actions, and materials that differentiate a program from business as usual. Intervention core components, operationalized in this way, are distinct from implementation drivers, such as training, ongoing technical support, performance evaluation, administrative support, and sustainability efforts that also bear on fidelity (Fixsen et al., 2009). Although the investigation of implementation drivers is a critical element of implementation research, their evaluation was outside the scope of this study focused on specific practices implemented by program users. Second, as is the case with analyses of the treatment-on-the-treated, our findings do not support a causal link between Academic Choice and achievement gains. There may be variables that explain both teachers' use of Academic Choice and achievement. However, correlational analyses such as those presented here are necessary to highlight components that deserve further experimental examination (Collins et al., 2005). Third, we analyzed four of the 10 intervention core components chosen because they represent key tenets of the intervention model, were assessed via observation and teacher report, and are grounded in educational psychology research. Still, a comprehensive understanding of any program's active ingredients would require analyses of all intervention core components in relation to a variety of outcomes before definitive conclusions about active ingredients or refinements to the program were to be made.

An important future direction would be addressing variability in fidelity across intervention core components. One hundred percent fidelity to the treatment model is rarely, if ever, achieved (Durlak & Dupre, 2008). Uneven fidelity to core components raises questions regarding implementer and contextual characteristics influencing implementation, especially those that are linked to consistency in uptake across components identified as potential active ingredients. For example, prior research has identified administrative/organizational support, implementers' self-efficacy to implement the program, and training and ongoing technical support (Rohrbach, Grana, Sussman, & Valente, 2006) as factors influencing fidelity. Future work should examine these factors, as they relate to fidelity to individual intervention core components.

Also important could be the systematic deconstruction of widely used curricular, preventive, and health promotion programs into their core component parts in order to evaluate their relation to measured outcomes, and ultimately identify active ingredients. To facilitate this process on a larger scale, evaluators are urged to be thoughtful in their creation of fidelity measures. To the extent that fidelity assessments can be used to readily transform ratings into indices of fidelity to intervention core components, the burden of figuring out how to create such indices post hoc will be lessened. Only when the identification of active ingredients is pursued on a comprehensive scale, developers, practitioners, and researchers might really begin to synthesize knowledge of active ingredients and apply them toward the creation of optimized interventions and effective practice. As researchers pursue these next steps, qualitative and mixed-method studies can contribute valuable insight into implementation processes.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. Inquiries regarding these data should be addressed to the third author.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through grant R305A070063 to the third author, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B090002 to the University of Virginia, and by the National Science Foundation through grant DRL#1252463 to the second author.

References

Abry, T., Brewer, A., Nathanson, L., Sawyer, B., & Rimm-Kaufman, S. E. (2010). *Classroom practices observation measure*. Unpublished instrument. Charlottesville, VA: University of Virginia.

Abry, T., Rimm-Kaufman, S. E., Larsen, R. A., & Brewer, A. J. (2013). The influence of implementation fidelity on teacher-student interactions in the context of a randomized controlled trial of the Responsive Classroom approach. *The Journal of School Psychology*, *51*, 437–453. doi: http://dx.doi.org/10.1016/j.jsp.2013.03.001

Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. doi: http://escholarship.org/uc/item/6cn677bx

Bruns, E. J., Suter, J. C., & Leverentz-Brady, K. (2008). Is it wraparound yet? Setting quality standards for implementation of the wraparound process. *The Journal of Behavioral Health Services & Research*, *35*, 240–252. doi:10.1007/s11414-008-9109-3

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. doi:10.1037/h0046016

Century, J., Cassata, A., Rudnick, M., & Freeman, C. (2012). Measuring enactment of innovations and the factors that affect implementation and sustainability: Moving toward common language and shared conceptual understanding. *The Journal of Behavioral Health Services & Research*, *39*, 343–361. doi:10.1007/s11414-012-9287-x

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*, 199–218. doi:10.1177/1098214010366173

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, *30*, 65–73. doi:10.1207/s15324796abm3001_8

Curby, T. W., Rimm-Kaufman, S. E, & Abry, T (2013). Do emotional support and classroom organization earlier in the year set the stage for higher quality instruction? *Journal of School Psychology*, *51*, 557–569. doi: http://dx.doi.org/10.1016/j.jsp.2013.06.001

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

Dobson, L., & Cook, T. (1980). Avoiding type III error in program evaluation: Results from a field experiment. *Evaluation and Program Planning*, *3*, 269–276. doi:10.1016/0149-7189(80)90042-7

Domitrovich, C. E., Bradshaw, C. P., Greenberg, M. T., Embry, D., Poduska, J. M., & Ialongo, N. S. (2010). Integrated models of school-based prevention: Logic and theory. *Psychology in the Schools*, *47*, 71–88. doi: 10.1002/pits.20452

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. doi:10.1007/s10464-008-9165-0

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development 82*, 405–432. doi:10.1111/j.1467-8624.2010.01564.x

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237–256. doi:10.1093/her/18.2.237

Embry, D. D., & Biglan, A. (2008). Evidence-based kernels: Fundamental units of behavioral influence. *Clinical Child and Family Psychology Review*, *11*, 75–113. doi:10.1007/s10567-008-0036-x

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128–141. doi:10.1207/S15328007SEM0801_7

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*, 531–540. doi:10.1177/1049731509335549

Greenberg, M. T. (2010). School-based prevention: Current status and future challenges. *Effective Education*, *2*, 27–52. doi:10.1080/19415531003616862

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118

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 Educational Effectiveness*, *2*, 88–110. doi: 10.1080%2F19345740802539325

Hulleman, C. S., Rimm-Kaufman, S. E., & Abry, T. (2013). Whole-part-whole: Construct validity, measurement, and analytical issues for fidelity assessment in education research. In T. Halle, A. Metz, & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 65–93). Baltimore, MD: Paul H. Brookes Publishing Co.

Irwin, M., & Supplee, L. H. (2012). Directions in implementation research methods for behavioral and social science. *The Journal of Behavioral Health Services & Research*, *39*, 339–342. doi:10.1007/s11414-012-9293-z

Li, J., & Julian, M. M. (2012). Developmental relationships as the active ingredient: A unifying working hypothesis of "What works" across intervention settings. *American Journal of Orthopsychiatry*, *82*, 157–166. doi:10.1111/j.1939-0025.2012.01151.x

Lunn, L. M., Heflinger, C. A., Wang, W., Greenbaum, P. E., Kutash, K., Boothroyd, R. A., & Friedman, R. M. (2011). Community characteristics and implementation factors associated with effective systems of care. *The Journal of Behavioral Health Services & Research*, *38*, 327–341. doi:10.1007/s11414-011-9244-0

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

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Nathanson, L., Sawyer, B., & Rimm-Kaufman. (2007a). *Classroom practices teacher survey*. Unpublished instrument. Charlottesville, VA: University of Virginia.

Nathanson, L., Sawyer, B., & Rimm-Kaufman. (2007b). *Classroom practices frequency survey*. Unpublished instrument. Charlottesville, VA: University of Virginia.

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*, *39*, 374–396. doi:10.1007/s11414-012-9295-x

Northeast Foundation for Children. (2007). *The responsive classroom level 1 resource book*. Turner Falls, MA: Author.

Northeast Foundation for Children. (2009). *The responsive classroom level II resource book*. Turner Falls, MA: Author.

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. doi:10.3102/0034654307313793

Pas, E. T., & Bradshaw, C. P. (2012). Examining the association between implementation and outcomes. *Journal of Behavioral Health Services & Research*, *39*, 417–433. doi:10.1007/s11414-012-9290-2

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*, 365. doi:10.3102/0002831207308230

Ponitz, C. C., Rimm-Kaufman, S. E., Brock, L. L., & Nathanson, L. (2009). Early adjustment, gender differences, and classroom organizational climate in first grade. *Elementary School Journal*, *110*, 142–162. doi:10.1086/605470

Rimm-Kaufman, S. E., Fan, X., Chiu, Y. J., & 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. doi:10.1016/j.jsp.2006.10.003

Rimm-Kaufman, S. E., Larsen, R. A., Baroody, A. A., Curby, T. W., Ko, M., Thomas, J. . . . DeCoster, J. (2013). Efficacy of the Responsive Classroom approach: Results from a three-year longitudinal randomized controlled trial. *American Education Research Journal*, *51*, 567–603. doi: 10.3102/0002831214523821

Rohrbach, L. A., Grana, R., Sussman, S., & Valente, T. W. (2006). Type II translation: Transporting prevention interventions from research to real-world settings. *Evaluation & the Health Professions*, *29*, 302–333. doi: 10.1177/0163278706290408

Salyers, M. P., Bond, G. R., Teague, G. B., Cox, J. F., Smith, M. E., Hicks, M. L., & Koop, J. I. (2003). Is it ACT yet? Real-world examples of evaluating the degree of implementation for assertive community treatment. *The Journal of Behavioral Health Services & Research*, *30*, 304–320. doi:10.1007/BF02287319

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

Sidani, S., & Sechrest, L. (1999). Putting program theory into operation. *American Journal of Evaluation*, *20*, 227–238. doi:10.1177/109821409902000205

Stipek, D. J., & Weisz, J. R. (1981). Perceived personal control and academic achievement. *Review of Educational Research*, *51*, 101–137. doi:10.3102/00346543051001101