American Journal of Evaluation
33(4) 547-565

DOI: 10.1177/1098214012452715
http://aje.sagepub.com
SAGE

How Does Fidelity of Implementation Matter? Using Multilevel Models to Detect Relationships Between Participant Outcomes and the Delivery and Receipt of Treatment

Keith Zvoch[1]

Abstract

Multilevel modeling techniques facilitated examination of relationships between fidelity indicators and outcomes associated with a summer literacy intervention. Three-level growth models were specified to capture the extent to which students experienced instruction and to demonstrate the ways in which dosage–response relationships manifest in program evaluation contexts. The observation that outcome-related deviations from program protocol occurred both at the provider and at the recipient levels suggests that evaluators will often need to conceptualize, measure, and model "treatment fidelity" as a multilevel, multidimensional construct.

Keywords

treatment fidelity, summer school, summer learning, early childhood literacy

Awareness of the importance of monitoring and documenting the implementation of educational, health, and social service interventions is on the rise. As evidenced by recent empirical review articles (Durlack & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003; O'Donnell, 2008), a spate of conceptual and integrative pieces in a journal special issue (Greenwood, 2009; Gresham, 2009; Hagermoser Sanetti & Kratochwill, 2009; Schulte, Easton, & Parker, 2009), and the updated guidelines of major funding agencies (e.g., Institute for Education Sciences, 2009), a sea change in attitude has ushered in a newfound appreciation of the limits imposed by ''black box'' program evaluation and context free field-based intervention research. Current awareness of ''the

[1] University of Oregon, Eugene, USA

Corresponding Author:
Keith Zvoch, University of Oregon, Department of Educational Methodology, Policy and Leadership, College of Education, 5267 University of Oregon, Eugene, OR 97403, USA
Email: kzvoch@uoregon.edu

implementation issue'' stems from the seminal thinking of methodologists (Sechrest, West, Phillips, Redner, & Yeaton, 1979; Yeaton & Sechrest, 1981) and policy analysts (Berman & McLaughlin, 1976) and the frontline work of evaluators (Fullan & Pomfret, 1977; Hall & Loucks, 1977; Hall, Loucks, Rutherford, & Newlove, 1975) who first recognized that the conceptualization and delivery of an intervention was often not consonant, and associated treatment effect estimates were thus ambiguous and misleading. Systematic examination of samples of the treatment intervention literature has also focused attention on the topic by documenting the absence of program implementation data in many published studies (Borrelli et al., 2005; Dane & Schneider, 1998; Hagermoser Sanetti & Kratochwill, 2008; Moncher & Prinz, 1991; Peterson, Homer, & Wonderlich, 1982). Finally, heightened attention to the development and specification of program theory also raises awareness among the evaluation community. Increasingly, comprehensive evaluation studies require linking program ingredients and outcomes and the explicit monitoring of key intervention components (Chen, 1990; Donaldson, 2007; Scheirer, 1987).

The methodological and inferential concerns first articulated several decades ago and more recent appreciation of the evaluative value offered by a richly contextualized treatment landscape has contributed to an effort to conceptualize, operationalize, and provide a framework for identifying and measuring program implementation components (Bellg et al., 2004; Century, Rudnick, & Freeman, 2010; Dane & Schneider, 1998; Mowbray, Holter, Teague, & Bybee, 2003; Schulte et al., 2009). In recent years, ''treatment fidelity'' has developed as a multidimensional construct that reflects not only the degree to which providers deliver an intended treatment, program, or service, but also the extent to which targets receive and interact with treatment components (Bellg et al., 2004; Schulte et al., 2009; Shadish, Cook, & Campbell, 2002). The delivery, receipt, and adherence/enactment conceptualization serves to outline the broad contours of the fidelity construct and highlight the unique role of providers and recipients in the implementation and use of intervention components. Associated subdimensions, including the extent to which a provider delivers the range of treatment components (integrity/adherence) along with the strength (dosage/exposure) and skill of delivery (quality), further identify and characterize the provider role (Dane & Schneider, 1998; Dusenbury et al., 2003). In modern conceptualizations, variation in treatment receipt and protocol enactment also matters, as an intervention can be delivered with a high degree of skill and integrity but program participants still may not receive or interact with the treatment as intended. Receipt and adherence/enactment breakdowns occur when program participants are not engaged during treatment delivery, fail to comprehend or follow through on treatment-related protocols, and/or intermittently attend treatment sessions (Bellg et al., 2004; Borrelli et al., 2005; Jones, Clark, & Power, 2008; Schulte et al., 2009).

In light of the number of identified provider and recipient level fidelity components and the diversity of fields in which program evaluations take place, it is important to note that there is not a general consensus regarding the use of terms (Hagermoser Sanetti & Kratochwill, 2009; O'Donnell, 2008; Schulte et al., 2009). For example, the health and prevention science fields may document a provider's fidelity to a treatment model and determine the extent to which a client received the treatment and followed an associated treatment protocol. Conversely, in an educational evaluation, the ''treatment'' may be a new curriculum or teaching practice, making the distinction between the delivery and receipt of an instructional treatment less clear, and reference to a treatment protocol less common. Across fields and evaluation contexts, there may also be a greater preference and need for documenting the skill and competence with which a treatment regime is delivered and more or less focus on capturing the adaptation of protocols by program implementers (Mowbray et al., 2003). Moreover, in cases where treatments or programs are compared with one another or a standard treatment regime, the differentiation of common and intervention specific components yields another dimension on which provider and receipt roles and actions must be clarified (Century et al., 2010; Dane & Schneider, 1998; Hulleman & Cordray, 2009).

Despite the complexity associated with uncovering and unpacking the ''black box'' of program operations, ongoing efforts to conceptualize and develop the treatment fidelity construct (Bellg et al., 2004; Century et al., 2010; Dane & Schneider, 1998; Schulte et al., 2009) have contributed to a broadened understanding of the manner in which implementation failures occur and stimulated the development and use of a variety of techniques for measuring provider and recipient-related implementation breakdowns. Across the social and health science fields, daily logs, self-report inventories, observational protocols, and provider interviews have generated a wealth of data on the delivery and receipt of critical program components (Durlack & DuPre, 2008; Dusenbury et al., 2003; Mowbray et al., 2003). The technical properties of various fidelity measures, including the reliability of observations and the construct validity of treatment fidelity scores, have also been a focus of investigation (e.g., McGrew, Bond, Dietzen, & Salyers, 1994; McKenna, Rosenfield, & Gravois, 2009; Mowbray, Bybee, Holter, & Lewandowski, 2006). In all, developments in the conceptualization and measurement of treatment delivery and receipt have highlighted a range of assessment advances and challenges and shed further light on the complex and varied nature of the fidelity construct. Yet, despite recent psychometric gains in measure quality, a divide between the conceptual and empirical currently remains, as less progress has been made integrating modern representations of the treatment fidelity construct with statistical models that allow the testing of hypothesized relationships fidelity components and program participants' outcomes (Century et al., 2010; Mowbray et al., 2003).

When data are collected on the delivery, receipt, and the adherence to/enactment of key program model components, evaluators have an opportunity to examine the degree to which specific variations in treatment fidelity directly or indirectly relate to recipient outcomes. For example, the availability of data that represents variation in the delivery of key program components enables evaluators to test the oft-considered hypothesis of whether provider differences in the fidelity to the program model associate with variation in recipient outcomes. Likewise, data on the degree to which program participants received the treatment and adhered to aspects of a treatment protocol enable examination of linkages between engagement levels and individual outcomes. The degree to which variation in program delivery moderates the relationship between the recipients' engagement and their outcomes may also be particularly relevant in many evaluation contexts. In recent years, some attempts at documenting statistical relationships between the components of treatment fidelity and program outcomes have been made (e.g., Bloom, Hill, & Riccio, 2003). The majority of effort has focused on estimating main effect relationships between treatment delivery indices and recipient outcomes (Durlack & DuPre, 2008; Noell, 2008; O'Donnell, 2008; Zvoch, Letourneau, & Parker, 2007) while others have treated variations in treatment fidelity as an outcome to be predicted by characteristics of treatment providers and/or treatment context (Durlack & DuPre, 2008; Dusenbury et al., 2003; Zvoch, 2009). These studies have contributed to a nascent understanding of when fidelity to the program model is particularly germane and offered insight into the range of individual and environmental characteristics that support or inhibit implementation fidelity. However, empirical studies designed to explicitly separate the components of fidelity into those associated with aspects of program delivery and those associated with aspects of program receipt for the purpose of examining interrelationships with recipient outcomes have been fewer in number (National Research Council [NRC], 2004).

The testing of hypotheses regarding fidelity components and recipient outcomes challenges evaluators because relatively complex statistical models are often necessary to estimate relationships between measures of the multidimensional fidelity construct and the individual outcomes of interest. At present, multilevel statistical models afford the best means to represent the data structures that typically arise when multiple fidelity indicators are used to characterize the manner in which providers deliver an intervention to recipients. The use of multilevel modeling techniques has several advantages over traditional single level regression or analysis of variance models. The ability to

separate and model variance occurring at each level of the data structure is a particular strength (Raudenbush & Bryk, 2002). The separation of recipient, provider, and site-level variance allows the evaluator to ascertain the manner in which outcome variation is distributed and to estimate within and between level relationships among key fidelity components and treatment outcomes. Another important advantage in treatment intervention contexts is the ability to represent longitudinal panel data (Raudenbush & Bryk, 2002). When the multilevel model is extended to incorporate outcome data obtained from the repeated observation of individual program participants over time, it is possible to estimate whether intra- and inter-individual change over the intervention period is related to the degree to which treatment was delivered and received.

Recent efforts to further conceptualize and measure the multilevel, multidimensional treatment fidelity construct have promoted greater awareness of the role that the delivery and receipt of treatment and adherence to/enactment of a treatment protocol plays in the evaluation of program effects. Yet, a divide between the measurement of key fidelity components and the empirical testing of hypothesized relationships remains. In response, this article demonstrates how multilevel modeling techniques can be used to estimate within and between level relationships between fidelity indicators and treatment outcomes. The analysis uses a variety of delivery, receipt, and adherence indicators to represent aspects of the fidelity construct. The applied example is that of a school-based 5-week summer literacy intervention delivered to struggling readers in a small group setting. As a school-based example, the ''treatment'' is an instructional intervention, the ''providers'' are teachers, and the ''recipients'' are students. Measured fidelity components captured in part the degree to which the literacy instruction was delivered and received as intended, and the extent to which an associated instructional ''protocol'' was enacted. Multilevel growth models were applied to program data to (1) estimate the initial status and literacy growth of summer school students, (2) model the intra-individual relationship between biweekly literacy performance levels and biweekly fluctuations in the receipt of intervention, (3) ascertain the extent to which individual differences in average treatment receipt and adherence levels were predictive of literacy growth, (4) determine the extent to which variation in the delivery of the instructional treatment was associated with the average growth rate of instructional groups, and (5) estimate the extent to which variation in the delivery of treatment was predictive of average group differences in the relationship between treatment receipt levels and literacy growth outcomes. The assumption underlying the demonstration was that higher fidelity levels whether in the delivery, receipt, or adherence to treatment would be associated with more positive literacy outcomes.

Method

Data Source

Data came from a university–school district collaboration to evaluate a district-sponsored summer literacy intervention. Both the university and the school district are located in a moderately sized city in the Pacific Northwest. The school district serves close to 6,000 students each year. The student body is approximately 75% White, 14% Latino, 3% African American, 3% Asian American, 3% Native American, and 2% Other. In recent years, about 44% of district students received a free or reduced price lunch and 3% of students were identified as English language learners.

Intervention

As a means to support struggling readers over the summer vacation period, the district established an academically intensive summer literacy program. Summer instructional programs are considered a targeted and cost-effective approach to offset the ''summer slide'' in achievement that is commonly observed among youth from disadvantaged backgrounds (Borman & Dowling, 2006; Cooper,

Charlton, Valentine, & Muhlenbruck, 2000; Zvoch & Stevens, 2011). The theory of change underlying the district's summer school initiative follows the "faucet theory" of learning, which suggests that distinct seasonal learning patterns stem from educational resources available at different points in the calendar year (Alexander, Entwisle, & Olson, 2001; Entwisle, Alexander, & Olson, 1997). During the academic year, the educational resource faucet is turned on allowing all children the opportunity to learn. However, the resource flow differs for students in the summer when school is not in session. While students of advantage continue to have access to a flow of home- and community-based educational resources, disadvantaged students tend to have reduced access to the educational opportunities that facilitate successful academic year learning. The result is the well-documented slowing or fallback in learning among disadvantaged youth over the summer vacation period (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Heyns, 1978, 1987).

In an attempt to offset the unequal distribution of educational resources and prevent the initial reading difficulties of struggling readers from progressing, the district provides targeted literacy instruction each summer to early elementary school students most in need of supplemental support. The supplemental literacy program runs over a 5-week period during the middle of the 3-month summer vacation period. Instruction is delivered 3.5 hr/day four mornings per week in one central school site. Students receive approximately 2 hr of teacher-directed daily literacy instruction in the critical beginning reading skills of phonemic awareness (oral blending and segmentation), alphabetic understanding (letter-sounds, decoding, phonic analysis), and fluency/automaticity (speed and accuracy in reading connected text). Literacy instruction is delivered in small group settings (~ 5 students per instructor) during two 45-min blocks separated by a short recess. Students are grouped based on skill and reading level, and remain with the same instructor throughout the summer. The instructional model closely aligns with the fundamental "big ideas" and best practices associated with early childhood literacy development (National Reading Panel, 2000).

Treatment Recipients

During the 2010 summer vacation period, supplemental literacy instruction was delivered to struggling readers ($N = 124$) who completed either the first ($n = 81$) or second grade ($n = 43$) during the 2009–2010 school year. Fifty-one percent of the sample was female ($n = 63$). In addition, 74% ($n = 92$) of summer school participants received a free or reduced price lunch during the previous academic year, 20% ($n = 25$) identified with an ethnic minority group, and 15% ($n = 18$) were classified as English language learners.

Treatment Providers

Literacy instruction was delivered to students in small group settings by 24 instructors located in one central school site. Twenty-two of the 24 instructional groups consisted of students from the same grade. In the other two groups, one very low scoring second grader was placed with first-grade students. The two second graders who were placed in a first-grade group had the lowest scores observed among the sample of second grade students. All but one of the 24 small group instructors was female. Instructors were regular classroom teachers and highly trained educational assistants. Program administrators selected summer literacy instructors through a competitive skill-based application process. Instructors in the summer literacy program underwent additional professional development training throughout the academic year and prior to the start of the program

Outcome Measure

Scores on the Test of Oral Reading Fluency (TORF; Children's Educational Services, 1987) determined a student's eligibility for a summer school placement and were used to measure subsequent changes in oral reading fluency. The TORF is a standardized, individually administered test of

accuracy and fluency with connected text that is designed to identify struggling readers and allow the monitoring of student progress over time. Reported test–retest and alternate-form reliability estimates have ranged from .89 to .97 (Tindal, Marston, & Deno, 1983). Each academic year, the school district administers the TORF to first- and second-grade students in September, January, and May. First-grade students who correctly read fewer than 30 words per minute and second-grade students who correctly read fewer than 70 words per minute on the May assessment are considered to be at heightened risk for future reading difficulty. These students are offered the opportunity to attend the summer instructional program. During the summer, program participants completed the TORF three additional times, during the first, third, and fifth week of the 5-week intervention period.

Individual-Level Predictor Variables

At the individual level, covariates included grade level, initial literacy status, and demographic characteristics of students as well as the extent to which summer school participants received and adhered to the summer instructional protocol. Dummy codes were used to identify girls, students from ethnic minority groups, English language learners, and free or reduced price lunch recipients. Students who completed second grade during the 2009–2010 academic year were also identified with a dummy code. The initial literacy status of students was defined by their TORF score ($M = 37.01$, $SD = 17.48$) at the onset of the intervention.

Dosage variables included the level of engagement in the daily literacy lessons and the proportion of homework assignments completed. Student engagement was indexed by a teacher rating of the extent to which each student in each small group was an attentive and active participant in each of the two daily literacy lessons. Student engagement was rated on a 10-point scale ranging from *not engaged* (1) to *highly engaged* (10). An average engagement score was computed by combining teacher ratings within and across instructional days (i.e., 32 literacy lessons across 16 instructional days where student engagement was recorded; $M = 7.77$, $SD = 1.32$). Homework completion was computed by dividing the number of daily homework assignments completed (one per student per day) by the total number assigned ($N = 16$; $M = .54$, $SD = .24$).

In the statistical models described below, student attendance was also specified as a dosage-related covariate. Student attendance was specified in two forms. Attendance was first treated as a time-varying covariate in an attempt to account for intra-individual deviations from expected performance at each TORF assessment point. The time-varying attendance covariate was represented as the proportion of days attended during each assessment period (week 1: $M = .85$, $SD = .27$; weeks 2–3: $M = .89$, $SD = .17$; weeks 4–5: $M = .84$, $SD = .26$). The proportion of days attended across the entire intervention period was also computed for each student ($M = .86$, $SD = .15$). Mean attendance served to represent inter-individual differences in instructional receipt.

Group-Level Predictor Variables

In addition to the range of individual-level background and dosage variables, variables representing instructor/instructional group characteristics and instructional process and practice (i.e., treatment delivery) served to index aspects of the treatment context. A dummy code was used to contrast the referent group of classroom teachers (25% of the sample, $n = 8$) with the group of highly trained educational assistants (75%, $n = 16$). The number of students per instructional group ($M = 5.17$, $SD = 1.05$) was used to capture the teaching ''load'' experienced by each instructor. The initial literacy status of instructional groups as defined by the group mean TORF score ($M = 36.11$, $SD = 15.47$) at the onset of the intervention served as a control for the skill-based grouping of students. Data on instructional process and practice came from observations of the fidelity with which literacy instruction was delivered to students. Each instructor was observed on two occasions during the course of the treatment period by a nationally certified direct instruction trainer

using an observational protocol with an anchored rating scale. A global fidelity of implementation score for each instructor was computed by combining the observational ratings across the two observation periods as described below.

Observational Protocol

The observational protocol enabled observers to monitor and record key aspects of program delivery and instructional presentation. School district personnel and university faculty with extensive knowledge of the common literacy components underlying each of three literacy programs created the observation protocol. Core components common to each program were those dealing with the organization and use of instructional materials, correct lesson wording and pacing, appropriate modeling of reading skills and strategies, allowance for student reading practice, and the provision of individual feedback to correct errors and positively reinforce correct student responses. In all, the observational protocol consisted of 13 items designed to capture the core instructional elements that were to be implemented during each literacy lesson. Items were represented on a 4-point scale where 0 = *program element absent or not observed*, 1 = *inconsistent level of implementation*, 2 = *high level of implementation*, and 3 = *expert level of implementation*. Each item and each scale point was operationalized in an accompanying scoring rubric. Example items include ''teacher follows explicit lesson wording,'' ''teacher models skills/strategies,'' ''teacher uses clear signals,'' ''teacher provides students opportunities to respond,'' and ''teacher corrects student errors.''

Observers

School district personnel ($N = 5$) conducted the small group observations. Observers were the nationally certified trainers who provided the professional development to literacy instructors during the academic year and prior to the start of the summer intervention. Observers worked closely with university faculty to construct the observational protocol and scoring rubric, and thus were well acquainted with the items and rating scales. Observations were conducted on two separate occasions (weeks 2 and 4) during the 5-week intervention period. Each observation was conducted during one of the 45-min daily literacy blocks. On three occasions, the entire five-member observation team observed the same instructor on the same day at the same time. Ratings from the joint observations were used to examine the degree of interrater agreement. The observation team was remarkably consistent in assigning ratings. Across the 13 observational items, a 1-scale point disagreement was recorded across observers on only three of the items. Otherwise, each observer assigned the same rating on each item during each instructional observation.

Implementation Index

Observers rarely assigned the lowest or highest implementation score on any of the items during either observation point. As a result, the absent and inconsistent (0 and 1) and the high and expert (2 and 3) level implementation ratings were combined into a low (0) and high (1) categorization before computing a global fidelity of implementation score for each instructor. With a total of 26 ratings across two observation periods (i.e., 13 items $\times$ 2 observations of each instructor), the implementation score was computed as the total number of instructional components that could have been delivered with high fidelity ($N = 26$) divided by the sum of the number of ''high'' fidelity ratings obtained across the two observation periods and multiplying by 100. The percentage of components implemented with high fidelity ranges from 0% to 100%. Computed implementation scores ranged from 42% to 100% with a mean of 84.9% and a standard deviation of 14.7%, reflecting a high fidelity, low variance delivery of instruction.

Analytic Procedures

Relationships between various treatment fidelity components and the reading fluency outcomes of students were examined by applying unconditional and conditional three-level growth models to the longitudinal panel data obtained on struggling readers over the course of the summer intervention period. A multilevel modeling approach explicitly accounted for the dependencies in the data that arose from the nesting of observations (i.e., summer literacy scores) within students and the nesting of students within skill-based instructional groups. The use of multilevel models enabled estimation of individual growth trajectories and separation of intra-individual, inter-individual, and group-level variance. The preservation of individual and group differences through the correct partitioning of variance components permitted analytic flexibility to control individual and group characteristics and estimate individual and group-level relationships between the delivery and receipt of instruction and summer literacy outcomes. The separation of variance also enabled examination of cross-level interactions between provider differences in the delivery of instruction and the relationship between student receipt of instruction and student learning.

Unconditional and conditional growth models were estimated using the Hierarchical Linear Modeling program, version 6.0 (Raudenbush, Bryk, Cheong, & Congdon, 2004). Full information maximum likelihood estimation was used in all modeling applications. An unconditional growth model was first specified to estimate the mean number of words correctly read at the onset of the intervention, the change in the mean number of words between assessment periods, and the amount of student and instructional group variation in each of the growth trajectory components. Equation 1 specifies the three-level unconditional growth model. In Equation 1, it can be seen that a linear regression function was fit to the reading fluency scores obtained on each student over the course of the invention period. More specifically, Y_{tij} is the reading fluency outcome at time t for student i in instructional group j, π_{0ij} is the reading fluency status of student ij at the onset of the intervention (i.e., the first week of summer school), π_{1ij} is the linear change in oral reading fluency across three summer assessment points for student ij, and e_{tij} is a residual term representing unexplained variation from the latent growth trajectory.

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\left(Time_{tij}\right) + e_{tij} \tag{1}$$

In the unconditional model, between-student variation in the initial fluency status (π_{0ij}) and change (π_{1ij}) of struggling readers (Level 2) was modeled in terms of either group mean status or group mean growth and student-level residuals while between-group variation (Level 3) in the initial fluency status and change of instructional groups was modeled as function of grand mean fluency status or grand mean fluency change and group-level residuals.

After estimating the parameters of the unconditional model, unexplained within-person (i.e., intra-individual) variation in reading fluency was conditionally modeled as a function of the person mean centered time-varying weekly attendance covariate while between-person (inter-individual) variation in the fluency status and change outcomes was modeled as a function of individual background characteristics and each of the individual-level dosage indicators. Between-group variation was conditionally modeled by adding terms representing instructor and instructional group characteristics as well as the treatment delivery indicator derived from the small group observations to the equation. The full conditional model is specified in Equations 2 to 7, whereby the inter-individual predictors of student initial status and growth are listed as terms in Equations 2 and 3, the between-group predictors of group mean growth and group differences in the slope relating student engagement to student TORF growth are represented in Equations 4 and 5, the intra-individual predictor of student TORF performance is presented in Equation 6, and the model's error structure is described in Equation 7.

$$Y_{tij} = \gamma_{000} + \gamma_{010}(\text{Female}_{ij}) + \gamma_{020}\left(\text{Ethnic Minority}_{ij}\right) + \gamma_{030}\left(\text{Free Lunch Recipient}_{ij}\right) + \gamma_{040}\left(\text{English Language Learner}_{ij}\right) + \gamma_{050}(\text{Second Grade}_{ij}) + \tag{2}$$

$$\gamma_{100}(\text{Time}_{tij}) + \gamma_{110}(\text{Time}_{tij} \times \text{Female}_{ij}) + \gamma_{120}\left(\text{Time}_{tij} \times \text{Ethnic Minority}_{ij}\right) + \gamma_{130}(\text{Time}_{tij} \times \text{Free Lunch Recipient}_{ij}) + \gamma_{140}(\text{Time}_{tij} \times \text{English Language Learner}_{ij}) + \gamma_{150}\left(\text{Time}_{tij} \times \text{Second Grade}_{ij}\right) + \gamma_{160}\left(\text{Time}_{tij} \times \text{Initial TORF Status}_{ij}\right) + \gamma_{170}\left(\text{Time}_{tij} \times \text{Mean Attendance}_{ij}\right) + \gamma_{180}\left(\text{TIME}_{tij} \times \text{Student Engagement}_{ij}\right) + \gamma_{190}\left(\text{Time}_{tij} \times \text{Homework Completion}_{ij}\right) + \tag{3}$$

$$\gamma_{101}\left(\text{Time}_{tij} \times \text{Treatment Fidelity}_{j}\right) + \gamma_{102}\left(\text{Time}_{tij} \times \text{Classroom Teacher}_{j}\right) + \gamma_{103}\left(\text{Time}_{tij} \times \text{Group Size}_{j}\right) + \gamma_{104}\left(\text{Time}_{tij} \times \text{Mean TORF Status}_{j}\right) + \tag{4}$$

$$\gamma_{181}\left(\textit{Time}_{tij} \times \text{Mean Engagement}_{ij} \times \text{Treatment Fidelity}_{j}\right) + \gamma_{182}\left(\text{Time}_{tij} \times \text{Mean Engagement}_{ij} \times \text{Classroom Teacher}_{j}\right) + \gamma_{183}\left(\text{Time}_{tij} \times \text{Mean Engagement}_{ij} \times \text{Group Size}_{j}\right) + \gamma_{184}\left(\text{Time}_{tij} \times \text{Mean Engagement}_{ij} \times \text{Mean TORF Status}_{j}\right) + \tag{5}$$

$$\gamma_{200}\left(\text{Weekly Attendance}_{tij}\right) + \tag{6}$$

$$r_{0ij} + r_{1ij}(\text{Time}_{tij}) + u_{00j} + u_{10j}(\text{Time}_{tij}) + u_{18j}\left(\text{Time}_{tij} \times \text{Mean Engagement}_{ij}\right) + e_{tij} \tag{7}$$

In the conditional model, the time-varying weekly attendance covariate was specified as a fixed effect at Levels 2 and 3, and the slope relating mean student engagement to reading fluency growth was specified to randomly vary across instructional groups. The engagement/fluency growth slopes were freed to vary in order to examine whether aspects of the treatment context served to moderate the relationship between the engagement of students and TORF growth. As can be seen in Equation 5, variation in the engagement/fluency growth slopes was modeled as a function of the same set of group-level variables that were used to predict the mean growth of treatment groups (i.e., treatment fidelity, instructor status, group size, initial group mean TORF).

Results

Unconditional Summer Reading Fluency Model

Table 1 presents the results of the unconditional model. Unconditional growth model estimates indicated that students were able to read approximately 35 words per minute on average at the start

Table 1. Three-Level Unconditional Summer Reading Fluency Model

Fixed effect	Coefficient	*SE*	*t*
Mean TORF status, γ_{000}	35.36	3.03	11.67*
Mean TORF growth, γ_{100}	4.55	0.55	8.21*
Random effect	**Variance component**	***df***	χ^2
Individual TORF status, r_{0ij}	46.23	97	259.58*
Individual TORF growth, r_{1ij}	4.31	97	148.93*
Level-1 error, e_{tij}	32.34		
Mean TORF Status, u_{00j}	207.04	23	378.13*
Mean TORF growth, u_{10j}	4.43	23	40.89*
Level-1 coefficient	**Percentage of variation between groups**		
Individual TORF status, π_{0ij}	81.7		
Individual TORF growth, π_{1ij}	50.7		

Note. *SE* = standard error.
*$p < .05$.

of the summer literacy intervention ($\gamma_{000} = 35.36$). Students also gained the ability to read an average of 4.5 additional words between each assessment period ($\gamma_{100} = 4.55$). Summer school participants thus gained the ability to read approximately nine additional words across the 5-week intervention period ($4.55 \times 2 = 9.1$). Variance estimates indicated that there were student and instructional group differences in entry status and growth across the treatment period. Calculation of the percentage of variation attributable to instructional groups indicated that 82% of the variability in entry status (i.e., 207.04/(46.24 + 207.04)) and 51% of the variability in oral reading fluency growth (i.e., 4.43/(4.31 + 4.43)) was due to group-to-group differences. All unconditional model estimates were statistically significant ($p < .05$).

Conditional Summer Reading Fluency Model

Table 2 presents the results of the conditional model. The coefficient associated with the time-varying attendance covariate ($\gamma_{200} = 0.46$, $p > .05$) indicates that intra-individual differences in weekly attendance were not associated with unexplained deviations from the estimated summer growth trajectories. In other words, there was no statistical evidence to suggest that students who had lower attendance between each of the assessment periods had simultaneously lower than expected reading fluency performance. With respect to the set of time invariant predictors, inter-individual (within group) coefficients indicated that grade level was associated with both reading fluency outcomes. Relative to their first-grade peers, second-grade students read an average of nine more words per minute at the onset of the intervention period ($\gamma_{050} = 8.99$, $p < .05$). However, the six-word per assessment period growth advantage held by first graders over the course of the intervention ($\gamma_{150} = -6.06$, $p < .05$) reversed the initial grade-level performance difference by the close of summer school. Students with higher TORF scores at the start of summer school also had more reading fluency growth than students with initially lower TORF scores ($\gamma_{160} = 0.39$, $p < .05$). In addition, average attendance ($\gamma_{150} = 0.98$, $p < .05$) and mean engagement ($\gamma_{150} = 1.40$, $p < .05$) during the summer intervention period were uniquely associated with summer reading fluency growth rates. Students who attended more frequently and students who were more engaged during the literacy lessons had higher growth rates than their more absent and less engaged peers. However, homework completion rates were not associated with summer reading growth and demographic characteristics were not associated with either oral reading fluency outcome.

Table 2. Three-Level Conditional Summer Reading Fluency Model

	Coefficient	*SE*	*t*
Mean TORF status, γ_{000}	30.76	3.36	9.14*
Inter-individual predictors			
Female, γ_{010}	2.90	1.65	1.76
Ethnic Minority, γ_{020}	–1.37	2.24	–0.61
Free Lunch Recipient, γ_{030}	–0.46	1.99	–0.23
English Language Learner, γ_{040}	1.22	2.37	0.82
Second Grade, γ_{050}	8.99	4.06	2.21*
Mean TORF Growth, γ_{100}	7.61	1.60	4.75*
Inter-individual predictors			
Female, γ_{110}	–1.17	0.99	–1.19
Ethnic minority, γ_{120}	0.12	1.32	0.09
Free lunch recipient, γ_{130}	0.66	1.16	0.57
English language learner, γ_{140}	–1.59	1.40	–1.14
Second grade, γ_{150}	–6.06	1.81	–3.35*
Initial TORF status, γ_{160}	0.39	0.05	7.63*
Mean attendance, γ_{170}	0.98	0.39	2.50*
Mean engagement, γ_{180}	1.40	0.66	2.13*
Homework completion, γ_{190}	1.50	2.38	0.63
Between-group predictors			
Treatment fidelity, γ_{101}	–0.36	2.91	–0.13
Classroom teacher, γ_{102}	–2.09	1.11	–1.88
Group size, γ_{103}	0.15	0.41	0.36
Mean TORF status, γ_{104}	0.57	0.04	14.52*
Engagement/Reading Growth Slopes, γ_{180}			
Between-group predictors			
Treatment fidelity, γ_{181}	0.94	0.44	2.15*
Classroom teacher, γ_{182}	–2.08	1.48	–1.41
Group size, γ_{183}	–0.25	0.51	–0.48
Mean TORF status, γ_{184}	0.13	0.04	3.43*
Mean Weekly Attendance, γ_{200}	0.46	0.37	1.26

Note. Results based on data from 124 students distributed across 24 instructional groups. *SE* = standard error.
*$p < .05$.

At the group level, the initial mean status of instructional groups was predictive of group mean growth over the summer intervention period. On average, for each unit increase in initial group mean TORF, mean TORF gains were approximately one half of a word-per-minute larger per assessment period ($\gamma_{104} = 0.57, p < .05$). However, contrary to expectation, instructor status, group size, and the fidelity with which instructors delivered the literacy lessons to students were not directly associated with mean growth rates ($p > .05$). Yet, along with group mean TORF ($\gamma_{184} = 0.13, p < .05$), fidelity to the instructional model ($\gamma_{181} = 0.94, p < .05$) was a unique predictor of the variation in the strength of relationship between student engagement and reading fluency growth. In direct instruction groups where the instructor maintained a high fidelity delivery of the instructional protocol, the relationship between student engagement and fluency growth was generally positive and strong. However, as fidelity of implementation decreased, the relationship between student engagement and reading fluency growth became less clear until a decidedly negative relationship between engagement and growth was observed in the lowest fidelity instructional group. These results suggest that engagement was beneficial when the instructor stayed on script, all else equal. Figure 1 displays the manner in which instructional fidelity was observed to moderate the adjusted relationship between

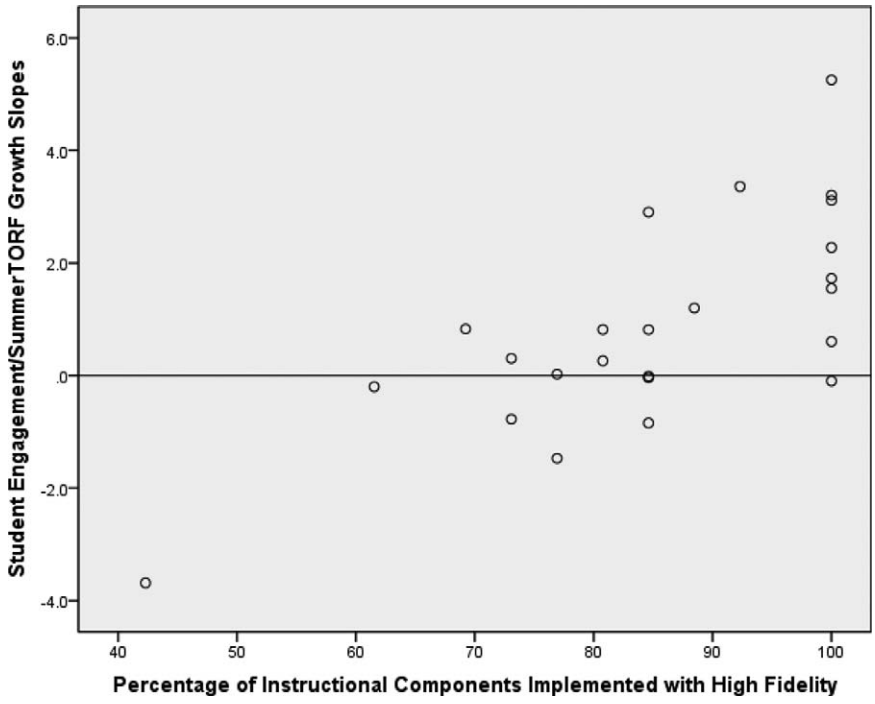

Figure 1. Relationship between student engagement and summer TORF growth as a function of instructional fidelity.

engagement and summer reading fluency growth. No other predictor of the engagement/growth outcomes was found to be statistically significant.

Discussion

The potential for the implementation of a treatment intervention to deviate from the intended design has led evaluators to identify essential program components and develop measures that capture the extent to which providers deliver and recipients receive and adhere to a treatment protocol (e.g., Bellg et al., 2004; Borrelli et al., 2005). Increased focus on tracking provider delivery of and recipient interaction with active program ingredients provides an opportunity to investigate whether deviations from protocol undermine or in some cases enhance recipient outcomes. However, the manner in which various graded deviations from a treatment protocol associate with recipient outcomes may not be always be readily apparent. In the current demonstration, while relationships between individual-level dosage variables were generally as expected (e.g., less receipt of instruction, less reading fluency growth), group-level main effect hypotheses regarding the relationship between instructional fidelity and mean fluency growth rates were not supported. Yet, while the fidelity by which literacy instruction was delivered was not predictive of the mean growth outcomes of groups, instructional fidelity was shown to moderate the relationship between student engagement and the reading fluency growth of individual struggling readers. Together, these results demonstrate that outcome relevant deviations from protocol may occur at a variety of levels and in a variety of ways both within and between levels. As a consequence, a more thorough understanding of program operations is likely to be achieved when evaluators conceptualize, measure, and model "treatment fidelity" as a multilevel, multidimensional construct.

Program Evaluation Applications

The integration of a multidimensional treatment fidelity conceptualization with a multilevel statistical model offers a powerful framework for investigating the manner and degree to which the delivery and receipt of treatment associates with individual outcomes. The present demonstration highlighted three applications germane to evaluation contexts. First, in evaluation settings where individual performance is tracked over time, it may be of interest to investigate whether a dosage variable is predictive of intra-and inter-individual outcome variation. For example, evaluators may want to determine if clients who are more engaged in individual therapy sessions have better than expected outcomes during the course of treatment and overall. Alternatively, an evaluator may want to examine whether weekly deviations from a strict dietary protocol associate with lower than expected weekly weight loss, but taken as an average predict higher mean weight loss overall. Results associated with the time-varying covariate (i.e., weekly summer school attendance) in the current demonstration indicated that students with higher weekly attendance tended to have better than expected reading fluency performance at each assessment point. However, the intra-individual relationship between summer school attendance and reading fluency growth was not strong enough to be statistically significant at a conventional alpha level. On the other hand, the inter-individual relationship between attendance and growth was positive and statistically significant (i.e., higher average attendance, more reading fluency growth), suggesting that conclusions regarding dosage ''effects'' can differ with respect to whether a within- or a between-treatment recipient variable specification is under consideration.

In addition to allowing for time-varying and time invariant individual demographic and treatment receipt covariates, the multilevel modeling framework facilitates estimation of relationships between contextual aspects of the treatment environment, process-related treatment delivery components, and program outcomes. The estimation of models that permit the separation of within (i.e., inter-individual variation) from between unit variation is relevant when recipients are delivered treatment by providers nested within a higher order organizational structure. In the current demonstration, students were delivered instruction in small groups in one school site. In other school-based intervention contexts, the nesting of students within classrooms and classrooms within schools would be common. Alternatively, in a community health context, the nesting of children within families and families within neighborhoods would be normative. In the latter scenario, if a multicomponent intervention where active treatment ingredients were delivered by parents in the home and by providers at local community sites, the delivery of treatment would occur at two levels. Yet, by conceptualizing and measuring implementation components at both the familial and the community site level, it would be possible to estimate the extent to which outcome variation was associated with each treatment modality and ascertain whether variations in program delivery or contextual factors in treatment environment were associated with treatment outcomes. Results associated with the sample of summer school instructors described herein indicated that a greater percentage of outcome variation was associated with instructional groups (82% of the variation in initial literacy status and 51% of the variation in literacy growth). The large amount of variance attributable to instructional groups reflects both the skill-based student grouping strategy employed by the district and provider differences in instructional effectiveness. Contrary to expectation however, predictors of the group mean outcomes were generally not available.

The relative lack of relationships between variations in the context and delivery of instruction and the fluency outcomes of instructional groups could be interpreted to mean that variation among providers was not directly relevant to group outcomes. However, in the multilevel modeling framework, it is also possible to test whether variations in the treatment delivery context indirectly moderate relationships between individual characteristics, including inter-individual differences in treatment receipt and associated treatment outcomes. Estimates from the current model revealed that while

group-level main effect relationships were limited to the association between mean initial status and growth, provider differences in the fidelity of instructional implementation were predictive of the strength of the observed relationship between student engagement and summer reading fluency growth. The cross-level relationship linking greater instructional fidelity with a stronger engagement-growth outcome indicates that the fidelity ''effect'' was transmitted through facilitation of a strengthened relationship between the engagement and learning rate of students in high fidelity instructional contexts. Cross-level fidelity relationships may be noteworthy in a variety of evaluation contexts. For example, heightened integrity to the delivery of a client-centered therapeutic model could be associated with a stronger relationship between client satisfaction and positive clinical outcomes. Conversely, it could be the case that deviations from a scripted treatment model associate with a decoupling of relationships between various demographic characteristics and observed treatment outcomes. In the former scenario, the provision of training to encourage fidelity to the program model would be a logical recommendation to convey to stakeholders while in the latter, further exploration of the nature of program deviations would be an area for inquiry.

Study Limitations

The examination of recipient outcomes as a function of variations in program delivery, receipt, and adherence within the context of supplemental literacy intervention provides a demonstration of the utility of conceptualizing treatment fidelity as a multidimensional construct. However, in order to better contextualize the findings reported above, it is necessary to discuss associated design limitations. First, the demonstration used intervention data obtained in conjunction with the localized delivery of supplemental summer instruction. Specific findings regarding the within- and between-level relationships between the components of fidelity and student learning outcomes observed in moderately sized Pacific Northwest school district may not be generalizable to dissimilar school sites, instructional settings, or other social or health science treatment contexts. Second, reported results do not directly speak to the relative efficacy of the summer literacy intervention as a comparison of outcomes with a similarly measured and equivalent group of nontreated struggling readers was not available. Moreover, the causal effect of a higher fidelity delivery (or receipt) of program protocol was not estimable as instructors and students self-selected into the graded implementation and participation levels currently observed. For stronger inferences regarding the general effect of providing supplemental summer instruction and/or the effect of providing a higher or lower instructional dosage, designs that allow for the random assignment of students and instructors to conditions (when ethical and feasible) and the estimation statistical models that account for noncompliance with treatment assignment (Jo, 2002; Jo & Muthén, 2001) are recommended.

Another aspect of the demonstration that merits consideration is the absence of large differences in the training and skill level of literacy instructors, the size of instructional groups, and ultimately the delivery of the instructional protocol. In many respects, the high quality and relative uniform delivery of instruction and associated consistency in observational ratings is testament to the organizational effectiveness and efficiency of the school district under study. However, limited variability in the delivery of instruction may have had the consequence of preventing the detection of additional relationships between measured fidelity components and student learning outcomes (see Schulte et al., 2009). A related concern stems from the lack of reliability data on the student engagement ratings. As reliability checks were only conducted on the observations of instructional delivery, the scores that were derived from instructor ratings of students' engagement during the daily lessons contained an unknown amount of error. Further, the demonstration was based on one measure of early childhood literacy that was administered multiple times over the 5-week intervention period. Use of an alternative literacy measure more or less sensitive to practice effects and the instructional

treatment may have resulted in different estimates of the relationship between the components of fidelity and summer learning outcomes.

Finally, it should also be acknowledged that while a multivariate treatment fidelity representation was used to demonstrate some of the analytic possibilities that are associated with the multilevel modeling framework, each of the individual- and group-level measures were unidimensional indicators of select intervention components. Although specifically designed to capture aspects of the intervention hypothesized conducive to the promotion of successful learning outcomes, it is possible that with more robust conceptualization and measurement of program components, multidimensional provider and recipient-level fidelity indicators could be identified and modeled. For example, in addition to documenting the extent to which intervention components were delivered, instructional quality could be tracked by trained observers. At the individual level, accuracy ratings could be conducted in conjunction with engagement ratings to also measure the extent to which students comprehend and correctly employ practices conducive to the acquisition of literacy skills during small group instruction. With multiple indicators representing different facets of a larger construct, it would be possible to specify and test a hypothesized measurement model using latent variable modeling techniques. The use of multilevel latent variable models to first represent the interrelations among fidelity indicators and then test within- and between-level linkages with conceptually relevant predictor and criterion variables is promising empirical direction for those conducting evaluations and/or pursuing the development and validation of treatment fidelity conceptual frameworks.

Issues to Consider in the Application of Multilevel Models

Multilevel models provide distinct advantages to evaluators seeking to estimate relationships and test hypothesis regarding within and between level linkages between fidelity indicators and program outcomes. However, the proper application of the multilevel model, as with any statistical procedure, is dependent upon the resolution of a range of methodological and statistical issues (see Dedrick et al., 2009). Key issues are those that have an analog in the estimation of traditional single-level models (e.g., model specification, missing data, distributional assumptions), as well as those that are more particular to and have wide-ranging implications in a multilevel modeling context (e.g., covariance structure, variable centering, estimation method; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). For evaluators working in settings that have a limited number of observations, participants, and/or sites, statistical power issues may also present a significant impediment to the application of the multilevel model. In nested designs where a treatment or program is delivered at the group or cluster level, statistical power at a set alpha level depends on the size of the group or cluster, the number of groups or clusters, the intraclass correlation (ICC) between grouping/clustering and the outcome measure, and size of the effect or relationship to be detected. In multilevel repeated measures designs, power is also affected by the amount of variance within and between persons as well as the number of measurement occasions. On average, statistical power is higher when cluster number and cluster size is large, the ICC is low, the detectable effect size is large, and in repeated measures designs, as the number of observations per individual increase (Raudenbush & Xiaofeng, 2000, 2001).

When recommending how to allocate resources in an evaluation design, evaluators should be aware that an increase in the number of groups/clusters has been shown to return a larger gain in statistical power than an increase in the number of individuals per unit especially as the ICC grows large. However, the practicality and cost associated with increasing the number of groups may be prohibitive in many evaluation contexts. Alternative strategies like measuring and specifying strong within- and between-group model covariates, increasing the number of individual observations in repeated measures designs, and where possible, encouraging the high fidelity delivery of stronger treatments may be required to enhance statistical power. For example, in an efficacy or effectiveness

trial, a higher fidelity implementation of program components can serve to maximize the differentiation of conditions, reduce within condition/cluster variance, and thereby increase the power to detect treatment effects. Yet, when a high fidelity treatment implementation is realized, it becomes more difficult to detect relationships between fidelity indicators and program outcomes. In general, a relative lack of variability in fidelity components may be desirable from a statistical power/causal inference vantage point in a summative evaluation of program effects, but may undermine the test of hypotheses regarding within and between level fidelity-outcome relationships in formative evaluations of program process.

Conclusion

As the conceptualization and measurement of the treatment fidelity construct has become more sophisticated, a coincident increase in the manner in which fidelity data can be used by evaluators and stakeholders for formative and summative program evaluation follows. However, the extent to which individual- and group-level dosage–response relationships can serve to identify and highlight the most critical treatment delivery components, focus provider training, and inform program redesign efforts has not been completely realized. In light of the current divide between the collection and empirical modeling of fidelity-related data, the purpose of the present article was to offer a demonstration of an analytic approach that optimizes the estimation of relationships between an outcome variable and variables representing a multilevel, multivariate treatment fidelity conceptualization. Results associated with the demonstration revealed the presence of a range of dosage–response relationships within and between levels of the data hierarchy. Although particular to a school-based instructional intervention, the basic analytic approach described herein can easily be adapted to fit a variety of organizational contexts where multiple clinicians and/or social service providers deliver a program of services to clients or participants. As a result, evaluators should generally be prepared to recommend evaluation designs that merge strong conceptualization and adequate measurement of key individual and provider level fidelity components with analytic models that enable thorough testing of the hypothesized multivariate, multilevel relationships thought to facilitate or undermine program operations. If supported by program stakeholders and sponsors, evaluators will then be better able to identify the conditions under which treatment outcomes are enhanced or degraded with respect to specific deviations from the program model and thereby offer empirically based recommendations to support ongoing program development efforts.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090369 to the University of Oregon. The opinions expressed are my own and do not represent the views of the Institute of Education Sciences or the U.S. Department of Education.

Declaration of Conflicting Interests

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

Funding

The author 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 R305A090369 to the University of Oregon.

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, *23*, 171–191.

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., … Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations for the NIH Behavior Change Consortium. *Health Psychology*, *23*, 443–451.

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *Educational Forum*, *40*, 345–370.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, *22*, 551–575.

Borman, G. D., & Dowling, N. M. (2006). The longitudinal achievement effects of multiyear summer school: Evidence from the Teach Baltimore randomized field trial. *Educational Evaluation and Policy Analysis*, *28*, 25–48.

Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Breger, R., … Orwig, D. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology*, *73*, 852–860.

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.

Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Children's Educational Services. (1987). *Test of Oral Reading Fluency (TORF)*. Eden Prairie, MN: Author.

Cooper, H., Charlton, K., Valentine, J. C., & Muhlenbruck, L. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monographs of the Society for Research in Child Development, 65*, Serial No. 260.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, *66*, 227–268.

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.

Dedrick, R. F., Ferron, J. M, Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., … Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69–102.

Donaldson, S. I. (2007). *Program Theory-Driven Evaluation Science: Strategies and Applications*. Mahwah, NJ: Lawrence Erlbaum.

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

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.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, schools, and inequality*. Boulder, CO: Westview Press.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction. *Review of Educational Research*, *47*, 335–397.

Greenwood, C. R. (2009). Treatment integrity: Revisiting some big ideas. *School Psychology Review 38*, 547–553.

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

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2008). Treatment integrity in behavioral consultation: Measurement, promotion, and outcomes. *International Journal of Behavioral Consultation and Therapy*, *4*, 95–114.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series, *School Psychology Review*, *38*, 445–459.

Hall, G. E., & Loucks, S. F. (1977). A developmental model for determining whether the treatment is actually implemented. *American Educational Research Journal*, *14*, 263–276.

Hall, G. E., Loucks, S. F., Rutherford, W. L., & Newlove, B. W. (1975). Levels of use of the innovation: A framework for analyzing innovation adoption. *Journal of Teacher Education*, *26*, 52–56.

Heyns, B. (1978). *Summer learning and the effects of schooling*. New York, NY: Academic Press.

Heyns, B. (1987). Schooling and cognitive development: Is there a season for learning*? Child Development*, *58*, 1151–1160.

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.

Institute for Education Sciences. (2009). *Request for Applications: Education Research Grants*. Retrieved from http://ies.ed.gov/funding/pdf/2009–84305A.pdf

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, *27*, 385–409.

Jo, B., & Muthén, B. O. (2001). Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 57–87). Mahwah, NJ: Lawrence Erlbaum.

Jones, H. A., Clark, A. T., & Power, T. J. (2008). Expanding the concept of intervention integrity: A multidimensional model of participant engagement. *Balance*, *23*, 4–5.

McGrew, J. H., Bond, G. R., Dietzen, L., & Salyers, M. (1994). Measuring the fidelity of implementation of a mental health program model. *Journal of Consulting and Clinical Psychology*, *62*, 670–678.

McKenna, S. A., Rosenfield, S., & Gravois, T. A. (2009). Evaluating the validity of instructional consultation: Level of Implementation Scale—Revised. *School Psychology Review*, *38*, 496–509.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*, 247–266.

Mowbray, C. T., Bybee, D., Holter, M., & Lewandowski, L. (2006). Validation of a fidelity rating instrument for consumer-operated services. *American Journal of Evaluation*, *27*, 9–27.

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.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Pub. No. 00-4754). Rockville, MD: National Institutes of Health.

National Research Council Committee for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials, Mathematical Sciences Education Board, Center for Education, Division of Behavioral and Social Sciences and Education. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academies Press.

Noell, G. H. (2008). Research examining the relationships among consultation process, treatment integrity, and outcomes. In W. P. Erchul, & S. M. Sheridan (Eds.), *Handbook of research in school consultation: Empirical foundations for the field* (pp. 315–334). Mahwah, NJ: Lawrence Erlbaum.

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.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis*, *15*, 477–492.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Raudenbush, S. W., & Xiaofeng, L. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199–213.

Raudenbush, S. W., & Xiaofeng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387–401.

Scheirer, M. A. (1987). Program theory and program implementation: Implications for evaluators. In L. Bickman (Ed.), *Using program theory in evaluation* (New Directions for Program Evaluation, No. 33, pp. 59–76). San Francisco, CA: Jossey-Bass.

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

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S. G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), *Evaluation studies review annual* (Vol. 4, pp. 15–35). Thousand Oaks, CA: Sage.

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

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Rep. 109). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, *49*(2), 156–167.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation*, *30*, 44–61.

Zvoch, K., Letourneau, L. E., & Parker, R. P. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation*, *28*, 132–150.

Zvoch, K., & Stevens, J. J. (2011). Summer school and summer learning: An examination of the short and longer term changes in student literacy. *Early Education and Development*, *22*, 649–675.