The Canadian Journal of Program Evaluation Vol. 24 No. 1 Pages 133–156
ISSN 0834-1516

DESIGNING AND APPLYING PROJECT FIDELITY ASSESSMENT FOR A TEACHER IMPLEMENTED MIDDLE SCHOOL INSTRUCTIONAL IMPROVEMENT PILOT INTERVENTION

Gary J. Skolits
Jennifer Richards
University of Tennessee
Knoxville, Tennessee

Abstract: This article argues that intervention pilot test evaluations have focused insufficient attention on the measurement of project fidelity and the subsequent use of fidelity results for (a) interpreting variations in project outcomes and (b) understanding the rationale for teachers' deviations from implementation protocols. The authors report on the establishment and application of an evaluation methodology for measuring and analyzing implementation fidelity for a middle school instructional improvement pilot project. The authors found that the highest implementation fidelity scores were not correlated with the most desirable project outcomes, as lower fidelity scores—in the 70–79% range—produced the most favourable gains on pre-post student outcomes. Moreover, application of the fidelity evaluation methodology provided insight into teacher deviation from implementation protocols; such deviation from the implementation protocols typically reflected meaningful professional classroom judgements.

Résumé : Cet article soutient que les évaluations de l'intervention des projets pilotes ont suscité trop peu d'attention dans la mesure de fidélité des projets et de l'utilisation des résultats de ces mesures pour (a) interpréter des variations dans les résultats du projet et (b) comprendre le rationnel des enseignants pour s'écarter des protocoles d'implantation. Les auteurs décrivent le développement et l'application d'une méthode de mesure et d'analyse de la fidélité de l'implantation d'un projet pilote qui vise l'amélioration de l'enseignement à l'école intermédiaire. Les auteurs ont trouvé que les scores de fidélité les plus élevés n'étaient pas corrélés aux résultats escomptés du projet, alors que les plus faibles, entre 70 % et 79 %, engendrent les gains les plus avantageux

Corresponding author: Gary Skolits, University of Tennessee, Educational Psychology and Counseling Department, A 503 Jane and David Bailey Education Complex, Knoxville, TN, USA 37996-3456; gskolits@utk.edu

sur les scores pré-post des étudiants. De plus, l'application de la méthodologie permet de mieux comprendre le rationnel des enseignants pour s'écarter du protocole d'implantation; ces écarts reflètent typiquement des jugements sensés des professionnels dans la salle de classe.

Educational problems and needs in schools generate systematic efforts by educational researchers and curriculum innovators to design, develop, and implement potential interventions capable of making desired improvements. The process used to assess the effectiveness of a new intervention is often a pilot test application of a proposed educational improvement under carefully specified conditions that are managed and controlled to the extent possible in a real-world school setting. The logic underlying a pilot test of an intervention is straightforward. An intervention model is designed to address a specific educational problem based upon the current knowledge base within the educational research literature. When the innovative intervention is implemented in a methodologically sound pilot test application, researchers assess subsequent changes in the problem condition of interest such as classroom student outcomes. If the condition improves upon implementation of the intervention, the model is viewed as an effective solution to the problem. If the condition fails to improve, the intervention model is labeled as ineffective.

While the logic of an intervention pilot is straightforward, the evaluation of an educational intervention rests on two challenging conditions that must be met for an intervention model to receive a credible pilot test. The first condition is internal validity—the actual implementation of an intervention must enable researchers to attribute change in outcomes to the pilot test intervention. The second condition is fidelity, which requires that the implementation of the pilot test application occurs in a manner that is consistent with the intervention design. In other words, the intervention must have a high level of consistency with the prescribed design being tested. Researchers have understandably focused substantial effort on the first condition, since intervention pilot tests that do not provide reasonable evidence of a linkage between outcomes and an intervention lack internal validity.

In contrast, less effort has typically been focused on the implementation fidelity condition (O'Donnell, 2008). In part, this lack of attention to fidelity is somewhat understandable. Fidelity assessment is always a challenge due to the inability of researchers to control school settings during implementation. Intervention designers and evaluators must think in terms of the degree of implementation fidelity since

absolute fidelity is virtually impossible in a changing and challenging real-world school setting. When an intervention design meets the contextual reality of the school, pressures for modifications to the prescribed intervention design can be expected to occur from multiple sources. For example, teachers can be expected to make some level of adjustments while implementing an intervention model to suit particular instructional needs of their classroom, or they may deem it necessary to make changes in response to constantly changing contextual situations (Leventhal & Friedman, 2004).

Unfortunately, failure to effectively monitor, measure, and account for adjustments made to an intervention during implementation diminishes researchers' ability to determine which specific variation of the intervention has actually been pilot tested as well as to determine how variation in implementation fidelity related to project outcomes (Bellg et al., 2004). Moreover, the lack of fidelity assessment interferes with the potential for replication of the research and diminishes the ability of researchers to make valid comparisons across multiple interventions (Hester, Baltodano, Gable, Tonelson, & Hendrickson, 2003). Of equal concern, when researchers do not carefully study how teachers modify interventions during implementation they miss an opportunity to understand how future interventions could be better designed to be more responsive to teachers' needs.

CHALLENGES IN ASSESSING IMPLEMENTATION FIDELITY FOR TEACHER-LED INITIATIVES

The fidelity assessment challenge is especially problematic for interventions promoting improved student academic outcomes that utilize classroom teachers to implement an intervention design. In these instances, developers typically provide professional development training to teachers in an effort to improve teacher content knowledge, pedagogical skill, and other aspects of instructional practice based upon various instructional improvement strategies reflected in the literature. At the conclusion of professional development activities, researchers often measure teachers' satisfaction with the training experience and may also assess gains in teacher content knowledge and/or pedagogical skills resulting from the professional development experience. In fewer cases, researchers seek to assess the extent of subsequent classroom implementation through teacher self-reports such as surveys or interviews. However, unless the project provides resources for the direct observation and monitoring of teachers' subsequent implementation of their newly acquired content knowledge

and/or pedagogical skills, there is no confirmable evidence of faithful implementation of the intervention model. When in-class assessment of implementation does not occur, it is often not possible to attribute project success or failure to the model intervention strategy since the prescribed model may not have been faithfully implemented or an unspecified and unknown variant of the intervention design may actually have been implemented in the classroom. Moreover, an opportunity is lost to better understand outcomes in light of actual implementation efforts.

If implementation fidelity is to be monitored and assessed, researchers must be able to directly and systematically observe teacher implementation efforts. While teacher self-reports can be used to roughly gauge fidelity, self-reports provide insufficient evidence of fidelity because they lack the critical objectivity that an independent observer can provide. However, observing teacher implementation efforts presents researchers with two immediate challenges. First, project designers need the time and staff resources to conduct subsequent classroom observations; such observations can be highly time intensive for improvement initiatives that are implemented across several days of instruction by even a small number of teachers. Second, researchers need the time, skill, and staff resources and support to plan for the development, testing, and use of credible classroom observation instruments and measurement protocols. Unfortunately, the time and resource challenge inherent in the assessment of implementation fidelity is particularly problematic for many public grant-funded initiatives.

While these time and staff resource issues can be daunting, they are not technically challenging. The major challenge of fidelity assessment (i.e., the development of a robust fidelity assessment methodology) presents a much more complicated array of technical obstacles for the researcher (Mowbray, Holter, Teague, & Bybe, 2003). Regardless of the intervention design, observation instruments and protocols need to be appropriately designed and subjected to the same rigorous assessment of reliability and validity as other measurement instruments. The validity of the measuring instruments needs to be assessed, even though the instrument may be nothing more than a well-crafted and systematically focused checklist of expected teacher behaviour mapped to the prescribed instructional protocol. Inter-rater reliability should also be addressed, if required, to create a common and reliable fidelity metric that is applicable to the overall data analysis of project outcomes. The technical aspects of measuring fidelity

is especially challenging for educational improvement projects that seek to advance more sophisticated intervention designs based on recent scientific-based research literature. For example, innovative intervention designs may be introducing several new pedagogical strategies that embed multiple learning activities that are integrated across math, science, language arts, and social studies content areas. Intervention models may also be based on a flexible, adaptive design, allowing teachers some range of specified implementation modifications based on changing student needs and other aspects of the learning content.

The purpose of this article is to report on a practical approach to assessing implementation fidelity and demonstrate how subsequent fidelity results derived from the assessment provided a deeper understanding of the project implementation and outcomes. These purposes are addressed within the context of a pilot test of a model curriculum designed to improve student outcomes in math, science, social studies, and language arts at five middle schools within two southeastern states in the United States. Three research questions guided this study: (a) How can project fidelity of an intervention be effectively assessed? (b) To what extent is there a relationship between ultimate project outcomes (student achievement) and variation in implementation fidelity? and (c) In what ways, and for what reasons, do teachers (implementers of the intervention) adjust or deviate from the prescribed project implementation? The findings from these questions will be used to suggest subsequent implications for research and practice.

REVIEW OF LITERATURE

Fidelity of implementation has been conceptualized as the degree to which a program model is implemented as intended by the program designers as well as the degree of consistency across multiple implementations (Dusenbury, Branningan, Falco, & Hanson, 2003; Fullan, 2001; Smith, Dunic, & Taylor, 2007). In reference to the effectiveness of efforts to train teachers to implement a specific educational intervention design, the educational research literature has long recognized the need to assess how closely aligned teacher implementation of the intervention strategy is with the original intervention design (Berman & McLaughlin, 1976). Accurately measuring fidelity is foundational for establishing the validity of program effectiveness research and critical for determining whether or not a pilot test represents a true measure of a program's design in real-world conditions

(Dumas, Lynch, Laughlin, Smith, & Prinz, 2001; Mowbray et al., 2003; Sanchez et al., 2007). The literature also reminds researchers that "failure to establish validity can severely limit the conclusions that can be drawn from any outcome evaluation" (Dumas et al., 2001, p. 39).

Understandably, project designers assume that high degrees of implementation fidelity are desirable, while lower levels of fidelity are correspondingly assumed to reduce the effectiveness of an intervention design (Chen, 2005). However, Leventhal and Friedman (2004) asserted that demands for pilot study participants to maintain extremely high levels of fidelity to an intervention design may actually tend to limit an intervention's effectiveness. In such instances, it is possible that the more attention is focused on following the letter of an intervention protocol, the more the overall spirit of the design underlying the model may be compromised. There is acknowledged pressure for a prescribed curriculum design to be modified and adapted by teachers when it is brought into the real-world setting of a classroom (Fullan, 2001; Shulman, 1990), and, as such, expectations for full adoption of curriculum designs are unrealistic (Rogers, 2003).

Curriculum designers and researchers should expect modification to a prescribed curriculum design, especially when initial efficacy studies (i.e., controlled testing of the internal validity of project design) proceed to effectiveness studies (more realistic field applications assessing the external validity of a project design). Chen (2005) argued that some modification from the original intents of program designers may be necessary to promote long-term change because program designs have to fit into changing and unpredictable real-world applications. However, while some degree of lower fidelity may be more appropriate and desirable, deviating too far from the original design is likely to dilute the program's effectiveness (Chen, 2005). Smith et al. (2007) suggested that while measurement of treatment fidelity was universally considered to be important, no specific threshold of fidelity has been identified in the current literature.

Despite the growing calls for greater scientific rigour in educational research that would reinforce the importance of the empirical assessment of fidelity and the use of fidelity assessment to better understand project outcomes (U.S. Department of Education, 2003), very few empirical studies of fidelity appear in the literature (see the review prepared by O'Donnell, 2008). A large proportion of the few empirical fidelity studies appearing in the literature reported a

positive relationship between fidelity and expected outcomes (Fuchs, Fuchs, & Karns, 2001; Penual & Means, 2004; Ysseldyke et al., 2003), but few studies reported a minimal or ideal fidelity range, especially from the perspective of the relationship between fidelity and project outcomes (Vaughn et al., 2006). O'Donnell (2008) concluded a detailed review of fidelity with the recommendation that "fidelity to critical components and processes should be captured quantitatively so that levels of fidelity can be related to outcomes" (p. 52).

The Treatment Fidelity Workgroup of the National Institutes of Health Behavior Change Consortium (BCC) recommended the consideration of five essential conditions when planning for assessment of fidelity in applied research settings (Bellg et al., 2004):

- Considering treatment fidelity from the beginning stages of the research design to help researchers anticipate and plan for the real-world contexts and possible roadblocks that may accompany any applied research project.
- Ensuring that the training of pilot test participants is standardized so that delivery of the intervention will be consistent across participants (Bellg et al., 2004; Hennessey & Rumrill, 2003).
- Monitoring the implementation of the intervention to determine the fidelity of treatment. Direct observation of the implementation is the monitoring method considered necessary to establish the strongest case for internal validity. However, the research literature also suggests that this can be costly (Smith et al., 2007).
- Monitoring treatment receipt and participants' acquisition of new knowledge and skills is also necessary. This is usually accomplished through some sort of pre/post assessment or student work analysis (Bellg et al., 2004; Smith et al., 2007).
- Collecting evidence, either through pre/post surveys or anecdotal evidence, that participants have applied their new knowledge and skills in their daily lives, where appropriate.

Fidelity analysis assumptions and operational considerations have also been suggested by Mowbray et al. (2003), namely the need to (a) focus on process and/or structural perspectives of fidelity, (b) establish a viable fidelity measurement strategy, and (c) assess the reliability/validity of the resultant measures. While the literature consistently reports on the need for assessing fidelity, little guidance is offered beyond Mowbray et al. and the Treatment Fidelity Work Group on how

to assess fidelity (i.e., methods), especially from the perspective of how to interpret fidelity assessment results to better understand project results. This study incorporated elements of both of these resources.

METHODS

Study Context

The present study was embedded in a larger educational intervention research project funded by the United State's Department of Agricultures (USDA) National Integrated Food Safety Initiative (award number TEN2005-02098). The *Food Safety in the Classroom* project evaluated the effectiveness of a newly designed, research-based food safety curriculum for seventh-grade students in five middle schools in the United States located in two southeastern states. The curriculum was an interdisciplinary food safety unit correlated with state standards (recommended content learning expectations/objectives for students at each grade level) for math, science, social studies, and language arts in the two targeted states involved in the initial pilot testing. The instructional unit of the model curriculum was designed to take approximately 6–8 class periods of instruction across all four core subject areas simultaneously. Richards, Skolits, Burney, Pedigo, & Draughon (2008) reported on the project design and overall favourable outcomes of the pilot test of this educational intervention, and this current research effort builds upon this project knowledge base by seeking to better understand the effects of project fidelity on reported project outcomes.

Study Methodology and Participants

This study uses a correlational design that depicts the relationship between implementation fidelity and student outcomes. Although correlation does not imply causation, several contextual factors tend to rule out potential rival explanations of student outcome results related to the pilot project. The project logic model provided a theoretical basis for linking the intervention to student outcomes. Moreover, participating project classroom students would have had few sources of food safety content knowledge beyond their classroom experience given the complete lack of food safety coverage in the school curriculum of the participating schools. Finally, the brief duration of time between the student pretest, teacher implementation of the instructional model, and the student posttest (i.e., approximately two

weeks) limited the opportunity for other exogenous variables such as history and maturation to greatly influence student outcomes. Taken as a group, these factors suggest the strong likelihood that the primary source of food safety knowledge on the part of project students resulted from the classroom unit sponsored by the project.

The participants in this study included 24 teachers and their 233 students in five middle schools, with one of the schools having two separate implementations to different groups of students by different teachers. Two schools in larger, growing communities near a medium-sized city, two schools in sparsely populated communities, and one school located in a medium-sized city were represented in the study. Three of the five schools were performing at or above state standards (learning expectations for students across all grade levels in the school) in math and four of five were performing at or above the state standards in reading. Three of the schools had greater than 50% of students classified as "economically disadvantaged." Participating teachers at each school were organized into teaching teams for the pilot test. These teams taught all four core content areas to a common group of students. Of the five participating schools, one school had a team of two teachers, one who taught language arts and one who taught math, science, and social studies. Another school had a team of five teachers (math, science, social studies, language arts, and reading). In this school the language arts component was co-taught by the reading and language arts teachers. The other schools had teams of four teachers each, with each teacher teaching one content area.

Data Collection Instruments

Researchers developed two instruments for this study: (a) an observational protocol used to guide classroom observations and measure the degree of implementation or fidelity and (b) a student outcomes assessment instrument. The observation protocol instrument reflected the instructional scope, sequence, and learning activities addressed within the model food safety curriculum that integrates food safety content throughout four core subject areas (mathematics, science, social studies, and language arts). The observation instrument is mapped to the instructional protocol, specifying the teacher's responsibility and the students' role in each lesson activity, the prescribed instructional materials and resources, and the timeframe. Raters check off (e.g., yes/no) the occurrence of the instructional activity from the perspective of the prescribed activity elements, its sequencing, and its associated time frame. Raters also comment on the nature of

each of the activity deviations. An external middle school assessment consultant examined the validity of the observation instrument for congruence with the implementation standards and recommended modifications. A revised observation instrument was then field tested by two trained raters who observed and scored a mock instructional event with the protocol instrument. Initial rater agreement exceeded 85%, but was less than the goal of 90% set by project staff. Raters discussed differences, resolved them, and then conducted a subsequent joint observation, resulting in agreement that exceeded 95%. The second instrument, measuring student outcomes, was developed under rigorous test development protocols. These protocols included the creation of a test blueprint tied to the curriculum objectives, expert review of test content, a field test of the instrument with an item analysis, and a subsequent second field test of the revised instrument (Richards et al., 2008). Instrument reliability was determined through a test/re-test design with an intact group of seventh-grade students who were similar to students who would be participating in the project. There was no significant difference ($p = .101$) between the means of the two assessment administrations, suggesting a high degree of test reliability. Internal instrument reliability was also high (Cronbach's alpha = .874).

Procedures

This study was conducted in a manner consistent with the five conditions described by the Treatment Fidelity Workgroup as essential for a fidelity study (Bellg et al., 2004). Each of the teacher study participants were trained to use *Food Safety in the Classroom* in one of five highly standardized training sessions conducting during May, June, and August 2006. Training sessions included a combination of modelling lessons in the curriculum, hands-on participation in activities, and a seminar-style discussion (Richards et al., 2008). This method of delivery allowed teachers an opportunity to apply, analyze, synthesize, and evaluate the new content knowledge as well as novel instructional strategies (Galbo, 1998). A detailed training agenda was used for all training sessions as a means of maintaining consistency between sessions. Teachers were taught how to implement the curriculum and the need to be consistent across all implementation efforts.

At each training session, teachers completed a teacher pretest and posttest assessment designed to measure changes in their content knowledge, pedagogical knowledge, and attitudes toward food safety.

In addition to evaluating the effectiveness of the training sessions, the pretest-posttest also helped assess consistency among the five training sessions. A one-way ANOVA comparing the pretests and posttest gains across all training sessions found no significant difference in the overall changes in participants' knowledge, attitudes, and behaviours between the various training sessions ($p = 0.626$), suggesting a high degree of training consistency from the perspective of participant outcomes.

At the conclusion of each training session, each school team of teachers set a date to implement the curriculum in their classrooms during the 2006–2007 school year. One week prior to implementation, each of the participating teachers' students participated in a pretest. Throughout the curriculum implementation, two researchers were on site to observe teachers' implementation of the unit and assess teacher implementation fidelity with the observational protocol. At the conclusion of the unit, students were administered a posttest that was identical to the pretest. This study is based on the data from the pretests/posttests, as well as the classroom observations of teacher implementation fidelity.

Fidelity Metric Used in This Study

Several intervention considerations led to the identification of the need for the researchers to address the issue of fidelity. The model curriculum was purposely designed to provide teachers with a limited range of flexibility in implementation to address the specific learning styles and other unanticipated needs of their students. Given the number of instructional activities within the instructional unit, the autonomy of 24 teachers in five schools across two states, and other contextual situations that might lead to deviations from the model curriculum design during implementation, project staff recognized the high potential for some level of variation in fidelity to occur across all instructional activities. However, fidelity was expected to occur within a fairly narrow range, especially since the highly standardized training of the participants addressed the need for consistency with implementation protocols. Overall, researchers expected fidelity of implementation to be around 90%.

A review of the available research literature did not result in a ready-to-use fidelity metric or much assistance in suggesting how fidelity should be measured; therefore, the research team had to design one *de novo*. This fidelity assessment methodology and associated metric

had to be designed to track the consistency of implementation effort to the model and determine if implementation changes were consistent with the underlying design elements of the model. The fidelity assessment also had to be capable of helping determine if there were trends in the teacher classroom implementation changes that suggested that some instructional activities needed revision. Finally, the fidelity methodology and assessment had to be capable of determining if there was a relationship between implementation fidelity and student outcomes on the posttest exam. Underlying these considerations was an operational definition of fidelity consistent with those suggested by Mowbray et al. (2003) that are addressed in the following paragraphs, namely (a) focusing on process and/or structural fidelity elements, (b) establishing a viable measurement strategy, and (c) assessing the reliability/validity of the resultant measures.

Because the researchers were focused on process fidelity aspects, they initially intended to rank the fidelity of each model instructional activity implementation on a 1–10 scale and then sum the scores across all instructional activities within each component (i.e., math, science, etc.). Given the 50 instructional activities within the intervention, initial field-testing of this first fidelity design suggested that this approach was too cumbersome and complicated for raters, and it provided little additional information of value over a simpler method using a binary decision (yes/no) regarding a fidelity threshold for each instructional activity. In the binary system, observers determined if the teacher followed the overall lesson plan parameters or did not. Observers noted all material implementation variations through written comments.

As part of the measurement strategy, classroom observers used the observation protocol to assess the implementation fidelity for each lesson in the curriculum unit and noted how lessons and activities were introduced, conducted, and concluded. All teacher modifications of activities were recorded, along with any significant teacher or student comments regarding the implementation. A simple scoring system was used. Each instructional activity within individual content components was allotted a maximum of 5 possible points. A score of 0 points was awarded only if an entire activity was omitted. From 1 to 5 points were awarded based on the ratio of “yes” to “no” checks given by the observers to all the instructional elements in each component. Comments recorded by the observer were also taken into account. In some situations the “letter” of the activity was followed (i.e., teacher followed all prescribed steps) but the “spirit” was not

(i.e., teacher did not facilitate discussion of critical or higher-order thinking questions). Points were deducted from activities where this was the case. Points were also deducted if the activity was performed out of sequence. The total points awarded through this process were divided by the total possible points to produce a percentage score for each component. For example, in the science component, there were a total of 50 possible points (10 instructional activities with a maximum of 5 possible points per activity). Therefore, the total points awarded based on the observation instrument were divided by 50 to produce a percentage score. This percentage score represented the teacher's overall fidelity score.

To help ensure measurement consistency, upon completion of each observation event the two observers jointly discussed their independent observations and resolved any differences in scoring (i.e., agreement on the yes/no decision as to whether the model instructional plan was followed during an instructional activity). Initial agreement between observers was high, correlated at 88% across all observations. For each score deviation, the observers discussed their rationale for awarding or deducting points and reached agreement on a consensus score. Fortunately, rating differences were few and the rating rationale for all points awarded were reflected in observational comments, and all observational comments were retained for subsequent analysis.

RESULTS AND DISCUSSION

Fidelity Findings

As reflected in Table 1, fidelity ranges for each of the four content area implementations in the participating schools ranged from 38.0 to 100 (out of a maximum score of 100). The mean fidelity score was 76.7 (SD = 16.17). None of the participating schools achieved consistently high or consistently low fidelity scores across the implementation of language arts, science, math, and social studies, suggesting the potential lack of overriding school effects. Moreover, the one school with two implementations achieved two different fidelity ratings (one in the 70s and one in the 80s). Regardless of the content implementation fidelity score values achieved, student pre/post gains were achieved for all content areas, and only one school was unable to produce statistically significant effect sizes on the pre/post content assessment gain scores (School 1 in language arts and social studies students, with corresponding fidelity scores of 54.4 and 91.1). However, this

particular school also had the lowest number of students (N = 10), limiting statistical power related to effect size calculations.

Table 1
Content Area Fidelity, Pre-Post Gains and Effect Sizes

School	Content area	Fidelity score	*N* students	Pre-post gain content test	Student outcomes effect size[a]
1	Science	80.0	10	13.85 (±18.05)	0.4283*
	Language Arts	54.4	10	15.5 (±35.03)	0.2960
	Math	100.0	10	20.0 (±16.32)	0.4325*
	Social Studies	91.1	10	12.7 (±33.19)	0.2242
2	Science	65.0	26	19.4 (±21.37)	0.5101*
	Language Arts	91.1	26	28.8 (±19.88)	0.6169*
	Math	38.0	26	10.7 (±18.7)	0.3667*
	Social Studies	72.2	26	22.8 (±20.16)	0.4942*
3	Science	75.0	47	27.0 (±19.19)	0.6776*
	Language Arts	57.8	47	26.0 (±24.16)	0.5979*
	Math	80.15	47	11.6 (±24.03)	0.3813*
	Social Studies	94.3	47	27.0 (±33.64)	0.5109*
4	Science	97.0	71	27.0 (±15.12)	0.7132*
	Language Arts	77.5	71	23.1 (±19.85)	0.5332*
	Math	62.5	71	13.3 (±18.31)	0.3502*
	Social Studies	65.0	71	13.0 (±25.53)	0.2975*
5A	Science	72.0	54	19.0 (±20.64)	0.4775*
	Language Arts	80.0	54	19.8 (±20.39)	0.4749*
	Math	80.0	54	12.9 (±23.02)	0.3176*
	Social Studies	100.0	54	20.2 (±21.44)	0.4589*
5B[b]	Science	62.0	25	24.0 (±18.93)	0.5413*
	Language Arts	100.0	25	19.6 (±29.79)	0.4005*
	Math	66.3	25	20.8 (±23.79)	0.5070*
	Social Studies	80.0	25	22.0 (±20.0)	0.5519*

[a] Student outcomes are for the corresponding science, language arts, math, or social studies portion of the student pre-post assessment. [b] Two implementations were provided to two different students at one school by two separate teams of teachers
* $p < .05$

The mean fidelity score of 76.7% fell significantly below the researchers' initial expectations of fidelity scores in the range of 90%. Given the existence of some flexibility built into the model curriculum and the potential for teachers to respond to individual learning contexts,

researchers' expectations of extremely high fidelity to the "letter" of the curriculum were clearly too high. However, despite the wide range of fidelity scores (38% to 100%), statistically significant student knowledge gains between the pretest and the posttest were achieved in all but two cases. The pretest and posttest instruments were identical, and they were given immediately before and after implementation. Possibly, the interdisciplinary and mutually reinforcing nature of the learning activities within the curriculum may have created the situation where low fidelity may not have necessarily led to a negative impact on student learning outcome; core concepts and skills were constantly reinforced throughout and across discipline lessons. These results may support the findings of previous research on interdisciplinary curriculum showing increased levels of student achievement as a result of interdisciplinary instruction (Caine & Caine, 1991; Clark, 1997; Vars, 1996).

School-Level Results

Fidelity was not statistically correlated with student pre/post content assessment gain and effect sizes at the school level. There was also a weak, negative, but statistically insignificant correlation (r = -0.320; p = .536) between total knowledge gain and fidelity of implementation scores of the schools. Similarly, there was a moderate, negative, but statistically insignificant correlation (r = -0.541; p = .268) between the effect size and fidelity of implementation scores of the schools. These findings may suggest that as teachers modified the curriculum to suit their personal teaching styles/preferences and to address particular instructional contexts, their students were likely to gain more knowledge. This finding is consistent with previous research suggesting that as instruction is tailored to meet the specific needs of learners, levels of student engagement, and, therefore, student achievement increase (Marzano, 2003; Wenglinsky, 2002).

Overall school-wide fidelity rates, representing the mean total fidelity scores for the science, language arts, math, and social studies implementations aggregated for each school, range from 66.58 to 83.0 (Table 2). Mean effect sizes on student pre-post assessment gains from a school-wide perspective were near or above .50, except for the one school with a low n of participants (n = 10), in which case the effect size achieved was slightly less than .30. Table 2 visually depicts the lack of a relationship between implementation fidelity and content assessment pre/post effect. For example, the school with the lowest fidelity score produced the second highest effect size. Correspond-

ingly, the school with the highest fidelity score ranked fourth (of six) in terms of effect produced.

Table 2
School Level Fidelity Scores and Pre-Post Gain Effect Sizes

School	Fidelity Score	*N* of Students	Pre-Post Gain	Mean Effect Size
5	83.0	54	7.19 (4)	.5405*
1	81.38	10	4.31 (6)	.2820*
3	77.48	47	9.28 (1)	.6658*
5[a]	77.08	25	5.08 (5)	.3798*
4	75.50	71	7.66 (3)	.5989*
2	66.58	26	8.17 (2)	.6451*

[a]Two implementations were provided to two different students at one school by two separate teams of teachers
*$p < .05$

Content Area Fidelity Results

From the perspective of the fidelity scores for content areas across the schools, social studies and language arts received the highest fidelity scores, while the lowest fidelity scores were earned for math and science respectively (Table 3). Overall, the third-highest content area fidelity score, the science content area, produced the highest pre-post assessment effect size. This may be due to the fact that the entire curriculum was designed to be science-based with lessons in other content areas extending and reinforcing science-related concepts and skills.

Table 3
Mean Fidelity Score and Pre-Post Gain Effect Sizes by Content Area (6 schools per content area)

Content Area	*N* of Schools[a]	Mean Fidelity Score	Pre-Post Gain	Mean Effect Size
Social Studies	5	83.77	2.21 (±0.4820)	.4229*
Language Arts	5	78.8	1.49 (±0.4379)	.4866*
Science	5	75.17	2.17 (±0.5194)	.5580*
Math	5	71.16	1.96 (±0.5697)	.3926*

[a]Two implementations were provided to two different students at one school by two separate teams of teachers, resulting in 6 total implementations
*$p < .05$

Fidelity scores for three of the four content areas were in the 70s (78.8, 75.17, and 71.16), and the highest subject-area fidelity score (social studies) was only 83.77. This highest fidelity score produced the next

to lowest pre-post assessment gain score effect size (.4229). Math had the lowest mean fidelity score and the lowest mean effect size. The participants of this study who taught math were acutely aware of their students' abilities and made regular modifications to the curriculum to address students' needs. For example, one teacher heavily modified a graphing activity because he felt his students had already mastered graphing. As such, he had students calculate measures of central tendency instead of graphing. Social studies had the highest mean fidelity score, but the third (out of four) lowest effect size suggesting that, even though modifications were less frequent, perhaps the lessons were not optimized to meet the particular contexts and needs of these student populations.

Fidelity Score Ranges and Pre/Post Assessment Effect Sizes

As depicted in Table 4, regardless of content areas, fidelity scores in the 70–79 range (out of a possible score of 100) produced the highest pre-post content assessment effect size. The three content areas with a perfect (100) fidelity score produced among the lowest effect sizes. School effects appear to be minimal, given the distribution of schools across the various fidelity ranges; at least three schools are represented in all but the two lowest fidelity ranges.

Table 4
Fidelity Score Ranges and Pre-Post Gain Effect Sizes

N Fidelity Scores	# of Content Areas in Range	Mean Effect Size For Pre-Post Gain	# of Schools in Range
100	3	.4306*	3
90-99	4	.5163*	4
80-89	5	.4308*	4
70-79	4	.5456*	4
60-69	5	.4412*	2
0-59	3	.4202*	1

*$p < .05$

A fidelity score in the range of 70–79% appears to have been the optimal range in this study, as it produced the highest mean effect size. This range is well below the established a priori fidelity range of 90–100%. While the mean pre-post content assessment effect size for implementations with fidelity in the 90–99% range is high (0.5163), there was also significant variation in the fidelity scores of that range (SD = 0.2115) suggesting the mean effect size may be an unreliable

statistical result in this instance. The fidelity score range of 70–79% had the greatest mean effect size (0.5456) with much less variation (SD = 0.0910), suggesting that modifications of the curriculum to better suit students' learning needs and abilities were reasonable and even desirable.

Even though fidelity of implementation was not a predictor of gains in student knowledge or effect size in this study, the simple metric employed here helped the project staff gain deeper insight into how the curriculum was being used and, more importantly, modified by teachers to meet classroom contexts and needs. By thoroughly analyzing the observations and resulting fidelity scores, project staff were able to identify specific areas in each lesson that were frequently modified. Subsequent interviews with participating teachers allowed the project staff to better understand why modifications were necessary and helped the curriculum design team make revisions to the curriculum that resulted in a stronger, more cohesive instructional unit.

Teachers' Deviation from Fidelity

It was not possible for the program designers to anticipate the needs and contextual situation of all the students in this study. As a result, teachers were encouraged throughout the professional development to modify the curriculum to meet the highly specific needs of their students. During the observation of teacher implementations of the model curriculum, researchers systematically recorded all material deviations from the curriculum. Upon completion of implementation, interviews were conducted with teachers to determine teachers' rationales for these deviations. Several themes emerged from these interviews, primarily including (a) teacher modifications to reflect instructional activities occurring prior to implementation of the model curriculum, (b) teacher modifications resulting from time constraints or convenience, and (c) teacher modifications to adjust to the learning abilities of students at the time of implementation.

Several teachers modified the curriculum to account for the scope and sequence of instructional activities prior to implementation of the model curriculum. For example, one science teacher had previously taught a unit on plant cells in which students developed strong familiarity with preparing slides and using microscopes. This teacher modified the model curriculum to eliminate an activity that introduced students to microscopes, taught the parts of a microscope, and reviewed safety in handling slides and operating the microscope.

The teacher indicated that this activity would be redundant to one previously taught and that instructional time would be better spent allowing students a longer period of time to complete laboratory activities. Similar modifications were made by other science teachers and in the math component.

Another implementation challenge for teachers related to the blocking of time available for instruction. The model curriculum was designed to be taught in approximately 6–8 class periods based on 55-minute periods. However, many teachers in this study only had 45 minutes per period. As a result, some of these teachers opted to modify the curriculum to fit this time constraint instead of spending additional class periods to cover the content. For example, several math teachers elected to aggregate data collected from the science labs in a computerized spreadsheet program so that students could quickly calculate measures of central tendency as opposed to having the students do it themselves by hand or with calculators. They rationalized that the more relevant mathematical skill was in analyzing findings and drawing conclusions, as opposed to calculating statistics.

Another unanticipated implementation challenge was related to Internet and computer accessibility issues. Several of the curricular activities required students to have access to the Internet. Two of the participating schools did not have Internet access available to entire classes. As such, these teachers were forced to make modifications to the curriculum. One social studies teacher did not have access to a computer lab where each student could use the Internet; however, she did have Internet access at her desktop computer in the classroom. She modified one of the social studies activities so that the required website was displayed on a large television screen in her classroom. The class worked as a whole to complete the activity, as opposed to working individually as specified by the intervention protocol.

Finally, a number of teachers made modifications to adjust to the learning abilities of their students. One language arts teacher, for example, modified a cooperative group learning activity in a substantial manner. Her class was composed of low-level learners, many of whom were identified as learning disabled in reading. This teacher did not feel that her students could successfully read the nonfiction informational texts provided without significant teacher support. Instead of completing the activity in small groups, the teacher read the text aloud to the students and then helped the work through the activity, step by step.

While the modifications that teachers made throughout the implementations of the model curriculum negatively impacted the fidelity scores of this study, the resulting gains in student knowledge were high (Richards et al., 2008). As such, it seems that the researchers' expectation of 90–100% fidelity was inappropriate. Moreover, it also appears that the fidelity instrument may have been too restrictive. Given that teachers were encouraged to modify the curriculum to reflect and accommodate learners' needs at the time of implementation, the fidelity metric used should have been able to account for these modifications in a manner that would only penalize fidelity if the modifications were not based on student learning needs.

STUDY CONCLUSIONS AND IMPLICATIONS

A few substantive conclusions can reasonably be drawn from this study. First, high levels of implementation fidelity are quite difficult to achieve, even when implementation is carefully planned and executed. The range of fidelity scores across this project was much larger than expected, and rarely were any implementation fidelity scores near 100%. This is consistent with the literature suggesting that full implementation of project designs are often unrealistic (Rogers, 2003). Second, the highest implementation fidelity scores were was not associated with the most desirable project outcomes. In fact, overall fidelity in the range of 70–79% tended to produce the most favourable outcomes, supporting the literature suggesting that high levels of fidelity may in fact limit intervention effectiveness (Leventhal & Friedman, 2004). Third, this study produced evidence that teachers modified the curriculum during implementation to address localized circumstances such as more immediate student needs, advantageous curriculum re-sequencing opportunities, or unexpected changes in resource availability (e.g., computer access). These changes appeared to reflect meaningful professional judgements on the part of teachers implementing the intervention model, and this lends support to aspects of the literature (e.g., Fullan, 2001; Shulman, 1990) suggesting that teachers retain a fundamental professional responsibility to respond to the real world and ever-changing contextual demands of the classroom.

A few related implications can be derived from this study. Although measuring and accounting for project implementation fidelity is complex and time-consuming, it is essential for internal validity purposes because linking project outcomes to an intervention requires that the actual intervention, as implemented, be known and specified. In

addition, project developers need to reconsider their expectations for fidelity, especially very high fidelity, since it appears to be counter-productive. Researchers and program developers also need to gain a deeper insight into the decisions of implementers (teachers in this study) when they choose to deviate from implementation protocols. This insight can be invaluable for formative evaluation purposes, leading to more contextually sensitive intervention designs. However, the most important implication of this study is that monitoring implementation fidelity offers potential insight into the resultant outcomes. Researchers who assess fidelity should be able to gain some understanding as to which implementation changes were associated with the most desirable outcomes. Finally, in this study the results may have been somewhat confounded by the interdisciplinary nature of the educational intervention, suggesting that future fidelity studies of classroom implementations should differentiate between interdisciplinary and single-discipline interventions.

REFERENCES

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 from the NIH behavior change consortium. *Health Psychology, 23*(5), 443–451.

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

Caine, R.N., & Caine, G. (1991). *Making connections: Teaching and the human brain*. Alexandria, VA: Association for Supervision and Curriculum Development.

Clark, S. (1997). Exploring the possibilities of interdisciplinary teaming. *Childhood Education, 73*(5), 267–272.

Chen, H.T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.

Dumas, J., Lynch, A., Laughlin, J., Smith, E., & Prinz, R. (2001). Promoting intervention fidelity. Conceptual issues, methods, and preliminary results from the early alliance prevention trial. *American Journal of Preventive Medicine, 20*, 38–47.

Dusenbury, L., Branningan, R., Falco, M., & Hanson, W.B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research Theory and Practice, 18*, 237–256.

Fuchs, L.S., Fuchs, D., & Karns, K. (2001). Enhancing kindergartens' mathematical development: Effects of peer-assisted learning strategies. *Elementary School Journal, 101*, 495–510.

Fullan, M. (2001). *The new meaning of educational change*. New York: Teachers College Press.

Galbo, C. (1998). Helping adults learn. *Thrust for Educational Leadership, 27*(7),13–15.

Hennessey, M.L., & Rumrill, P.D. (2003). Treatment fidelity in rehabilitation research. *Journal of Vocational Rehabilitation, 19*, 123–126.

Hester, P.P., Baltodano, H.M., Gable, R.A., Tonelson, S.W., & Hendrickson, J.M. (2003). Early intervention with children at risk of emotional/behavioral disorders: A critical examination of research methodology and practices. *Education and Treatment of Children, 26*(4), 362–381.

Leventhal, H., & Friedman, M.A. (2004). Does establishing fidelity of treatment help in understanding treatment efficacy? Comment on Bellg et al. (2004). *Health Psychology, 23*(5), 452–456.

Marzano, R.J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mowbray, C.T., Holter, M.C., Teague, G.B., & Bybe, D. (2003) Fidelity criteria: Development, measurement., and validation. *American Journal of Evaluation, 24*(3), 315–340.

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*(1), 33–84.

Penual, R., & Means, B. (2004). Implementation variation and fidelity in an inquiry science program: Analysis of GLOBE data reporting systems. *Journal of Research in Science Teaching, 41,* 294–315.

Richards, J., Skolits, G.J., Burney, J., Pedigo, A., & Draughon, F.A. (2008). Validation of an interdisciplinary food safety curriculum targeted at middle school students and correlated to state educational standards. *Journal of Food Science Education, 7*(3), 54–61.

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

Sanchez, V., Stekler, A., Nitirat, P., Hallfors, D., Cho, H., & Brodish, P. (2007). Fidelity of implementation in a treatment effectiveness trial of Reconnecting Youth. *Health Education Research, 22*(1), 95–111.

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

Smith, S.W., Dunic, A.P., & Taylor, G.G. (2007). Treatment fidelity in applied educational research: Expanding the adoption and application of measures to ensure evidence-based practice. *Education and Treatment of Children, 30*(4), 121–134.

U.S. Department of Education. (2003, December). *Identifying and implementing educational practices supported by rigorous evidence: A user-friendly guide.* Washington, DC: Institute for Education Sciences.

Vars, G.F. (1996). The effects of interdisciplinary curriculum and instruction. In P.S. Hlebowitsh & W.G. Wraga (Eds.), *Annual review of research for school leaders* (pp. 147–164). Jefferson City, MO: Scholastic.

Vaughn, S., Cirino, P.T., Linan-Thompson, S., Mathes, P.G., Carlson, C.D., & Hagan, E.C. (2006). Effectiveness of a Spanish intervention and an English intervention for English-language learners at risk for reading problems. *American Educational Research Journal, 43,* 449–487.

Wenglinsky, H. (2002, February 13). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, *10*(12). Retrieved June 12, 2005, from http://epaa.asu.edu/epaa/v10n12/

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using curriculum-based instructional management systems to enhance math achievement in urban schools. *Journal of Education for Students Placed at Risk, 8,* 247–265.

Gary Skolits, Ed.D., is the Director of the Institute for Assessment and Evaluation at the University of Tennessee, Knoxville. He is also an assistant professor and coordinator for the Evaluation, Statistics, and Measurement Ph.D. program, engaging in research focused on evaluation methods, organizational accountability, and professional development interventions.

Jennifer Richards, Ph.D., is an experienced middle school teacher who now directs a multi-state educational research project at the University of Tennessee, Knoxville. Her areas of research include professional development models for teachers, interdisciplinary curricula, and effective instructional strategies.