APPROACHES TO MEASURING IMPLEMENTATION FIDELITY IN SCHOOL-BASED PROGRAM EVALUATIONS

Leonard Bickman
Vanderbilt Universiy

Manuel Riemer
Wilfrid Laurier University

Joshua L. Brown
Harvard University

Stephanie M. Jones
Fordham Universitys

Brian R. Flay and Kin-Kit Li
Oregon State University

David DuBois
University of Illinois at Chicago

William Pelham, Jr. and Greta Massetti
University of Buffalo

This article focuses on issues related to implementing school-based social and character development programs in the context of a large multiprogram evaluation study funded by the Institute of Education Sciences and the Centers for Disease Control and Prevention. Implementation analysis is a relatively new but important research area. The first section describes why this analysis is especially important when null results are obtained. It is also noted that not only does the intervention need to be documented but activities at the comparison site require explication. The wide range of implementation methods used at four of the seven sites in the Social and Character Development Research Program is described and difficulties in studying implementation at multiple sites are discussed.

Assessing program implementation is a critical element of an evaluation. Berman and McLaughin (1976, p. 349) describe it as, "The bridge between a promising idea and its impact on students is implementation;" however, "innovations are seldom implemented as planned," (as cited in Dusenbury, Brannigan, Falco, & Hansen, 2003). Studies and reviews that have investigated the diffusion of efficacious innovative educational interventions

• **Leonard Bickman,** Center for Evaluation and Program Improvement, Vanderbilt University, Peabody College, Peabody #151, 230 Appleton Place, Nashville, TN 37203-5721. Phone 615.322.8694. E-mail Leonard.Bickman@vanderbilt.edu

Journal of Research in Character Education, 7(2), 2009, pp. 75–101 ISSN 1543-1223

have consistently found a serious lack of implementation fidelity (Dusenbury et al., 2003; Hallfors & Godette, 2002; Rohrbach, Graham, & Hansen, 1993). Without efforts to enhance ongoing implementation, such as typically found in efficacy studies, fidelity also appears to deteriorate the longer a project is in existence. Rohrbach et al., for example, found that 78% of the teachers, who participated in the Adolescent Alcohol Prevention Trial, implemented at least one program lesson during the first year. In the second year this dropped to one out of four. The effectiveness of the program cannot be adequately judged if a program is not implemented with fidelity to the theoretical model and program benchmarks. This assumes that there are theoretical models and program benchmarks, which rarely exist. This article will describe the analytic and conceptual problems of conducting implementation evaluations of four unique character education programs.

One of the most cited reasons for assessing implementation fidelity is the need to account for null-effects or ambiguous results should they occur (Bickman & Athay, 2009; Hohmann & Shear, 2002). It is also important to document that the treatment and control or comparison conditions were indeed similar at the start of the study and different once the study is completed (Bond, Evans, Salyers, Williams, & Kim, 2000; Mowbray, Holter, Teague, & Bybee, 2003). Measures of fidelity can be used in a correlational approach to predict outcomes, especially when there are no significant differences in outcomes between the treatment and control groups. In essence this is a form of a dose-response analysis that relinquishes the higher internal validity of an experiment to examine if there is any conceptual support for the notion that the amount or quality of the independent variable correlates with the outcome (Dane & Schneider, 1998; Mowbray et al., 2003). In multisite studies fidelity assessments are critical to ensure that the interventions across sites are similar and significant differences are captured (Paulson, Post, Herincks, & Risser, 2002). Finally, ongoing assessment of fidelity is the prerequisite for continuous quality improvement and the prevention of program drift (Bickman & Peterson, 1990; Riemer, Rosof-Williams, & Bickman, 2005).

Although the study of implementation is in its infancy, specific strategies have been promoted to help understand the degree to which programs are implemented as intended, such as (a) linking program theory to teacher training and support, (b) assessing training quality—whether program elements are implemented, the quality of implementation, and (c) assessing teachers' attitudes toward implementation (Domitrovich & Greenberg, 2000). Utilizing implementation information may not only inform evaluators in their interpretation of evaluation findings, but may also inform interventionists in their quest for program improvement.

GENERAL UNDERSTANDING OF IMPLEMENTATION FIDELITY IN THE LITERATURE

Implementation fidelity, also referred to as treatment or program integrity, has been defined in several different ways. Dusenbury and colleagues (2003), for example, define it as "the degree to which teachers and other program providers implement programs *as intended by the program developers*" (p .249; italics in original). Thus, the extent to which fidelity can be assessed is closely related to the intention of the program developers. Ideally, these intentions are elaborated as part of a program theory or model (Bickman, 1987, 2000; Chen, 2005; Riemer & Bickman, in press). Bond and colleagues (2000), therefore, refer to fidelity simply as "the degree to which a particular program follows a program model" (p. 75).

Fidelity criteria described in the literature can be grouped into five major categories (Dane & Schneider, 1998; Dusenbury et al., 2003): (1) adherence to the program, (2) dose (the amount of the program delivered), (3) quality of program delivery, (4) participant responsiveness and (5) program differentiation. Fidelity criteria are also differentiated by whether they describe structure (i.e., the framework for service delivery) or process (i.e., the way in which services are delivered) (Mowbray et al., 2003).

However, it is important to recognize that program theories or models seldom are specified in sufficient detail to allow the researcher to draw reliable and valid conclusions about specific implementation questions. For example, quality of implementation remains a nebulous concept that is difficult to operationalize (Bickman & Salzer, 1997). It is also not easy to specify the amount of input that is necessary to produce the desired outcome. For example, is 4 hours of training sufficient or are 12 hours needed before the training could be expected to have an effect? Too rarely are parametric studies conducted on the program elements that allow such insights. The same problem occurs in determining the length of any intervention or how long an effect should last. Programs usually are complex mixes of components but it is not clear if all are truly necessary or if they need be implemented to the same degree of intensity, order or quality. It is also incumbent on the program developers to be as specific as possible in describing the contexts in which they think their programs will be effective. Such details are seldom provided by program developers because there are typically no extant data by which to make those judgments. Thus, tying implementation quality to the intent of the developers is not an entirely satisfactory solution to the problem of defining fidelity.

HOW FIDELITY IS TYPICALLY MEASURED

In a recent NIMH workshop on advancing implementation science the participating experts described "several major principles of a research approach for implementation, which include qualitative and quantitative methods, multiple time points, and multiple viewpoints" (Chambers, 2008, p. 7). In general, studies frequently use mixed methods in investigating the process of implementation. Studies have used semistructured interviews (Elliott & Mihalic, 2004; Torill & Oddrun, 2007), case studies (Clapp, Burke, & Stanger, 1998), review of secondary documents and/or administrative databases (Weiner, Helfrich, Savitz, & Swiger, 2007; Zinn, Mor, Feng, & Intrator, 2007), direct observation (Hitt et al., 2006), focus groups (Stevens, Bourdeaudhuij, & Oost, 2001), questionnaires (Inouye, Baker, Fugal, & Bradley, 2006), and participatory action research techniques (Waterman et al., 2007). The most common methods are (a) ratings by (program, research, or content) experts based on interviews, classroom observations, video taping, program documents such as activity checklist or logs, site visits, and record reviews, and (b) surveys or interviews completed by individuals delivering the services (e.g., teachers) or receiving them (e.g., students) (Mowbray et al., 2003). In general, Mowbray and colleagues (2003) propose three steps in establishing criteria for assessing implementation fidelity: (1) identify possible indicators or critical components of the program model, (2) collect data to measure the indicators, and (3) examine the indicators in terms of their reliability as well as validity. However, these are methods of data collection and do not resolve the conceptual problems in defining quality or quantity of fidelity of implementation.

ASSESSING FIDELITY IN THE CONTEXT OF THE SACD RESEARCH PROGRAM

The Social and Character Development (SACD) Research Program is a randomized trial of seven universal school-based programs that aim to promote social competence, reduce problem behavior, and improve school climate, including a multiprogram evaluation. Details about this research program are described in Haegerich and Metz (this volume). As part of this multisite evaluation the members of the research consortium tried to find ways of assessing implementation fidelity that will do justice to the unique aspects of each program but also provide a general indicator of fidelity that can be used for the national evaluation across programs. As discussed above, assessing the level of implementation is critical for the validity of the evaluation findings. At the level of each program there was no specific common fidelity evaluation approach taken by all seven projects in the consortium. However, the national evaluator used the same measures of implementation at all sites as described later. Each evaluator decided what approach best fitted his or her project and the data to be collected. The national level, where results from the seven programs were aggregated, used common definitions described in the next section. Four evaluation groups volunteered to contribute a description of the approach they used for this article. The following description of four examples of measuring fidelity of SACD program implementation, and discussion of an attempt to integrate these assessments into a general indicator of fidelity that can be used across programs, will provide researchers and practitioners with valuable insights into some strategies as well as challenges of assessing fidelity. While there seems to be an agreement that measuring fidelity is critical, there is a continuous struggle how to best go about that. The efforts of the SACD research program are no exceptions in this regard. It is important to note that impact findings of the multiprogram evaluation will not be presented in this article, but will be disseminated in a SACD Consortium-authored report released by the Institute of Education Sciences (IES). In addition there will be a series of empirical articles published in the scientific literature by the different Consortium members.

HOW IMPLEMENTATION FIDELITY WAS DEFINED, OPERATIONALIZED, AND MEASURED IN THE SACD RESEARCH PROGRAM

The SACD Consortium has defined fidelity as a measure of how well a program is implemented in practice compared to the developer's intentions for the program, and consists of the amount (content) and quality (process) of the intervention delivered. Amount of implementation refers to the extent to which teachers and schools implement the intended components of the program, whereas quality of implementation refers to the extent to which teachers and schools implement the program *in the manner in which it was intended*. Furthermore, amount and quality may be related but are not fully dependent on each other. For example, some teachers may implement all components of a program (fidelity of amount), but they may not do so with full integrity or quality. Likewise, some teachers may implement only some components or elements of an intervention, but those components that are in place are done with very high quality and engagement on the part of the teacher. It is usually unknown whether amount or quality will have a larger effect on the outcomes. As pointed out earlier, it is also unlikely that the developers have defined some minimal value that, if not achieved, then the intervention should be considered not implemented and thus the program theory not testable. Reporting statistically significant differences between experimental conditions is a good start, but begs the question of whether the differences were meaningful. Amount and quality of implementation can differ within a school (e.g., some teachers might be better able, more skilled, or more motivated to deliver a program), as well as across schools (e.g., some school administrators might be more committed to supporting program implementation than others).

Because each program included different intervention components and strategies, investigators at each site designed the approach to measurement of fidelity in accordance with their approach to measuring implementation and their program model. The models for assessing fidelity will be briefly described for four of the sites described in this article.

At the national level (used in all sites) the multisite workgroup developed a fidelity model that characterized program elements across sites based on the intended agents and targets of the intervention (e.g., program elements delivered *by* teachers *to* students). Each site identified intervention components provided to: (1) teachers (e.g., teacher training, faculty boosters, weekly coaching); (2) students by teachers (e.g., daily curriculum, daily behavior management strategies); (3) parents (when applicable; e.g., parent meetings and distribution of program materials); and (4) all members of the school (e.g., assemblies, school-wide rules and expectations). Across sites, measures utilized included counts of trainings attended, teacher reports on program implementation, and teacher observations, among others (although measures were tailored to the individual programs under evaluation). Program fidelity was judged by benchmarks defined by program developers.

In order to have the fidelity data in a format that is comparable across sites, each local evaluation team identified two individuals with the most experience in monitoring program implementation in the SACD treatment schools and interpreting the site-specific fidelity data. These local "experts" were instructed to base their fidelity ratings on the site-specific fidelity data collected as part of the complementary study, the program benchmarks, and the rater's own judgment. No training or further instructions were provided to these individuals nor was there training required ensuring that reliable and valid judgments were being made or that there was consistency across the seven projects. Using a 5-point scale (substantially below = 1 to substantially above = 5) they rated the degree to which the local program benchmarks were met. Then, their ratings were compared and they were asked to come to an agreement in regard to each rating and provide a justification for each rating. These final ratings were then used by the independent national evaluator in the analysis of the multiprogram evaluation data.

In the national multiprogram evaluation, each of the schools were characterized as either "high implementers" or "low implementers". In the first year the high and low implementers outcomes were compared, but in subsequent years the patterns of implementation (e.g., high, high, low) were used to categorize the degree of implementation for each school across years.

The greatest challenge for the national multiprogram evaluation approach to fidelity is the difficulty in comparing the degree and fidelity of implementation across programs. The seven programs in this evaluation were not selected because they share important features in their program theory (see Haegerich & Metz, this volume, for a description of the SACD Research Program selection process). In addition, these programs differ significantly in their approach to character education as well as the types of character traits they targeted. They also differ in their intensity, breadth and cost. For example, some programs have a few hours of teacher training while others require several days. Thus, the main aspects they shared are their focus on improving social development outcomes and reducing problem behaviors and that they were all school-wide (delivered in all grades). Thus, it is doubtful that a program that is very intensive and has been implemented somewhat below their program's benchmark would be comparable to a program with low intensity that also has been implemented somewhat below their program's benchmarks. These benchmarks were not equated across the seven programs. Further, it is not known, or possibly even knowable, if the fidelity scores are comparable across the seven programs. For example, if a program with three program components that are known to be very effective but only two were implemented can that program be compared to another program that has one only one program component of unknown effectiveness but which was fully implemented? Also, what does it mean for policy and funding decisions if two programs produce comparable outcomes but one is much harder to implement with greater fidelity than the other? Clearly, the multiprogram evaluation faced challenges in regard to fidelity assessment that may be insurmountable given the design of the study. Of course these problems are not unique to the SACD Research Program.

Comparable social and character education programs and activities at control schools during time period of the evaluation were also assessed by the national multiprogram evaluator. The information about these activities was gathered in two ways. First, the *Teacher Report on Classroom and School* (TRCS) is a self-administered questionnaire completed at each data collection point by the third, fourth, and fifth grade teachers in a school. This instrument included items asking the teachers to report the level of SACD activity in the classroom and school. Second, interviews with principles were conducted once a year asking them about SACD programs and activities at their school. These data emphasize that the control condition for the SACD project is not a "no treatment" control but a "standard practice" control. Standard practice at the control schools included high use of "SACD-like" activities in the classroom and the school, use of a variety of teaching materials and instructional practices to accompany these activities, and high staff attendance at related professional development. Some of the research teams for the individual site evaluations administered additional measures of SACD-like activities in the control group as mentioned in the respective sections below.

The analytic and conceptual problems at the national multiprogram level were not as severe at the local level where there is variation in evaluation approaches taken by each site. Thus, in the following section, authors contributing to this article will summarize, for their respective sites, how fidelity was defined and measured, how successful the measurement of implementation was, what the implementation results were, how the local team is planning to use the implementation data in further analyses, and what some of the lessons are that they learned in regard to implementation and the measurement of fidelity. This will be followed by a discussion of the similarities and differences and the implications of the findings for the field.

PROJECT-SPECIFIC DESCRIPTIONS OF THE STUDY OF IMPLEMENTATION

Reading, Writing, Respect & Resolution (4Rs)

The 4Rs program, tested in NY City schools, is a universal, school-based intervention in literacy development, conflict resolution, and intergroup understanding that integrates social and emotional development into the language arts curriculum for grades K–5. The 4Rs uses high quality children's literature as a springboard for helping students gain skills and understanding in the areas of handling anger, listening, assertiveness, cooperation, negotiation, mediation, building community, celebrating differences and countering bias. By highlighting universal themes of conflict, feelings, relationships, and community, the 4Rs curriculum is designed to add social and emotional meaning and depth to rigorous literacy instruction. The 4Rs Program provides a pedagogical link between the teaching of conflict resolution and the teaching of fundamental academic skills, thereby capitalizing on their mutual influence on successful youth development (see Flay, Pelham, Berkowitz & Bier, this volume, for more details).

Measurement of Fidelity

School-wide implementation of the 4Rs Program began in September 2004 and proceeded throughout three consecutive school years with no major disruptions or significant problems. Implementation of the primary components of the 4Rs Program were systematically tracked and monitored using several measures designed to capture both the quantity and quality of implementation (fidelity) as well as the degree to which students received the 4Rs program (dosage). As noted by Flay, Pelham, Berkowitz, and Bier (this volume), the 4Rs program has 2 primary components: (1) 25 hours of training (5 days x 5 hours per day) followed by ongoing coaching of teachers to support them in teaching the 4Rs curriculum with a minimum of 12 contacts in one school year; (2) a comprehensive 7-unit, 35–50 lesson, literacy-based curriculum in conflict resolution and social-emotional learning. Each unit also includes additional activities and related readings. Procedures for completing the implementation measures for each of the 4Rs components are described below.

Training. (1) Training Logs were completed for each day of training delivered by 4Rs Staff Developers to school teachers. Three separate forms (Introductory Training Checklist, Teacher Attendance Form, Teacher Stipend Form) together provide the following data: school information, names of teachers present, time in/time out of training (hours spent), content (specific topic/s of training agenda covered and time spent on each), and format of session (small/large group, workshop, individual). (2) Anonymous Training/Workshop Evaluation forms were completed by teachers evaluating the usefulness and quality of each session attended. These forms were used by the program partners, the Morningside Center for Teacher Social Responsibility, to monitor, reflect on, and rapidly respond to the degree to which the training sessions were reported as effective, useful, and of high quality by teacher participants.

Ongoing coaching. (1) 4Rs Staff Developer logs were completed at each point of contact with a teacher, principal or parent detailing school/teacher/other information, school period, types of contact, contact time (minutes), unit/theme of contact, discussion of goals completed and planning of goals/next steps/follow-up. These logs were completed at each point of contact and were delivered to program senior staff on a monthly basis and were used to facilitate quality management at monthly 4Rs Staff Developer meetings. (2) 4Rs Staff Developers completed “Ongoing Coaching” ratings of quality of teacher delivery of the curriculum; and quality of the “process” of staff developer work and engagement with the teacher. These ratings were also completed on a monthly basis.

4Rs curriculum. (1) Weekly logs were completed by teachers detailing school/teacher information, content/type of 4Rs activity/lesson and day of delivery, time spent on 4Rs activities/lesson each day, checklist of activity/lesson content, and rating of degree of student comprehension and engagement in the 4Rs activities/lesson that week. Teachers completed logs each week, 4Rs School Liaisons circulated and collected logs at the end of each week, and 4Rs Staff Developers collected logs each month to be delivered to program senior staff for recording and use in monthly quality management meetings.

Sample and Participants

With regard to the sample on which the brief implementation results presented below are based, a total of 292 teachers across Grades K–5 are included with 67 or 23% present in Year 1 only, 158 or 54% present in both Years 1 and 2, and 67 (23%) present in Year 2 only (note, Year 3 is not yet available). In general teachers in this sample were predominantly Caucasian (53%) followed by AfricanbAmerican (28%) and Hispanic (18%). The vast majority of teachers were female (88%). Teachers reported on average 6.8 years experience as a teacher (with a standard deviation of 6.6 years) and on average 4.7 years teaching *in this school* (with a standard deviation of 4.4 years). Finally, approximately 75% of the teachers held a MA degree at the time of their participation.

Success of Implementation Measurement and Implementation Results

As expected there was variability in 4Rs implementation among teachers, grades, and schools. Moreover, as is common in urban public schools, there is turnover between years in teachers, with teachers leaving and new teachers entering schools. This is not inconsistent with similar programs and evaluation studies that focus on public schools (e.g., Kam, Greenberg, & Walls, 2003). Table 1 describes for Years 1 and 2 (note, Year 3 is not yet available), and by school, the amount of (a) teacher training and ongoing coaching provided to teachers by 4Rs staff developers, (b) the average number of 4Rs activities completed by teachers per week, and (c) the percentage of weeks in the school year in which teacher logs were returned.

According to the implementation data from Year 1, teachers in the 9 treatment schools received (a) on average 2.4 days of training in the delivery of the 4Rs curriculum, and (b) an average of 38 staff developer coaching days. On average, teachers delivered three-quarters of a lesson per week, with the majority closer to the benchmark of one lesson per week. Perhaps more interesting, and likely reflective of the enormous pressures facing public school teachers today, the majority of teachers appear to have spent on average between 20–25 (~40 minutes/week) total hours during Year 1 on 4Rs. Year 2 implementation data reveals a slight decrease in training days, and a slight increase in coaching days and the average classroom lessons per week, and the amount of time spent on 4Rs per week. Moreover, the data indicate that teachers who were trained in the first year of the study, and who remained in the school the following year, were even closer to program benchmarks (i.e., on average they implemented one lesson/week and spent ~50 minutes on 4Rs per week). Unfortunately, 4Rs implementation data for all of Year 3 is not yet available; however, data from the early units of the 4Rs curriculum in Year 3 indicate that teachers completed 0.92 acts per week on average, an increase from 0.75 in Year 1 and 0.80 in Year 2. In addition, while the average number of training days decreased in Year 3 to approximately 1 day per school, the average number of coaching days increased to 44 days per school. In addition, we calculated the average number of weeks teachers completed logs, but reported “No 4Rs” for that week. In Year 1 the average number of weeks teachers reported “No 4Rs” ranged from 1.3 in School C to 10.3 in School B. In Year 2 the average number of weeks teachers reported “No 4Rs” ranged from 2 in School F to 8 weeks in School H. Despite this variability across schools within year, by the end of elementary school, children in the research cohort were exposed to the 4Rs Program for 3 consecutive years, both directly through lessons delivered by their classroom teachers, and indirectly through exposure to other students across all grades who were also exposed to 4Rs.

TABLE 1
Summary of 4Rs Implementation Through Year 2

	Year 1 Teachers (n = 273)				*Year 2 Teachers (n = 273)*			
School	*Train Days*	*Coach Days*	*Class Acts Per Wk*	*% Wks Logs Returned*	*Train Days*	*Coach Days*	*Class Acts Per Wk*	*% Wks Logs Returned*
A	2.5	45	.82	63	2.0	23	.88	100
B	2.5	30	.59	100	2.0	48	.71	99
C	2.5	45	.99	100	2.0	37	.77	93
D	2.0	21	1.00	100	2.0	20	.91	95
E	3.0	35	1.00	100	2.0	30	.95	85
F	2.5	52	1.00	80	3.0	48	.81	78
G	2.0	45	.52	49	2.0	53	.59	90
H	2.5	34	.86	98	2.0	40	.55	96
I	2.0	34	.57	78	2.0	53	1.00	98
Mean	2.4	38	.75	85	2.1	39	.80	93

Planned Use of Implementation Data

As described above, there was variability in teacher's fidelity of implementation of the 4Rs curriculum. To capitalize on this variability and examine the impact of treatment dosage on child outcomes, background teacher demographic and experiential characteristics will be examined in relationship to 4Rs implementation using propensity scores methods (e.g., Lochman, Boxmeyer, Powell, Roth, & Windle, 2006; Hill, Brooks-Gunn, & Waldfogel, 2003). In short, in the treatment schools a benchmark criterion representing either low or high quantity/quality implementation by 4Rs teachers will be modeled as a function of teacher demographic and experiential background characteristics. We will then use these propensity profiles to draw a comparable group of teachers from the control condition with the goal of producing an unbiased, experimental assessment of the treatment impact on children for high or low implementers. To account for the important issue of teacher change between each year of the intervention (i.e., teacher turnover between Grades 3, 4, and 5) we will consider conducting a set of piecewise regressions in which a teacher/classroom high or low implementer dummy variable is included as a predictor of within year change in elementary school. Teachers classified as "high 4Rs implementers" are those who reach expected program benchmarks of lessons delivery. The basic benchmarks are that teachers complete approximately one 4Rs activity per week, spending approximately 40 minutes on the activity. In Year 1, the lowest implementation year, approximately half the third grade teachers (n = ~45) across the 9 intervention schools met this criterion. Our assumption is that we will be able to draw a similar number of propensity-matched teachers from the control group resulting in ~90 teachers in total.

Lessons Learned

Three sets of interrelated lessons can be drawn from our examination of the implementation of the 4Rs Program over the first 2 years of our study. First, multiyear implementation practices must account for high rates of teacher turnover through intensive training of new teachers each year combined with ongoing implementation mentoring and support networks by previously trained teachers, both of which will maximize the likelihood of effective program sustainability. Second, documentation of program implementation needs to be better integrated into ongoing program practice and perceived as a tool to monitor and promote effective implementation rather than as an additional practice burden. Third, the integration and ongoing monitoring of program implementation by the program practitioners needs to be combined with regular presentation and discussion of implementation data with teachers and school administrators. Incorporating a data driven implementation model in these ways will facilitate the cross-year training and sustainability of the program.

POSITIVE ACTION (PA)

The PA program is a multicomponent, SACD school-based program designed to improve academics, student behaviors and character. It is grounded in a broad theory of self-concept (Purkey, 1970), is consistent with comprehensive theories of health behavior like the Theory of Triadic Influence (Flay & Petraitis, 1994), and is described in detail elsewhere (Flay & Allred, 2003; Flay, Allred, & Ordway, 2001; Flay, Pelham, Berkowitz & Bier, this issue; and at the program website (www.positiveaction.net). Briefly, the PA program tested in this study consists of K–8 classroom curricula, school-wide climate changes undertaken by the principal and a PA coordinator/committee, and family involvement components. The sequenced elementary curriculum consists of 140 lessons per grade, per academic year, offered in 15–20 minute periods (approximately 1 hour/week). Lessons cover six major units on topics related to self-concept, mind and body positive actions (e.g., nutrition, physical activity, decision-making skills, motivation to learn), social/emotional actions for managing oneself responsibly (e.g., emotional regulation, time management), getting along with others (e.g., empathy, respect, treating others as one would like to be treated), being honest with yourself and others, and self-improvement (e.g., goal setting, courage to try new things, persistence). The program utilizes an interactive approach, whereby interaction between teacher and student is encouraged through the use of structured discussions and activities, and interaction between students is encouraged through structured or semi-structured small group activities, including games, role plays and practice of skills. The school-climate and parent kits encourage and reinforce the six units of *PA* school-wide.

Measurement of Fidelity

In a prior trial of PA, we used a combination of qualitative and quantitative data on teacher implementation of the PA program to inform and further develop a working model of influences on the amount and quality of teacher implementation of the curriculum and other classroom-based program components (Beets, Flay, Vuchinich, Acock, Li, & Allred, 2008; Fagen, 2003; Fagen & Flay, 2006). Preliminary analyses of qualitative and quantitative data identified several factors that appear to be influential in shaping integrity of teacher implementation. These include the extent to which teachers receive support from their principal, collaborate with and receive support from other teachers when implementing the program, as well as teacher's own attitudes and beliefs regarding the need for schools to do SACD and the likely effectiveness of SACD programs.

At the school level, the most promising prevention programs positively impact school climate and these effects appear to promote better student outcomes (Adelman & Taylor, 2000; Griffith, 2000; Kuperminc, Leadbetter, Emmons, & Blatt, 1997; Roeser, Eccles, & Sameroff, 2000; Greenberg, Domitrovich, & Bumbarger, 2001). Because school climate effects of SACD programs such as PA are most likely to accrue through school-wide program components (e.g., coordinating committee, assemblies, use of common terminology, reinforcement of positive behaviors, involvement of family), it is critically important to assess the integrity with which these activities are implemented and the factors that affect integrity.

High levels of implementation integrity are necessary for individual students and classrooms of students within schools to receive high levels of exposure to program activities (i.e., dosage). Program effects typically appear greater when focusing on students with greater levels of program exposure and participation. Selection effects (e.g., those teachers who are already prone to elicit positive student outcomes also tend to deliver the program at higher levels) may seriously bias such analyses, although data analytic procedures such as propensity score analysis can be used to attempt to control for these types of confounds (Foster, 2003; Rosenbaum & Rubin, 1983).

The measures used in this study are summarized in Table 2. Measures of implementation included both local program-specific measures and multisite measures. Program-specific measures provided information on the implementation of the PA program among the program schools, while multiprogram evaluation measures assessed the general SACD materials and teaching strategies used in both program and comparison schools. Multiprogram evaluation measures provided less specific information about the implementation of the program schools, but created a common ground for comparison between program and comparison schools in terms of level of use of SACD-related activities.

Sample and Participants

The PA program was tested in 14 elementary schools in the Chicago Public Schools (CPS) system during the 2004-05 through 2006–07 academic years, 7 in the control con-

TABLE 2
Instruments Used to Assess Implementation Fidelity—Positive Action

Instrument	*Source*	*Description*
Weekly Implementation Report (web-based)	All teachers	Teachers' self-report on amount and quality of classroom PA activities
Unit Implementation Report (web-based; at the end of each of the 7 units)	All teachers	Same as above plus teachers' perceived effectiveness of and attitudes toward program
End-of-year Process Survey	Cohort teachers	Same as Unit Implementation Report
End of year Process Survey	Principals, PA coordinators	School-level implementation activities

dition and 7 in the intervention condition. The study followed a single cohort of students who were third graders when treatment schools began implementing the program in the 2004–05 academic year. Five assessments were conducted during the study period: Fall 2004 (baseline assessment), Spring 2005, Fall 2005, Spring 2006, and Spring 2007 (end of Grade 5). A high rate of return for parental consent forms by students (98%) was achieved by a class incentive and class visits and the affirmative rate (79%) was satisfactory. Rates of consent and assent and of data being obtained at baseline and at the final assessment were not significantly different across treatment and control conditions. Assessments and procedures were approved by the Institutional Review Boards of MPR, the University of Illinois at Chicago and Oregon State University.

Depending on the measure (see below), the implementation data discussed in this article were collected from all teachers in the study schools (N ~ 420), teachers of the study cohort students (N ~ 35), PA Coordination Committee members (N ~ 40), and principals from participating schools. Sample numbers varied a little across years due to fluctuations in school enrollment. Moreover, as is common in urban public schools, there was turnover between years in teachers (Kam et al., 2003).

Success of Implementation Measurement

By the use of extensive reminders and incentives, we were able to obtain Weekly Implementation Reports from an average of 74% of teachers, and Unit Implementation Reports from an average of 85% of teachers by the third year of program implementation. In addition, the rates of response for all the End-of-Year implementation reports from cohort students, cohort teachers, and the principals and site-coordinators were high (range 94% to 100%), except for the Year 3 cohort teachers (64%). This low response rate by teachers at year 3 was due to missing responses from two schools; however, most of the other implementation records from those schools were completed. The completion rates for the multisite measures of implementation (i.e., reported SACD-like activities) across the three years were satisfactory (range 91% to 100%).

Implementation Results

There was wide variability between schools in all of the above implementation indices, especially in year 1, with improvements over time. For this report, we limit our results to summaries of data from the Unit Implementation Reports and the multiprogram evaluation reports of use of SACD-like activities. The results of the data reported by the teachers on the Unit Implementation Reports for years 1, 2, and 3 of program implementation are summarized in Table 3, where we show the percentage of reporting teachers meeting program benchmarks (e.g., distribute 5 or more Word of the Week Cards per week, play PA Music on 2 or more days per week). The percentages were averaged among all the seven program schools. Even though there was improvement in pro-

TABLE 3
Percent of Teachers Meeting Implementation Benchmarks on Unit Implementation by Year—Positive Action

Program Benchmarks	*2004–05*	*2005–06*	*2006–07*
Teaching at least 4 lessons per week	61	68	66%
Distribute 5+ Word of the Week Cards per week	22	35	35%
Distribute 5+ PA Stickers per week	30	39	41%
Read 5+ notes from ICU Box	42	48	49
Play PA Music 2+ days per week	19	35	33
Spoke with 2+ Parents about PA per week	25	45	43
Identified Academic Learning Standards in PA	81	93	91
Attended a PA assembly each unit	17	47	47
Teacher believes s/he delivered program quite well or very well	64	71	71
Teacher believes continued use of PA is very or extremely likely to improve student character	61	63	68
Teacher believes continued use of PA is very or extremely likely to improve student academics	49	53	58

gram implementation over time, the growth was slow. After 3 years of implementation, most of the general implementation benchmarks (lessons taught, identified academic learning standards in PA, teacher believes s/he delivered program quite well or very well, teacher believes continued use of PA is very or extremely likely to improve student character, teacher believes continued use of PA is very or extremely likely to improve student academics) were met by more than 50% of the teachers on average. Nevertheless, there was considerable variability among the program schools in terms of program implementation. By the end of year 3, one school was still implementing at a low level (meeting program benchmarks, on average, across all benchmarks below 50%), four at a moderate level (meeting program benchmarks at levels between 50 and 60%), and two at a moderate to high levels (meeting program benchmarks at a level between 60 and 70%). Nevertheless, by the end of Grade 5, children in our research cohort were exposed to the PA Program for 3 consecutive years, both directly through lessons delivered by their classroom teachers, and indirectly through exposure to school-wide program activities and other students across all grades who were also exposed to PA.

Use of Implementation Data

We plan on producing one or more articles investigating the relationship between various measures of program implementation, and between levels/quality of implementation and program impact. For example, relevant to the former, we will conduct analyses to replicate our findings from our Hawaii trial of the same program (Beets et al., 2008). Relevant to the latter, we plan on conducting propensity score analyses of program effects that adjust for level and quality of program implementation. Program outcomes include various components of social and character development, as well as behavioral outcomes such as drug use and violent behaviors.

Lessons Learned

From this trial and another trial in Hawaii schools, we have learned that it takes much more time for many low-performing schools to fully adopt and implement a comprehensive program than may have in prior decades. Along with other comprehensive program developers and researchers (e.g., Fixsen, Naoom, Blase, Friedman, & Wallace, 2005), we believe that under current conditions (e.g., the heavy focused demands of No Child Left Behind [NCLB]), many underperforming schools need 5–7 years to fully adopt and implement a comprehensive program and see substantial benefits from it. In addition, the hypothesized benefits of the PA program may be attributed to both the increased use of SACD-like activities as well as the better organized delivery of the activities. However, the use of SACD-like activities by comparison schools will limit the size of the program effects that can be observed.

THE ACADEMIC AND BEHAVIORAL COMPETENCIES (ABC) PROGRAM

The ABC Program, a school-wide program to reduce classroom disruption and encourage rule following and social competencies includes a comprehensive package of teacher training, mentoring, and direct service strategies. The ABC Program involves teacher training in behavioral strategies (time out, contingency management, implementation of rewards, social reinforcement), mentoring and ongoing consultation in implementation of program strategies and related skills, and direct services to students, including an after-school social skills program, individualized programming for more severe children, and social competencies instruction and reinforcement (see Flay et al., this issue, for a full description).

Measurement of Fidelity

In developing the model for measurement of implementation fidelity, we defined fidelity as a measure of how well a program is implemented with respect to two key constructs: implementation content, and implementation process. For implementation *content*, we attempted to assess the extent to which teachers and schools implemented the intended components of the program. For implementation *process,* we assessed the extent to which teachers and schools implemented the program *in the manner in which it was intended.* Both implementation process and implementation content indicators for each program component were developed to key into specific characteristics of each component. For several components, more than one indicator was collected to reflect the full picture of implementation. Indicators included archival information, ratings collected on an ongoing basis from program staff, and observations. The information about fidelity was continually monitored and at times fed back to consultants to improve implementation quality.

Baseline surveys were collected in the fall of the first year of implementation, and teachers were asked to report their practices during the year prior to the current year. Implementation surveys were then collected at the end of years 1, 2, and 3 of implementation (approximately 190 teachers per year). Teachers were asked a range of questions regarding their use of a variety of classroom behavioral management strategies over the past school year. For a list of common strategies (such as a daily home-school note, weekly rewards for appropriate behavior), teachers were asked whether they had used each strategy on a 1 (“Don’t use”) to 4 (“Use most/all of the time”) scale. For the same list of strategies, teachers were also asked to report whether each strategy was effective on a 1 (“Not effective”) to 4 (“Very effective”) scale.

Classroom observations were conducted with all intervention and comparison teachers in years 2 and 3 of implementation. Observations were conducted between February and May of the school year in all kindergarten through fifth grade classrooms. Observations were scheduled for one hour during an instructional activity. Prior to the observations, teachers completed Rules and Procedures forms that asked them to specify the expectations and routines in their classrooms. For example, teachers were asked to note whether children were expected to ask permission before getting out of their seats to sharpen a pencil or to retrieve materials from their desks. This information was taken into account during the observation procedures. Disruptive behaviors in the classroom were coded using a modified observational code. During the 40-minute active observation period, observers coded student behaviors that met specific operationalized criteria according to the coding manual. Teacher behaviors coded during the observations included frequency of teacher use of social reinforcement and frequency of teacher commands. In addition to the frequency categories, teacher responses to students' disruptive behaviors were coded; coding categories included not observed by teacher, not acknowledged, appropriate or inappropriate acknowledgment, and appropriate or inappropriate consequence greater detail on observational codes is available.

Following the observation, observers completed a Post-Observation Rating form that evaluated the classroom environment with respect to tone and style of the teacher, and her use of social reinforcement, commands, and her responses to disruptive behavior. The five questions on the Post-Observation Rating were scored on a Likert scale from 1 to 7 and included: (1) overall effectiveness of the teacher's use of social reinforcement; (2) overall effectiveness of the teacher's use of commands; (3) overall effectiveness of the teacher's use of behavioral management strategies ; 4) teacher's tone of voice (1 = "Harsh," 4 = "Neutral," and 7 = "Pleasant"); and 5) overall classroom climate (1 = "Very negative," 4 = "Neutral," and 7 = "Very positive").

Staff implementation ratings were completed monthly only for intervention teachers, as these ratings were completed based on consultation meetings with teachers, which only took place in intervention schools. Staff implementation ratings included monthly Likert ratings on 14 items related to specific intervention components and weekly tracking of teacher use of components (such as Daily Notes and Fun Friday activities).

Success of Implementation Measures

The fidelity model for the ABC Program identified intervention components provided to: (1) teachers; (2) students by teachers; (3) parents; and (4) students by schools. Table 4 provides information about the fidelity content and process indicators for each program component, including some information about completion rates.

Sample and Participants

Fidelity data were collected for all 3 years of implementation for all teachers in grades kindergarten through Grade 5 in 7 intervention and 7 comparison schools. While data were collected from special education teachers, they are excluded from the majority of fidelity analyses, as their data will be examined separately. In addition to teacher data collected, the study followed the cohorts of students who were first and third graders when treatment schools began implementing the program in the 2004–05 academic year. Five assessments were conducted during the study period: Fall 2004 (baseline assessment), Spring 2005, Fall 2005, Spring 2006, and Spring 2007 (end of Grade 5).

Implementation data discussed in this article were collected from all teachers in the study schools who returned signed consent forms ($N \simeq 250$), observers who conducted classroom observations, and behavioral consultants assigned to intervention schools.

Implementation Results

While teacher self-report ratings were also collected *annually*, we had concerns about the validity of teacher self-reports as measures of implementation fidelity. For example, preliminary analyses for the teacher self-report surveys and classroom observations found correlations that ranged from .02-.10 (Massetti, Pelham, & Waschbusch, 2007). Therefore, we did not rely on self-report ratings to assess implementation. Preliminary analyses of implementation data suggest that teachers use classroom management strategies at high rates across both intervention and comparison schools, with over 85% of teachers in both intervention and comparison schools reporting using such strategies. For intervention teach-

TABLE 4
Instruments and Implementation Measurement Used to Assess Implementation Fidelity—
Academic and Behavioral Competencies Program

Services Provided	*Program Component*	*Content Measure*	*Process Measure*	*Completion Rate for Measures*
To teachers	Teacher training in classroom management	Archival; training attendance records	Archival; supervisor ratings for content and delivery	100%, Intervention schools only
	Teacher consultation	Archival; session completion logs	Archival; monthly consultant ratings	100%, Intervention only
To students by teachers	Classroom management program: response/ cost system	Archival; daily teacher logs (collected weekly for minimum of 4 weeks) and annual consultant ratings	Classroom observations	Process: 53%-83% (mean = 68%), intervention only; Content: 100% (2−5 observations per teacher), both intervention and comparison
	Classroom management program: Daily Notes, Fun Friday activities	Archival; daily and weekly teacher logs	Archival; monthly consultant ratings	Process: same as response/ cost system; Content: 100%, intervention only
	Classroom management program: Social reinforcement, student-teacher interactions	Classroom observations	Postobservation ratings	100% both intervention and comparison
	Classroom social skills program	Archival; teacher social skills logs, annual consultant ratings	Archival; annual consultant ratings	100%, intervention only
	Individualized programs	Archival; monthly and annual consultant ratings	Archival; monthly and annual consultant ratings	100%, intervention only
To parents	Parent workshops	Archival; records of workshops and materials	Archival; supervisor ratings	100%, intervention only
To students by schools	Recreational skills program	Archival; attendance and program records	Archival; supervisor ratings	100%, intervention only

Note: In addition to observations and archival records, questions about each teacher-specific component were also included in annual teacher surveys. The "completion" column above denotes the number of measures completed or collected successfully.

ers, there was some variability in implementation across program components. For example, 98% of teachers implemented rule tracking systems; 91% used a Fun Friday activity weekly throughout the year; 63% used a daily note or stamp procedure; 72% had individual programs or daily report cards for students in the classrooms. Program components with low implementation included the daily social skills program (25% of teachers) and posted behavior charts (50%). Additionally, while implementation content was high for certain components of the program, the implementation process or quality of delivery of those components was highly variable across teachers. For example, staff monthly implementation ratings ranged from 0% to 100% within the same school for some months; observations scores likewise were highly variable across teachers on all observations variables, including frequency of social reinforcement, quality of classroom management, and use of appropriate consequences in response to student disruptive behavior. These findings suggest that when implementing a program teachers may be willing to engage in a particular activity, but the quality of engagement will not be guaranteed. Observation data suggest high variability across teachers in implementation of the interactive components of the program, such as positive reinforcement (which ranged from 0 instances per 40-minute observation to more than 50) and commands (which ranged from 2 instances to 60). Furthermore, there were significant associations found among the interactive components, such that teachers who used more positive reinforcement also used more appropriate commands and provided more neutral, appropriate feedback to students regarding disruptive behavior.

Planned Use of Implementation Data

Implementation data will be used to describe the process of implementation of the program over time, as classroom observations and monthly ratings can provide sensitive assessments of the implementation strategy within schools and across schools. For example, questions regarding which aspects of implementation content were more easily implemented by teachers, and the variation in content over time can be addressed. Furthermore, changes in implementation process, or the quality of implementation across classrooms and across time will be explored. Predictors of implementation process and content, at both the teacher and classroom level, can be examined, with a particular focus on the relationship between quality of implementation of behavioral classroom management strategies (as measured in observations) and levels of disruptive behavior in the classroom in both observations and teacher-report surveys. Finally, the relationship between implementation content and process can be explored, to address ways to best define the concept of implementation of behavioral classroom management strategies, and which aspects of implementation are most closely linked with decreases in students' disruptive classroom behavior.

Lessons Learned

While the fidelity model provided comprehensive information regarding implementation within and across intervention schools (with some data also collected in comparison schools), data collection required a substantial allocation of resources. This was relevant once the poor relationship between teacher self-report measures and observational measures were examined, as greater emphasis was then placed on more costly and demanding observational strategies and archival records. Decisions regarding the plan for fidelity data collections had to be made prior to implementation, as plans for archival data collection and ratings forms had to be developed and piloted. Furthermore, ABC Program staff assigned to each school had to allocate time to completing ratings for each teacher on a monthly (or more frequent) basis, and most importantly archival records had to be collected from teachers daily or weekly. Low return rates for weekly teacher logs were addressed by providing incentives to teachers in the form of lotteries and gift cards. Personal relationships between ABC Program consultants and teachers were an important mechanism for collecting archival records, as well.

LOVE IN A BIG WORLD (LBW)

LBW is character education program that uses a school-wide music assembly, morning announcements, and a teacher delivered curriculum to promote positive student development. Over the course of the school-year the program covers a different character trait each week with several activities such as reading a story, group discussions, and journaling. The program was developed and implemented by a group independent of the evaluation conducted by Vanderbilt University and led by Leonard Bickman and Manuel Riemer. There was no previous or ongoing association between the evaluators and program developers/implementers.

Measurement of Fidelity

In working with the developers of LBW we had to first develop an evaluable program theory with them before we could create implementation measures. There was no existing logic model and there were no fidelity measures. The program developers had no information by which to judge the quality of implementation. For example, there was no standard by which to judge whether a teacher is presenting the curriculum lesson with competence. Thus, we limited our assessment in this local evaluation to the *amount* of implementation fidelity. In this article we will focus on program adherence. Treatment differentiation (treatment vs. control group) is presented elsewhere (Riemer & Bickman, 2008).

As described in an earlier article in this special issue (Flay et al., this volume) LBW has several program components that are implemented at the school level and a curriculum that is administered by a teacher at the classroom level. Some of the school-level activities are delivered by LBW program staff while others are executed by school staff. The teacher training, the assembly, the morning announcements, and classroom curriculum are considered essential while the other program components are there to supplement and support these essential components. The benchmarks for each program component are listed in Table 5.

As can be seen in Table 5 we used a combination of logs and checklists to assess adherence to the different program components. At the core of our assessment of implementation fidelity is the daily activity checklist completed by teachers. On this checklist teachers reported for each school day whether or not they did one or more character education (CE) activities that day. For each CE activity they reported what the specific activity was, whether it was an activity from the curriculum, and what character trait was targeted with the activity. The teachers also reported how much time they spent on the activity and their

TABLE 5
Benchmarks and Measures by Program Component—Love in a Big World

Program component	*Implemented by*	*Essential*	*Benchmark*	*Measure*
Teacher training	Program staff	Yes	90% of all teachers must attend trainings in all 3 years	Attendance log
Assembly	Program staff	Yes	One assembly per school per year	Report by program staff
Follow-up calls	Program staff	No	Monthly 15-20 minute call to program coordinator	Log by program staff
Kids club	Program staff	No	none	none
Family newsletter	Program staff	No	none	none
Morning announcements	School	Yes	At least 22 of the available 24 announcements have been played	Checklist completed by school-based program coordinator (year 1 only)
Faculty boosters	School	No	At least 22 of the available 24 weekly faculty boosters were provided to the teachers	Checklist completed by school-based program coordinator (year 1 only)
Service project	School	No	One project per year	none
Curriculum	Teacher	Yes	A teacher must deliver at least 4 curriculum-based activities for the corresponding character trait each week. The activities should last at least 10 minutes.	Daily activity checklist

rating of effectiveness. A teacher was expected to deliver at least four curriculum-based activities for the corresponding character trait each week. Each of these activities should last 10 to 15 minutes. During the first year teachers completed the checklist on the computer using an online data entry system. As an ancillary study half of the teachers in the program schools received weekly feedback reports regarding their implementation fidelity. Because of technical problems with the schools' internet connections and the problems the teachers had in completing the checklist online, paper versions were used in year 2 and 3. The completed forms were faxed weekly to local evaluation team. Over the course of the evaluation we had to make slight changes to the daily activity checklist because the program was modified after the first year. Teachers in the control schools completed a similar weekly log.

Sample and Participants

The LBW program was tested in 13 elementary schools in two Public Schools systems in medium-sized cities in Tennessee during the 2004–05 through 2006–07 academic years. In the first year there were 6 schools with a total of 30 classrooms and teachers in the treatment condition and 6 schools with a total of 23 classrooms in the control condition. In the second year another school was added resulting in 7 schools and a total of 29 classrooms while there were 21 classrooms in the control school. In the third year the respective numbers of classroom were 28 and 20. SACD activity and program data were collected from both the intervention and the control school teachers.

Success of Implementation Measurement

To ensure compliance with the completion of the measures needed for the research and the evaluation including the fidelity logs and checklists we clearly differentiated the research from the program activities. In an initial research training that was provided separately from the program training, teachers received information regarding the importance of providing information regarding implementation even if they did not like or use the program. They were also instructed in how to complete the daily activity checklist. Teachers also received payments for both participating in the research in general as well as for each completed daily activity checklist. Similarly, the program coordinator received payment for participating in the research and providing information on implementation on a regular basis.

The completion rates differ by implementation measure as can be seen in Table 6. Those program components delivered by the program staff at the beginning of each school year (the teacher training and the assembly) were reported 100% of the time. However, the logging of the follow-up calls turned out to be unfeasible for the program staff and was stopped after the first year. The program components delivered by the school were reported for the majority of weeks with a rate of 83% for the morning announcements and 77% for the faculty boosters. This was assessed only in Year 1. The compliance with completing the daily activity checklist was poor with only 51% in the first year but improved in year 2 and 3 (73% and 76% respectively).

Implementation Results

As Table 6 shows the benchmarks for two of the program components delivered by the program staff were met in the first year, but then attendance in the teacher training decreased. All teachers were trained in the first year, while in year 2 it dropped to 83%, and 54% in year 3. Each intervention school received at least one assembly per year. However, the absence of logs from the program staff regarding the follow-up calls suggests that these either did not happen at all or only sporadically. There is not enough data, however, to determine this conclusively.

TABLE 6
Completion Rates and Average Implementation Results—Love in a Big World

Program Component	*Benchmark*	*Measure*	*Completion Rate*	*Results (Average)*
Teacher training	90% of all teachers must attend trainings in all 3 years	Attendance log	Year 1: 100% Year 2: 100% Year 3: 100%	Year 1: 100% Year 2: 83% Year 3: 54%
Assembly	One assembly per school per year	Report by program staff	Year 1: 100% Year 2: 100% Year 3: 100%	One assembly per school year
Follow-up calls	Monthly 15-20 minute call to program coordinator	Log by program staff	Incomplete and only during first year	Not determinable
Morning announcements	At least 22 of the available 24 announcements have been played	Checklist completed by school-based program coordinator (year 1 only)	Year 1: 83%	Year 1: 20 announcements Std Dev =1.67
Faculty boosters	At least 22 of the available 24 weekly faculty boosters were provided to the teachers	Checklist completed by school-based program coordinator (year 1 only)	Year 1: 77%	Year 1: 18.5 faculty boosters Std Dev =3.02
Curriculum	A teacher must deliver at least 4 curriculum-based activities for the corresponding character trait each week. The activities should last at least 10 minutes.	Daily activity checklist (DAC)	Year 1: 51% Year 2: 73% Year 3: 76%	Year 1: 12.4 Wks, SD 7.8* == 41.1 % out of total school weeks Year 2: 12.3 Wks, SD 9.9 == 34.9% out of total school weeks Year 3: 10.4 Wks, SD 9.06 == 28.7 % out of total school weeks

* The format of the Daily Activity Checklist was changed after year 1. Thus, the comparability of the activity level of year 1 and year 2 and 3 is limited.

The implementation results for the program elements delivered by the schools during year 1 indicate that schools were close to meeting benchmarks. Schools delivered between 71% and 93% of the morning announcements with an average of 20 announcements played (*SD* = 1.67). The benchmark was 22 announcements. The implementation of the faculty boosters were a little bit lower with only 18.5 (*SD* = 3.02) reported instances of providing the materials to teachers. The range was also greater with the lowest performing school providing 54% of the boosters compared to the 88% of the top school.

The expectations for the implementation of the curriculum were more complex. Two criteria were used: the number of activities per week and the average time spent on each activity. Table 7 shows that in most schools the benchmark of four activities per week (independent of the trait targeted) were reportedly met in less than half of the weeks. It is especially noticeable that in year 3 five out seven schools met the benchmark in less than a third of the weeks. The lowest implementation level is 12% of the weeks while the highest with about 64% of the weeks is clearly unsatisfactory.

TABLE 7
Percentage of Effective Weeks* Per School Per Year—
Love in a Big World

School ID	*Year 1*	*Year 2*	*Year 3*
3	32	41	31
4	56	34	14
6	57	52	12
7	20	64	30
8	22	17	46
12	59	28	41
14	.	14	24

*Effective Week is a week with 4 activities or more.

While the teachers do not do as many of the activities as expected, it appears as if they do spend the expected amount of time on each activity. As can be seen in Table 8 the median time spent on the character education activity is about 15 minutes as reported by the teachers. However, there is also quite a large variability and wide range in regard to the time spent on each activity. This is especially true in year 2 and 3. In retrospect one may ask whether the benchmark of at least four activities per week was unrealistic. In addition, there may be certain types of activities that are more effective than others. Thus, missing those would lower the impact of the program more than others. Without a strong theory and empirical evidence about the effectiveness of specific components of a program in addition to the needed dosage, the designation of benchmarks is arbitrary.

Planned Use of Implementation Data

The Vanderbilt team is planning to use the implementation data for two major purposes. First, we will judge whether a summative judgment about the effectiveness program based on the comparison of the experimental and control groups can be made. Because we collected very similar data from the control and treatment schools we are able to compare the level of CE activities at baseline as well as over the three years of implementation. If there is not much differentiation then a report of the comparison between the two groups on outcomes would be misleading and inappropriate. The results would be inconclusive because of implementation of activities were equivalent at both treatment and control schools. Second, we are planning to use the implementation data to learn more about the implementation process itself. We are especially interested in studying the relationship between teacher-level variables such as self-efficacy, attitudes, and motivation on implementation. We are also testing a specific intervention (providing implementation feedback to teachers) intended to improve implementation fidelity.

Lessons Learned

Implementation is a complex process that is difficult to assess with rigorous scientific methods. This is especially true if a program is not well defined and the target itself is elusive such as a young child's character. Nevertheless, it was clear in the case of this evaluation how important it is to attempt such measurement. The key ingredient of the LBW program according to its developers is the set of character education activities delivered by the teachers based on the LBW curriculum. Other program components such as the teacher training and the faculty boosters are there primarily to support this core element. But, it is this core element that was implemented way below

TABLE 8
Average Number of Minutes Spent Per Activity Per School—Love in a Big World

School ID	*Year 1*					*Year 2*					*Year 3*				
	Mean	*Median*	*Std*	*Min*	*Max*	*Mean*	*Median*	*Std*	*Min*	*Max*	*Mean*	*Median*	*Std*	*Min*	*Max*
3	16.08	15	6.08	4	35	16.85	15	7.96	0	45	30.34	30	19.49	0	120
4	13.81	15	6.86	2	100	19.39	15	14.07	5	180	12.08	10	10.86	0	60
6	17.72	15	6.56	5	45	14.11	15	7.06	0	90	17.08	15	8.3	2	40
7	14.87	15	8.33	5	60	13.94	15	5	5	30	19.24	15	13.47	5	90
8	11.59	10	8.4	1	115	22.01	20	14.74	0	95	12.44	10	7.23	0	65
12	14.99	12	6.46	5	40	13.79	15	4.56	1	30	18.72	15	16.25	0	90
14	.	.	.	.	.	15.4	15	6.62	5	30	15.64	15	10.68	0	45

benchmarks in most schools. Thus, there should be little reason to expect any differences in outcomes in the implementation and control schools. However, the fact that a program could not be implemented as intended also suggests that there is a mismatch between the program characteristics and the context within which it is being implemented. This in itself has important policy implication.

DISCUSSION

Accountability for Implementation Evaluation is Critical

The SACD consortium seriously attempted to measure implementation at each site. Each group creatively designed measurement systems in an attempt to capture the fidelity of the programs they were studying. However, comparisons at the multiprogram level were much more difficult to make. Given the wide variations in program content, processes and structure, it was not feasible to produce an approach that would apply to each site and yet generalize across sites. There was no key theoretical construct that tied all the programs together. The absence of any prior empirical evidence on how to measure quality and intensity of implementation and to determine which of the many activities were critical to determining outcomes was a major handicap that all evaluations faced. Thus, the measurement of fidelity and implementation quality in the present SACD multisite evaluation provides a set of clear examples of the need for measurement of implementation to be idiographic to the structure and content of the specific program evaluated.

SMOKE AND MIRRORS

Implementation measurement has been neglected until recently. But even with the increased sensitivity, most attention and resources are directed at measuring outcomes and, to lesser extent, mediators and moderators. This results in not only inadequate measurement but also measurement of questionable validity. The ABC investigators reported that they found practically no relationship between what teachers reported and observation of their behavior. In order for observation at the individual level to be reliable many such observations are required with concomitant high costs. However, if self-report is not a valid measure of implementation then we must direct our resources to more direct measures. One major limitation of the ABC investigators' findings is that their self report was done once a year. In contrast Vanderbilt's was done weekly. Whether higher frequency just produces more invalid data or data that are more valid remains to be seen. It should be noted that dependence on self report was not unique to measuring implementation. Almost all the measures used for outcomes, mediators and moderators at all sites were self report.

THE PARADOXICAL EFFECT OF INTENSE IMPLEMENTATION MEASUREMENT

The more documentation of implementation that is required, the more difficult implementation becomes. Most implementation data are provided by the persons implementing the intervention. Greater demands placed on implementers to document what they are doing leads to greater likelihood of problems with implementation, as resources may be redirected from implementation to measurement of implementation. Documentation takes up time and is typically not as valuable to the person implementing the program as it is for researchers. Unless the new program reduces the workload (highly unlikely) or there is sufficient payment for documentation then, as shown in many of the tables in this article, implementation documentation will be spotty. Finally, accountability for whether the program "works" is a problem for the developers. The implementers, however, are accountable for the quality of implementation. Accountability is typically resisted, thus teachers may not be enthusiastic about collecting data that reflects on their performance.

WHAT HAPPENS WHEN THE TREATMENT DOESN'T STAY IN THE TREATMENT GROUP?

In the real world we typically cannot isolate the treatment from the control group. Some of the authors noted that they studied SACD-like activities in the control sites as well as the treatment sites. Such measurement should be a requirement of every evaluation. Not only is there spread of program activities through reading, meetings, etc. but in many cases SACD activities may become required by state law, as in the case of Tennessee, or because social and character development has become popular among educators. The intervention cannot be adequately tested if the treatments—in this case the SACD activities—occur with similar frequency and quality in both the treatment and control groups. These local history artifacts threaten the internal and construct validity of the design.

FORMATIVE OR SUMMATIVE?

One group of researchers noted that they provided regular feedback on implementation to program personnel. Other groups felt that it would detract from the realistic testing of the intervention if such feedback was part of the evaluation and not a typical part of the program. In some of the sites the program developers and the evaluators were one and the same thus enhancing the possibility that feedback would occur. There was no consensus within the SACD consortium on a uniform approach to whether formative feedback should be provided as part of the evaluator's responsibility. A related issue is the potential problem caused when the evaluators were the program developers or intimately involved in the implementation of the program. In some sites the evaluators were totally independent of the developers and implementers, while in others they were the same individuals. One may expect that implementation fidelity could be affected by this factor. Moreover, it is possible that in these situations an" allegiance effect" may operate in which better outcomes occur when the evaluations are conducted by the program developers.

SIMILARITIES AND DIFFERENCES AMONG PROGRAMS AND SITES

An earlier article in this special volume (Flay et al.) describes in more detail the similarities and differences among the programs. It is evident that, although the age groups, common measurement core, target behaviors (social and emotional competencies, prosocial and problem behavior, school climate, and academic achievement) and the emphasis on whole school interventions were similar for all sites, there were still significant differences in program elements. Some programs were highly focused on one problem area and others were very broad. Some had intensive teacher training and others had minimal training. The programs had different benchmarks, program theories and resources. There were also important differences in regard to the study sites. For example, some were urban and others were suburban or semirural. Concomitant with location were differences in race, culture and SES of the children (see Massetti et al., this volume, for examples). These program and site differences had implications not only for interpreting the outcomes but as illustrated in the data presented by the four programs how implementation was studied.

TO STUDY IMPLEMENTATION YOU NEED TO IMPLEMENT

The planning year provided in the grant mechanism was very helpful but we could not anticipate the major problems in implementing a complex program in a live environment. Regardless of careful and detailed preparation, it can be expected that implementation will take longer than expected, change will be resisted more than anticipated and what seemed like a really great idea will not be appreciated universally. Moreover, it is likely that stronger forces will emerge, such as NCLB, that will reduce the importance and time teachers can allocate to the intervention. Given the inevitability of the unpredictable, it is wise to learn from these events by including a careful study of implementation in evaluating programs.

IMPLEMENTATION FAILURE IS COMMON—GET USED TO IT

Implementation failure is the mode and our expectations may just be too high. Edison tried 10,000 filaments before he found one that worked. However, perspective is important. He is credited with saying "I have not failed. I've just found 10,000 ways that won't work." Learning from our putative failures is critical to the development of effective real world interventions.

AUTHOR NOTE

The findings reported here are based on research conducted as part of the Social and Character Development (SACD) Research Program funded by the Institute of Education Sciences (IES), U.S. Department of Education, under contract ED-01-CO- 0039/0006 to Mathematica Policy Research (MPR), Princeton, NJ, in collaboration with the Centers for Disease Control and Prevention's Division of Violence Prevention (DVP), and the recipients of SACD cooperative agreements. The SACD Consortium consists of representatives from IES, DVP, and the national evaluation contractor (MPR), and each cooperative agreement site participating in the evaluation. Research institutions in the SACD program (and principal researchers) include: *IES* Amy Silverman, Edward Metz, Elizabeth Albro, Caroline Ebanks; *DVP* Tamara M. Haegerich (previously, IES), Corinne David-Ferdon, Le'Roy Reese (Moorehouse School of Medicine; previously DVP); *MPR* Karen Needels, John A. Burghardt, Heather Koball, Laura M. Kalb, Peter Z. Schochet, Victor Battistich (University of Missouri—St. Louis); *Children's Institute* Deborah B. Johnson, Hugh F. Crean; *New York University* J. Lawrence Aber, Stephanie M. Jones (Harvard University), Joshua L. Brown (Fordham University); *University at Buffalo, The State University of New York* William Pelham, Greta M. Massetti (CDC), Daniel A. Waschbusch; *Oregon State University* Brian R. Flay, Carol G. Allred (Positive Action), David L. DuBois (University of Illinois at Chicago), Michael L. Berbaum (University of Illinois at Chicago), Peter Ji (University of Illinois at Chicago), Vanessa Brechling, (University of Illinois at Chicago); *University of Maryland* Gary D. Gottfredson, Elise T. Pas, Allison Nebbergall; *University of North Carolina at Chapel Hill* Mark W. Fraser, Thomas W. Farmer (Penn State University), Maeda J. Galinsky, Kimberly Dadisman; and *Vanderbilt University* Leonard Bickman, Catherine Smith.

Notice of potential conflict of interest: The research described herein was done using the program and the training and technical support of Positive Action, Inc. Dr. Flay's spouse holds a significant financial interest in Positive Action, Inc.

Acknowledgment: The authors wish to thank the Consortium reviewers who commented on earlier drafts of this article. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Institute of Education Sciences, Centers for Disease Control and Prevention, Mathematica Policy Research, Inc., or every Consortium member, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Correspondence concerning this article should be addressed to Leonard.Bickman@Vanderbilt.edu

REFERENCES

Adelman, H. S., & Taylor, L. (2000). Moving prevention from the fringes into the fabric of school improvement. *Journal of Educational and Psychological Consultation, 11*, 7–36.

Beets, M. W., Flay, B. R., Vuchinich, S., Acock, A.C., Li, K. K., & Allred, C. (2008). School climate and teachers' beliefs and attitudes associated with implementation of the Positive Action program: A diffusion of innovations model. *Prevention Science.* DOI: 10.1007/s11121-008-0100-2.

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

Bickman, L. (1987). The functions of program theory. In L. Bickman (Ed.), *Using program theory in evaluation. New Directions for Program Evaluation, No 33* (pp. 5-17). San Francisco: Jossey-Bass.

Bickman, L. (2000). Summing up program theory. In P. Rogers, T. Hacsi, A. Petrosino, & T. Huebner (Eds.), *New Directions for Evaluation, 87,* 103–112.

Bickman, L., & Athay, M. M. (2009). The worst of all possible program evaluation outcomes. In A. R. Stiffman (Ed.), *The field research survival guide*. New York: Oxford University.

Bickman, L., & Peterson, K. A. (1990). Using program theory to describe and measure program quality. In Bickman, L. (Ed.), *Advances in program theory. New directions for program evaluation, No. 47* (pp. 61–72). San Francisco, CA: Jossey-Bass.

Bickman, L., & Salzer, M. S. (1997). Measuring quality in mental health services. *Evaluation Review, 21*, 285–291.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research, 2,* 75–87.

Chambers, D. (2008). Advancing the science of implementation: A workshop summary. *Administration and Policy in Mental Health and Mental Health Services Research 35,* 3–10.

Chen, H. T. (2005). *Practical program evaluation: Assess and improve program planning, implementation, and effectiveness.* Thousand Oaks, CA: SAGE.

Clapp, J. D., Burke, A. C., & Stanger, L. (1998). The institutional environment, strategic response and program adaptation: A case study. *Journal of Applied Social Sciences, 22*, 87–95.

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.

Domitrovich, C., & Greenberg, M. (2000). The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children. *Journal of Educational and Psychological Consultation, 11,* 193–221.

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.

Elliott, D. S., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science*, *5*, 47–53.

Fagen, M. C. (2003). Practice notes: Strategies in health education. The Aban Aya Sustainability Project. *Health Education & Behavior, 30*, 641–643.

Fagen, M. C., & Flay, B. R. (2006 December 15). Sustaining a school-based prevention program: Results from the Aban Aya Sustainability Project. *Health Education & Behavior, 20*, 1–15.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* Tampa: National Implementation Research Network, Louis de la Parte Florida Mental Health Institute, University of South Florida. FMHI publication 231. Retrieved March 5, 2009, deom http://www.fpg.unc.edu/~nirn/resources/detail.cfm?resourceID=31

Flay B. R., & Allred, C. G. (2003). Long-term Effects of the Positive Action Program. *American Journal of Health Behavior, 27,* S6.

Flay, B. R., Allred, C. G., & Ordway, N (2001). Effects of the Positive Action program on achievement and discipline: Two matched-control comparisons. *Prevention Science, 2(2),* 71–89.

Flay, B. R., & Petraitis, J (1994). The theory of triadic influence: A new theory of health behavior with implications for preventive interventions. In G. S. Albrecht (Ed). *Advances in Medical Sociology, Vol IV: A Reconsideration of models of health behavior change* (pp 19–44). Greenwich, CN: JAI Press.

Foster, M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care, 41*, 1183–1192.

Greenberg, M. T., Domitrovich, C., & Bumbarger, B. (2001). The prevention of mental disorders in school-aged children: Current state of the field. *Prevention & Treatment, 4,* Article 1.

Griffith, J. (2000). School climate as group evaluation and group consensus: Student and parent perceptions of the elementary school environment. *Elementary School Journal, 101*, 35–61.

Hallfors, D., & Godette, D. (2002). Will the "Principles of Effectiveness" improve prevention practice? Early findings from a diffusion study. *Health Education Research, 17*, 461–470.

Hill, J., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birthweight premature infants. *Developmental Psychology, 39*, 730–744.

Hitt, J. C., Robbins, A. S., Galbraith, J. S., Todd, J. D., Patel-Larson, A., McFarlane, J. R., et al. (2006). Adaptation and implementation of an evidence-based prevention counseling intervention in Texas. *AIDS Education and Prevention*, *18 (suppl A)*, 108–118.

Hohmann, A. A., & Shear, M. K. (2002). Community-based intervention research: Coping with the "noise" of real life in study design. *American Journal of Psychiatry, 159,* 201–207.

Inouye, S. K., Baker, D. I., Fugal, P., & Bradley, E. H. (2006). Dissemination of the Hospital Elder Life Program: Implementation, adaptation and success. *Journal of the American Geriatric Society*, *54*, 1492–1499.

Kam, C., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science*, *4,* 55–63.

Kuperminc, G. P., Leadbetter, B. J., Emmons, C., & Blatt, S. J. (1997). Perceived school climate and difficulties in the social adjustment of middle school students. *Applied Developmental Science, 1*, 76–88.

Lochman, J. E., Boxmeyer, C., Powell, N., Roth, D. L., & Windle, M. (2006). Masked intervention effects: Analytic methods for addressing low dosage of intervention. *New Directions for Evaluation*, *110*, 19–32.

Massetti, G. M., Pelham, W. E., & Waschbusch, D. A. (2007, June). *Teacher fidelity of use of behavior management strategies: Relationships among observations, self-report, and children's disruptive behavior.* Poster session presented at the annual meeting of the Institute of Education Sciences Research Conference, Washington, DC.

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.

Paulson, R. I., Post, R. L., Herincks, H. A., & Risser, P. (2002). Beyond components: Using fidelity scales to measure and assure choice in program implementation and quality assurance. *Community Mental Health Journal, 38,* 119–128.

Purkey W. W. (1970). *Self-concept and school achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Riemer, M. & Bickman, L. (in press). Using program theory to link social psychology and program evaluation. In M. M. Mark, S. I. Donaldson, & B. Campbell (Eds.), *Social psychology and program/policy evaluation*. New York: Guilford.

Riemer, M., & Bickman, L. (2008, June). *The importance of assessing treatment differentiation.* Poster session presented at annual meeting of the Institute of Education Sciences Research Conference, Washington, DC.

Riemer, M., Rosof-Williams, J., & Bickman, L. (2005). Theories related to changing clinician practice. *Child and Adolescent Psychiatric Clinics of North America*, *14*, 241–54.

Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (2000). School as a context of early adolescents' academic and social-emotional development: A summary of research findings. *Elementary School Journal, 100*, 433–471.

Rohrbach, L. A., Graham, J. W., & Hansen, W. B. (1993). Diffusion of a school-based substance abuse prevention program: Predictors of program implementation. *Preventive Medicine, 22,* 237–260.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55

Stevens, V., Bourdeaudhuij, I. D., & Oost, P.V. (2001). Anti-bullying interventions at school: Aspects of programme adaptation and critical issues for further programme development. *Health Promotion International*, *16,* 155–167.

Torill, L. & Oddrun, S. (2007). Implementing Second Step: Balancing fidelity and program adaptation. *Journal of Educational and Psychological Consultation*, 17, 1–29.

Waterman, H., Marshall, M., Noble, J., Davies, H., Walshe, K., Sheaff, R., et al. (2007). The role of action research in the investigation and diffusion of innovations in health care: The PRIDE project. *Qualitative Health Research*, *17*, 373–381.

Weiner, B. J., Helfrich, C. D., Savitz, L. A., & Swiger, K. D. (2007). Adoption and implementation of strategies for diabetes management in primary care practices. *American Journal of Preventive Medicine*, *33*, S35–S49.

Zinn, J. S., Mor, V., Feng, Z., & Intrator, O. (2007). Doing better to do good: The impact of strategic adaptation on nursing home performance. *Health Service Research*, *42*, 1200–1218.