Hammill Institute on Disabilities

Journal of Emotional and Behavioral Disorders
21(1) 66–78

DOI: 10.1177/1063426611429571
http://jebd.sagepub.com
SAGE

Assessing the Effectiveness of First Step to Success: Are Short-Term Results the First Step to Long-Term Behavioral Improvements?

W. Carl Sumi[1], Michelle W. Woodbridge[1], Harold S. Javitz[1], S. Patrick Thornton[1], Mary Wagner[1], Kristen Rouspil[1], Jennifer W. Yu[1], John R. Seeley[2], Hill M. Walker[3], Annemieke Golly[2], Jason W. Small[2], Edward G. Feil[2], and Herbert H. Severson[2]

Abstract

This article reports on the effectiveness of First Step to Success, a secondary-level intervention appropriate for students in early elementary school who experience moderate to severe behavior problems and are at risk for academic failure. The authors demonstrate the intervention's short-term effects on multiple behavioral and academic outcomes as delivered off-the-shelf in a variety of classrooms and schools across the country—effects that were mitigated by fidelity of implementation. Furthermore, the authors assess the body of evidence on First Step to Success against the standards for effectiveness and widespread dissemination promulgated by the Society for Prevention Research, thereby suggesting directions for further research.

Keywords

evidence-based interventions, externalizing behavior, elementary school, fidelity of implementation, effectiveness study

An increasing number of children in elementary schools have behavior issues that compromise their ability to meet the expectations of their teachers, such as self-regulation, academic focus, and positive interactions with peers and adults (Walker, Ramsey, & Gresham, 2004). These behavioral attributes, defined as academic enablers (DiPerna & Elliott, 2002), are fundamental to the development of social competence and effective learning and achievement (Gresham, Cook, Crews, & Kern, 2004). Children who begin their school careers with serious limitations in these competencies often fail to get off to a good start in school or to derive maximal benefits from their schooling.

With these children in mind, the School Mental Health Alliance, a group of more than 50 experts in school mental health research and policy, released *Working Together to Promote Academic Performance, Social and Emotional Learning, and Mental Health for All Children* (Hunter et al., 2005). This position paper provides an action plan for addressing the mental health needs of behaviorally at-risk children in school and mental health settings. A major recommendation of this group is to emphasize wider implementation of promising cost-effective interventions at the point of school entry and to forge collaborative partnerships among educators, parents, mental health experts, and community agencies.

One promising exemplar of this approach is First Step to Success (hereafter referred to as First Step), a school–home intervention that has a solid evidence base in achieving positive outcomes for behaviorally at-risk children in the primary grades (Walker et al., 1997, 1998). First Step is considered a secondary-level intervention (i.e., used when children do not respond to primary, schoolwide universal prevention strategies); it is appropriate for students who experience moderate to severe behavior problems early in their school careers and, thus, may be at risk for academic failure. First Step is a manualized program packaged for wide dissemination, and it is considered evidence-based, having been adopted and implemented with successful

[1]SRI International, Menlo Park, CA, USA
[2]Oregon Research Institute, Eugene, USA
[3]University of Oregon, Eugene, USA

Corresponding Author:
W. Carl Sumi, SRI International, 333 Ravenswood Avenue BS182, Menlo Park, CA 94025 USA
Email: carl.sumi@sri.com

results in a number of school districts across the country (see Beard & Sugai, 2004; Lien-Thorne & Kamps, 2005; Overton, McKenzie, King, & Osbourne, 2002; Sprague & Perkins, 2009; Walker et al., 2009).

First Step has three linked modular components—universal screening, classroom-based intervention, and in-home parent education known as homeBase—that are designed to be implemented in concert. Throughout the duration of First Step implementation, which generally takes about 3 months, these program components are modeled in the classroom and delivered in the home by a behavior coach. Coaches are typically drawn from the ranks of school psychologists, counselors, behavioral specialists, and resource teachers, but paraprofessionals with proper training can fill this role. During implementation, the coach works closely with participating teachers and parents/caregivers, providing them with ongoing technical assistance (TA) and support.

The coach's main focus is to provide teachers and parents with skills to teach students replacement behaviors and to distribute rewards when those behaviors are used appropriately and consistently. During initial implementation of the classroom-based intervention, the coach gives the student visual cues (i.e., a green or red card) to indicate whether he or she is on task and using appropriate behaviors, and the student accrues points toward a behavioral goal. If the student earns the daily goal, he or she is allowed to choose an enjoyable activity for the whole class. On about the 6th day of the program, the teacher assumes control over the activities, with supervision and support from the coach. The teacher's role also includes providing parents with daily feedback about the student's progress. Parents, in turn, are encouraged to reward the student's positive behavior with an activity at home, such as playing a game or taking a walk together. In addition, parents participate in the homeBase component of First Step, during which the coach meets with the parents weekly for 6 weeks to conduct lessons designed to strengthen parenting skills and to encourage a collaborative home and school working relationship whose focus is on joint problem solving and the development of school success.

First Step is well grounded in a social-ecological model (Bronfenbrenner, 1979; Schalock, 1989), which conceptualizes the individual as embedded within a system, organization, and setting. An appealing feature of the social ecology framework, mirrored in the key features of First Step, is that solutions to behavioral problems can be achieved by teaching individuals (i.e., students, teachers, and parents) new behaviors and ways to respond to behaviors, and by altering the environment (i.e., home, classroom) to promote positive behavior.

To date, evaluations of First Step have been small in scale, although a recent randomized controlled trial (RCT) supported the efficacy of First Step with a diverse population of students in a large urban school district (Walker et al., 2009). Although the researchers designed and conducted this study with rigor, they also exerted considerable control over the conditions in which the intervention was delivered. For example, the level of intensity of the training and the coaches' consistency of follow-up and supervision were optimal—not typical—conditions. Although efficacy studies are a critical step toward widespread dissemination of evidence-based programs (Flay et al., 2005), such studies do not consider the intervention's effects in real-world situations (Clarke et al., 1995; Flay et al., 2005; Hoagwood, Hibbs, Brent, & Jensen, 1995). When interventions move from a controlled experimental environment to naturalistic settings, effects may diminish markedly (Payne, 2008; Scott, 2001; Supovitz & Weinbaum, 2008; Weisz & Jensen, 2001).

An effectiveness study that demonstrates how to transport and integrate an intervention into schools' routine practices and achieve desired outcomes is the required next step to justify its widespread implementation (Frey, Nolen, Edstrom, & Hirschstein, 2005; Schoenwald & Hoagwood, 2001). To achieve effectiveness, Schoenwald and Hoagwood (2001) asserted that researchers must demonstrate an intervention's sustained effects on multiple behavioral and academic outcomes with students who represent a diverse range of demographic and functional characteristics in a variety of real-world school settings. In addition, participating families and schools should accept and support the intervention, and implementers should be able to deliver the intervention off-the-shelf with acceptable fidelity. More recently, an expert panel appointed by the Society for Prevention Research (SPR) has promulgated similar standards to be met by a body of research to assert that an intervention meets any of three increasingly stringent levels of evidence: efficacy, effectiveness, and readiness for broad dissemination (Flay et al., 2005). These standards specify that effectiveness studies must focus on factors such as quality of implementation under natural conditions and program adaptations that may contribute to variations in expected outcomes. To be ready for broad dissemination, a program must not only be proven effective, but it must also ensure that it can be adopted, implemented, and sustained in the field (Flay et al., 2005).

The purpose of this study was to conduct a large-scale RCT of the effectiveness of First Step. In this article, the authors (a) describe the methods, intervention procedures, and measures used to implement and evaluate First Step as typically delivered under naturally occurring conditions across multiple school districts; (b) report the intervention's effects on behavioral and academic outcomes among a diverse population of high-risk elementary students and according to different levels of implementation fidelity; and (c) assess the body of evidence on First Step against the SPR standards for effectiveness and widespread dissemination (Flay et al., 2005), thereby suggesting directions for further research.

Method

Participating Schools

This effectiveness study was conducted in 48 elementary schools in five geographically diverse settings across the United States, including Eugene and Springfield, Oregon; Huntington, West Virginia; Cook County, Illinois; San Jose, California; and Tampa, Florida. Researchers made on-site presentations about First Step and the study design to administrators at 84 schools in the districts, with a goal of enrolling 48 schools prior to randomization to ensure adequate analytic power. During the presentations, administrators were informed that their school had a 50% chance of being assigned to the intervention group if they participated in the project. Each school agreed to participate before school-level randomization occurred.

The average student population of the participating schools ranged from 356 to 772, and the proportion of students receiving free or reduced-price lunches ranged from 51% to 79%. The subsample of students from schools that participated in the study reflected the larger population of students in the districts on these factors.

Study Design

The 48 participating elementary schools entered the study in a staggered fashion, with 10 schools in San Jose, California, and 10 schools in Cook County, Illinois, enrolling in the 2006–2007 school year. The following year, 10 elementary schools in Huntington, West Virginia, 10 schools in Tampa, Florida, and 8 schools in Eugene and Springfield, Oregon, joined the study. Schools were randomly assigned to study groups, resulting in 24 intervention and 24 comparison schools. Researchers randomized groups of students at the school level to minimize the contamination of intervention that can occur when randomization occurs at the classroom level.

Participants and Procedures

In the fall and spring semesters, three students (one each in first through third grade) participated in the study, resulting in a total of six student participants across six classrooms during the school year. Students in the intervention group participated in the First Step program during the semester of their enrollment; researchers collected data on behavioral and academic outcomes at their enrollment (baseline) and immediately following completion of First Step (post test). Researchers collected data on the students in the comparison group at approximately the same temporal periods, with baseline collected at enrollment and approximately 3 months later (post test).

Assessing student eligibility. Participating teachers in intervention and comparison schools used a modified version of Stages 1 and 2 of the Systematic Screening for Behavior Disorders (SSBD) procedure (Walker & Severson, 1990) to identify eligible students. The SSBD has excellent psychometric characteristics, is nationally normed, and has been used in a number of research studies (e.g., Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007; Walker & Severson, 1990; Walker, Severson, & Seeley, 2007). The SSBD includes a multigated screening approach to detect if early elementary students have an elevated risk for internalizing or externalizing school behavior problems. To be eligible for this study, students had to demonstrate an elevated risk for externalizing school behavior problems. Although the SSBD has an optional Stage 3 that involves classroom and playground observations, this procedure was deemed too labor intensive to complete for the purposes of this study.

In Stage 1, after becoming familiar with students for at least 30 days, teachers rank-ordered up to five students in the classroom who exhibited the highest levels of externalizing behaviors (e.g., arguing, disturbing others, fighting). For the three highest ranked students, teachers completed brief ratings of students' behaviors in Stage 2 of the SSBD. The Adaptive Behavior Index (ABI), a 12-item scale ($\alpha = .82$), and the Maladaptive Behavior Index (MBI), an 11-item scale ($\alpha = .84$), assessed the frequency of students' adaptive and maladaptive behaviors as perceived by the teacher on a 5-point scale (from 1 = *never* to 5 = *frequently*). Teachers also completed the Critical Events Index (CEI) for each student, indicating how many of 30 high-saliency, low-frequency maladaptive indicators (e.g., stealing, setting fires, physical aggression) occurred during the past 30 days.

Students who were invited first to participate in the study had the highest average ranking on the ABI, MBI, and CEI of the three students rated by the teacher. If two students in the same classroom had the same average rank, the student with the higher raw score on the CEI was selected first. If the parents of the first selected student refused to consent to participate in the study, the research team sought consent from the parents of the next highest ranked student.

When the SSBD is used as a universal diagnostic screening tool in its entirety, a student is considered at an elevated risk for developing school behavior problems if the student meets specific scoring thresholds: (a) a score of 30 or lower on the ABI, a score of 35 or more on the MBI, and one to four critical events endorsed on the CEI or (b) five or more critical events endorsed on the CEI. Teachers' nominations of students with the poorest externalizing behaviors determined eligibility for this study, even if the student would not have officially qualified as at risk under the official diagnostic scoring criteria. This adapted procedure ensured that all participating classrooms included one participating student in the study.

SSBD scores were analyzed to determine how study eligibility characteristics—which prioritized teacher rankings—compared with official SSBD diagnostic eligibility

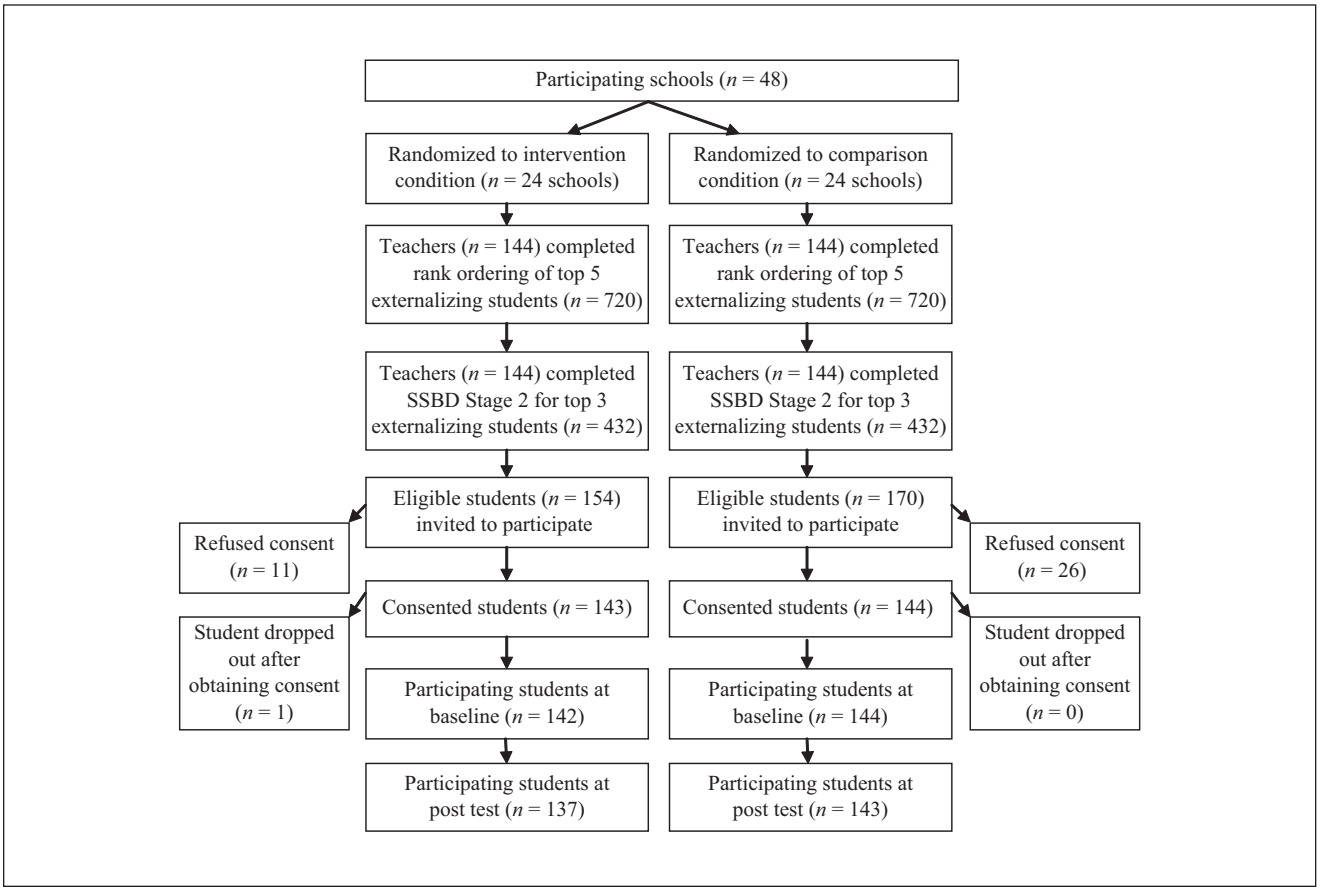

Figure 1. Schematic overview of the flow of research participants through randomization, screening, consent procedures, and data collection intervals (baseline and post test) of the First Step randomized controlled trial.
Note: SSBD = Systematic Screening for Behavior Disorders.

criteria—which prioritized scoring thresholds. Of the 286 eligible students, 223 (78%) met official SSBD Stage 2 diagnostic eligibility criteria. Of the 63 students who did not meet SSBD Stage 2 criteria, all had elevated CEI scores (i.e., one to four critical events) or met threshold criteria on either the ABI (i.e., a score of 30 or lower) or MBI (i.e., a score of 35 or higher). Overall, however, there was a statistically significant difference between study condition; more students in the intervention group met the official scoring thresholds (83% vs. 73%), χ^2 (1, 285) = 4.32, p = .04. This was not unexpected, as it was more difficult to obtain consent to participate in the comparison group, requiring researchers to invite students with lower rankings and initial severity levels to participate in the comparison group after students with higher rankings and behavior severity declined participation.

Consent rates. Across the 48 schools, 288 teachers completed Stages 1 and 2 of the SSBD for 1,440 students. A total of 22 teachers refused to complete the ABI, MBI, and CEI on all 3 highest ranking students in their classroom, so not all potentially eligible students (n = 22 students) received SSBD subscale scores, although all students invited to participate in the study had SSBD subscale scores.

Across both conditions, 324 students were ultimately invited to participate in the study. The number of invitations exceeded 288, because researchers invited the second- or third-ranked student if consent was not obtained for the higher ranked student. The parents of 287 students provided their consent for study participation (88%). After consent was obtained for 287 students, 1 student dropped out prior to baseline data collection. The final number of participating students was 286, with 142 students enrolled in the intervention group and 144 students enrolled in the comparison group (see Figure 1).

Behavior coaches. The First Step program requires the services of a behavior coach to work directly with participating students during the first 5 days of implementation, providing modeling and consultation to the classroom teachers and peers for the duration of the intervention (approximately 8 weeks). Behavior coaches can range in characteristics and qualifications, from paraprofessionals trained exclusively in First Step techniques to graduate-level

professionals with extensive behavioral support experience and training beyond First Step.

The approach the participating district administrators used to supply behavior coaches for the First Step intervention differed across sites. Coaches in Illinois were paraprofessionals who served schools within the borders of a special education cooperative serving six districts. As part of their regular duties, the paraprofessionals provided support services to schools in the cooperative while simultaneously acting as the First Step coach for one student each semester. The San Jose school district used school-based dropout prevention counselors as coaches. The counselors typically acted as the coach for one to two students each semester while fulfilling their regular school responsibilities (i.e., monitoring attendance, counseling students, coordinating student services). Coaches in West Virginia were drawn from a pool of retired teachers who often volunteered in the local elementary schools. The volunteers usually served as the First Step coach with one to two students a semester. School staff such as guidance counselors, social workers, school psychologists, and behavior specialists served as coaches in Tampa's schools (typically one student per semester for each staff member). University of Oregon graduate students served as coaches in the Eugene and Springfield sites. It was common practice in these school districts for graduate students to provide behavioral support in the schools, and for the effectiveness study they served as the coach for one to two students a semester.

Regardless of their background, all coaches attended a 2-day training provided at their locale by a First Step program developer. Coaches were assigned specific students with whom to implement First Step over a period of about 3 months, and were paid US$600 for each student who completed the program. Throughout the project, teachers and coaches were encouraged to take advantage of TA made available to staff from any school that purchased the First Step program. TA consisted of emails or conference calls with the First Step trainer to problem solve implementation issues.

Outcome Measures

A site coordinator was hired in each district to manage data collection activities, including supervision of six to eight research assistants (RAs). All site coordinators and RAs participated in extensive trainings conducted by the research team on the data collection procedures. Site coordinators conducted frequent reliability checks on all data collection measures to ensure adherence to study protocols. Students were temporally grouped across intervention and comparison groups, so that RAs administered outcome measures with students in both groups at about the same time in the school year.

Social skills rating system (SSRS). The SSRS–Teacher version (Gresham & Elliott, 1990) is a 57-item scale that assesses students' social skills, problem behaviors, and academic competence as perceived by the classroom teacher. The 30-item Social Skills subscale (SS; $\alpha = .88$) measures cooperation, assertion, and self-control on a 3-point scale (0 = *never*, 1 = *sometimes*, or 2 = *very often*). The 18-item Problem Behavior subscale (PB; $\alpha = .85$) measures students' internalizing and externalizing problem behaviors on the same 3-point scale. The 9-item Academic Competence subscale (AC; $\alpha = .91$) measures students' reading and math performance, motivation, intellectual functioning, and parental support on a 5-point percentage cluster scale (from 1 = *the lowest 10%* to 5 = *the highest 10%* compared with other students in the classroom).

The SSRS–Parent version (Gresham & Elliott, 1990) assesses students' social skills and problem behaviors as perceived by their parents/caregivers. The 38-item SS ($\alpha = .88$) measures social competence in day-to-day activities and interactions at home. The 17–item PB ($\alpha = .88$) measures internalizing and externalizing problem behaviors. The parent version is scaled and scored on the same 3-point scale (0 = *never*, 1 = *sometimes*, or 2 = *very often*) as the teacher-reported measures.

SSBD. As described above, a modified version of the SSBD (Walker & Severson, 1990) was used to identify students who had an elevated risk for school behavior problems and could be candidates for the First Step study. Teachers completed the three SSBD indices at baseline—the CEI, ABI, and MBI—and then the ABI and MBI again at post test for each student involved in the study for an additional behavioral outcome measure.

Academic engaged time (AET). RAs directly observed students in their classrooms for two 15-min academic periods to collect measures of AET, an indicator of students' academic involvement and adjustment to classroom expectations (Walker & Severson, 1990). Procedures used to observe and score the sessions mirrored those manualized for SSBD Stage 3 (Walker & Severson, 1990). Using a stopwatch recording procedure, RAs measured the proportion of time that each student (a) attended to the material and task, (b) made appropriate motor responses, (c) asked for assistance in an acceptable manner, (d) interacted with the teacher and classmates about academic matters, and/or (e) listened to teacher instructions and direction.

Prior to data collection, RAs attended a 2-day AET training session on standardized observation and coding procedures. All RAs were required to demonstrate and sustain a minimum .80 interobserver agreement level before and during data collection. Site supervisors monitored AET data collection and recorded reliability estimates for 33% of conducted observations; RAs were retrained as necessary throughout the study to minimize drift and ensure adequate reliability of recorded observations. The overall intraclass

correlation (ICC) of AET interrater reliability was excellent (ICC [3, 1] = .80).

RAs collected AET data on different days within a week of one another; these data were averaged to compute the proportion of observed time the student exhibited academic engagement. RAs collected AET observations during whole-class instruction (60%), individual seatwork (32%), or small-group work (8%), during which time the participating students were engaged in instruction related to language arts (59%), math (27%), or other topics (e.g., science, history, social studies). To minimize the effects of varying classroom contexts on student engagement, RAs attempted to collect post test data at the same time of day and during a similar classroom activity. There were no statistically significant differences in instructional setting, χ^2 (2, 1888) = 0.15, p = .93, or classroom subjects, χ^2 (4, 954) = 6.05, p = .20, across baseline and post test AET data collection occasions.

Woodcock-Johnson III Letter-Word Identification (WJ III LWI). To assess one aspect of students' literacy skills, RAs administered the LWI subtest from the WJ III Diagnostic Reading Battery (Woodcock, Mather, & Schrank, 2004). The WJ III LWI subtest measures a student's ability to identify isolated letters and words and has a median reliability of .91 for students 5 to 19 years of age (Woodcock et al., 2004).

Oral reading fluency (ORF). To gain another perspective on students' literacy skills, RAs administered two different 300- to 400-word first-grade level reading passages selected from a series of passages previously used in national studies (Fuchs, 2003). RAs computed a total ORF score at baseline and post test based on the average number of words read correctly by the student in 1 min.

Process Measures

Implementation fidelity measures of the First Step class- and home-based components, teacher and coach alliance, and social validity data were collected from the intervention group. The following sections describe the process measures in more detail.

Implementation Fidelity Checklist (IFC). The IFC (Walker et al., 2009; α = .94) was used to document the extent to which the coach and teacher delivered First Step components as intended and with high quality as perceived by an external observer (i.e., the RA).

RAs observed the implementer's adherence and quality in the classroom on three occasions: once for the coach during the first 5 program days and twice for the teacher at or around Program Days 10 and 15. RAs rated the implementer's adherence (either *yes* or *no*) according to 18 implementation components (e.g., whether the implementer announced the number of points needed for a reward, elicited cooperation from the class, provided positive feedback to the target student). Simultaneously, RAs rated the quality of implementation on a 5–point scale (0 = *very poor,* 0.25 = *poor,* 0.50 = *okay,* 0.75 = *good,* and 1.0 = *excellent*) on the same implementation components. The ICC assessing interrater reliability for 33% of the fidelity observations was excellent (ICC [3, 1] = .94).

Adherence scores were calculated as the proportion of procedures implemented. The means of the coach and teacher adherence scores and quality ratings were combined to estimate overall classroom adherence and quality.

Classroom monitoring form (CMF). First Step is organized into 30 distinct program days, and a student must meet daily performance criteria to proceed to the next day. If the student fails to meet the criteria, he or she must repeat that day's intervention. Teachers used the CMF (Walker et al., 2009) daily to document implementation of First Step and the students' compliance with daily goals. Teachers tabulated the points possible and earned by the students, and whether a recycle day (i.e., repeating a program day) was necessary due to the students' not meeting goal criterion.

Classroom dosage was calculated as the proportion of program days delivered by the coach and teacher (out of 30 possible), and student compliance was calculated as the proportion of days when the number of points earned was equal to or greater than the number needed to earn the daily goal.

homeBase monitoring form (HMF). Coaches used the HMF (Walker et al., 2009) to document the extent to which they perceived the parents/caregivers were engaged in the homeBase component of First Step. Coaches completed the HMF after each 1-hr session with the parents/caregivers, rating their level of fidelity on a 3-point scale (0 = *low,* 0.5 = *medium*, and 1.0 = *high*). A high fidelity rating indicated that the parent/caregiver was engaged with homeBase training and procedures, followed through on assignments, and applied procedures with skill, sensitivity, and confidence. Session topics included communication and sharing, cooperation, limits setting, problem solving, friendship skills, and confidence.

Dosage of the homeBase component was calculated as the proportion of sessions (out of six possible) in which the parent(s) participated. The homeBase mean dosage was combined with the classroom dosage (as measured on the CMF) to calculate an overall measure of First Step dosage. The homeBase mean fidelity rating was combined with the classroom quality mean rating (as measured on the IFC) to calculate an overall quality measure of First Step delivery.

Alliance survey. To assess alliance at the conclusion of First Step, teachers in the intervention group and their coaches completed a 10-item rating scale (α = .95) developed by the research team. Aspects of alliance addressed included the degree to which the teacher or coach perceived their relationship as characterized by trust, collaboration, and shared goals; whether the teacher/coach sincerely desired to understand and improve the behavior addressed; and whether the time spent working with the teacher/coach

was effective and productive. Respondents rated each item on a 5–point scale (from 1 = *never* to 5 = *always*).

Satisfaction survey. To assess satisfaction with the First Step program, teachers and parents in the intervention group completed rating scales developed by the research team. The 13–item Teacher Satisfaction Survey (α = .90) assessed teachers' perceptions of the training and support received, the effectiveness of First Step in changing student behavior and peer interactions, and their willingness to use and recommend First Step in the future. The 12-item Parent Satisfaction Survey (α = .93) assessed the parent's perceptions of the effectiveness and value of the First Step program based on its impact on the child's behavior at home. Satisfaction items were scored on a 5–point scale (from 1 = *strongly disagree* to 5 = *strongly agree*).

Statistical Analysis

Missing data. The proportion of missing data on outcomes measures administered with the student or teacher (i.e., AET, ORF, WJ III LWI, SSRS-Teacher, ABI, MBI) ranged from 1.4% to 4.5% at baseline and from 3.8% to 6.6% at post test. The proportion of missing data on the outcome measure administered to the parents/caregivers (i.e., SSRS-Parent) was 4.5% at baseline and 15.7% at post test.

To avoid losing cases and to reduce potential bias, missing values were imputed using fully conditional specification models (i.e., logistic, polytomous, or linear regression) applied iteratively using Stata's "ice" procedure (which implements multiple imputation by chained equations) (Royston, 2004; Van Buren, Brands, Groothuis-Oudshoorn, & Rubin, 2006). Separate imputations were conducted for intervention and comparison group students, and five values were imputed for each missing value. These results were combined to provide estimates of the variability and *p* values for regression coefficients. Up to 12 variables were included in each model as predictors, including observations of the dependent measure at one or more points in time when values were available. Other student-specific variables in one or more imputation models included student age, grade level, gender, race/ethnicity, language, English language learner status, lunch program status, and special education status. Teacher-specific variables were a summary measure of teacher self-reported knowledge and skills in working with students with behavior problems (Cheney, Walker, & Blum, 2004), years of teaching experience, and whether the teacher was fully credentialed.

Analysis of intervention effects. Hierarchical linear modeling (HLM) regressions were performed on each set of imputed data to estimate intervention effects. The dependent variables were measures of student academic ability, social skills, or behavior as measured by direct assessment (i.e., AET, ORF, or WJ III) or teacher and parent reports (i.e., SSRS-Teacher subscales, SSRS–Parent subscales, ABI, or MBI). The independent variables included a constant, the baseline measure for that dependent variable, and a group indicator. For comparison purposes, regressions were also run on the data with and without imputed values including independent variables for student age at baseline, grade, gender, race/ethnicity, lunch program status, special education status, teacher self-reported knowledge and skills, and baseline MBI. Levels in the model included student and school, with additive random effects for each.

Results from the HLM models on each imputed data set were combined using the Stata "mim" procedure for working with multiply imputed data sets that implements Rubin's method (Rubin, 1987). Because multiple measures were tested for intervention effects, the Benjamini–Hochberg (BH) correction for Type 1 error rate was applied to the 10 univariate tests (see Schochet, 2008). That is, for any given test, the reported *p* value was the smallest false discovery rate (FDR) value for which the corresponding null hypothesis was rejected.

Effect sizes were reported as Cohen's *d* statistic (Cohen, 1988) and were calculated by dividing the intervention indicator coefficient by an estimate of the pooled between-student standard deviation at post test. The latter was obtained using an HLM regression on the imputed data in which the dependent variable was the outcome at post test, the independent variables were a constant and the intervention indicator, and there were random additive effects for student and school. In addition, the What Works Clearinghouse Improvement Index (WWC II; Valentine & Cooper, 2003; What Works Clearinghouse, 2008) was reported as a measure of the practical significance of the findings by translating the effect sizes into an improvement index that represents "the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the comparison group mean (i.e., the 50th percentile) in the comparison group distribution" (p. 25). The index can be interpreted as the expected change in percentile rank for the median comparison group student if that student had participated in First Step.

Results

Baseline Equivalence Analysis

The equivalence of the participants in the intervention and comparison groups was examined at baseline. For continuous and ordinal variables, the statistical significance of the difference between the two groups at baseline was determined from linear HLM regressions. The regressions included a variable indicating whether the student was in the intervention or comparison group and random effects for students and schools. For dichotomous variables, statistical significance was determined by using a logistic HLM regression test.

Table 1. Student Demographic and Behavioral Characteristics at Baseline

Characteristic	Total (N = 286)	Intervention (n = 142)	Comparison (n = 144)	p
Age *M* (*SD*)	7.90 (0.99)	7.87 (1.00)	7.93 (0.97)	.63
Percentage				
Male	77.13	75.07	79.17	.46
Grade level				
First	31.82	33.10	30.56	.64
Second	35.32	33.10	37.50	.44
Third	32.16	32.39	31.94	.94
SSBD rank				
First-ranked student	60.26	61.97	58.57	.56
Second-ranked student	22.69	23.24	22.14	.73
Third-ranked student	17.05	14.79	19.28	.32
Race/ethnicity				
Black	23.98	32.39	15.69	.07
Hispanic	27.06	19.58	34.44	.26
White	44.61	43.38	45.83	.63
Primary language				
Spanish	11.68	7.04	16.25	.13
ELL	13.01	7.18	18.75	.12
Eligible for lunch program	72.52	73.66	71.39	.75
With a behavior support plan	23.08	25.07	21.11	.67
With an IEP or 504 plan	15.31	17.32	13.33	.38
Mean scores				
SSBD-ABI	33.95	33.22	34.68	.19
SSBD-MBI	31.79	33.13	30.47	.01*
SSRS-SS-Teacher	83.66	82.78	84.53	.37
SSRS-SS-Parent	89.36	88.21	90.50	.25
SSRS-PB-Teacher	120.19	120.95	119.45	.31
SSRS-PB-Parent	114.42	114.65	114.18	.87
SSRS-AC-Teacher	87.22	87.02	87.43	.85
AET	0.59	0.59	0.59	.99
WJ III LWI	99.80	99.71	99.89	.92
ORF	62.77	59.82	65.68	.26

Note: SSBD = Systematic Screening for Behavior Disorders; ELL = English language learner; IEP = Individualized Education Program; ABI = adaptive behavior index; MBI = maladaptive behavior index; SSRS = Social Skills Rating System; SS = Social Skills subscale; PB = Problem Behavior subscale; AC = Academic Competence subscale; AET = academic engaged time; WJ III LWI = *Woodcock-Johnson III Letter-Word Identification subtest;* ORF = oral reading fluency (words per minute).
*$p < .05$.

Results indicated that First Step participants were not significantly different from comparison group students on demographics (e.g., age, language, race/ethnicity) or school factors (e.g., special education or lunch program status). In addition, no significant differences between groups were found on 9 out of 10 baseline behavioral or academic measures; however, teachers rated students in the intervention group with significantly more maladaptive behaviors on the initial MBI (M_{MBI} = 33.13, SD = 7.71) than did teachers of students in the comparison group (M_{MBI} = 30.47, SD = 8.89; p = .01; see Table 1).

Intervention Fidelity, Therapeutic Alliance, and Satisfaction

Overall adherence to First Step implementation protocols was satisfactory: Coaches and teachers implemented a majority of procedures as intended (M = 0.76, SD = 0.15). Implementation quality was in the *good* to *excellent* range (M = 0.78, SD = 0.15) for overall classroom implementation and in the *medium* to *high* range (M = 0.63, SD = 0.33) for the homeBase component. With respect to dosage, students received a majority of the intended classroom program

Table 2. Means and Standard Deviations for Baseline and Post test Outcome Measures and HLM Results

	Intervention (n = 142)		Comparison (n = 144)					
Domain/measure	Baseline *M* (*SD*)	Post test *M* (*SD*)	Baseline *M* (*SD*)	Post test *M* (*SD*)	Treatment coefficient (*SE*)	p[a]	d	WWC II
Prosocial/adaptive behavior								
SSBD-ABI	33.2 (7.9)	38.1 (9.3)	34.7 (7.6)	35.3 (8.7)	3.81 (1.04)	.00	.42	+16.3
SSRS-SS-Teacher	82.8 (11.8)	92.4 (13.2)	84.5 (11.5)	85.3 (11.7)	8.35 (1.22)	.00	.67	+24.7
SSRS-SS-Parent	88.2 (15.3)	94.7 (16.2)	90.5 (15.9)	90.4 (16.8)	5.86 (1.67)	.01	.33	+12.8
Problem/maladaptive behavior								
SSBD-MBI	33.2 (7.7)	28.1 (9.0)	30.5 (8.9)	29.7 (9.5)	−3.30 (1.06)	.01	−.36	+13.9
SSRS-PB-Teacher	120.9 (10.1)	115.7 (12.9)	119.4 (11.7)	119.2 (10.9)	−4.60 (1.11)	.00	−.38	+14.9
SSRS-PB-Parent	114.6 (14.0)	108.8 (15.2)	114.2 (14.3)	111.5 (13.2)	−3.03 (1.64)	.07	−.21	+8.4
Academic								
SSRS-AC-Teacher	87.0 (10.9)	88.1 (10.9)	87.4 (11.2)	86.3 (11.3)	2.11 (.70)	.01	.19	+7.5
AET	.60 (.20)	.73 (.19)	.59 (.20)	.66 (.19)	.061 (.03)	.02	.35	+13.5
WJ III LWI	99.7 (13.4)	100.6 (12.9)	99.9 (13.3)	102.4 (17.5)	−1.60 (1.42)	.26	−.10	+4.1
ORF	59.8 (40.6)	71.1 (42.9)	65.7 (45.9)	71.9 (43.8)	4.71 (2.0)	.02	.11	+4.3

Note: HLM = hierarchical linear modeling; WWC II = What Works Clearinghouse Improvement Index; SSBD = Systematic Screening for Behavior Disorders; ABI = adaptive behavior index; SSRS = Social Skills Rating System; SS = Social Skills subscale; MBI = maladaptive behavior index; PB = Problem Behavior subscale; AC = Academic Competence subscale; AET = academic engaged time; WJ III LWI = *Woodcock-Johnson III Letter-Word Identification subtest;* ORF = oral reading fluency (words per minute).
[a]p value after applying the Benjamini–Hochberg correction for multiple comparisons.

days (M = 0.84, SD = 0.25), and parents participated in most of the homeBase sessions (M = 0.80, SD = 0.33). Finally, students completed a high proportion of the program days according to criterion without needing to repeat a session (M = 0.95, SD = 0.11).

Alliance was rated highly overall by coaches (M = 4.42, SD = 0.59) and teachers (M = 4.44, SD = 0.50). Parents' overall satisfaction ratings also were favorable (M = 4.21, SD = 0.62); however, teachers reported moderate overall satisfaction ratings (M = 3.54, SD = 0.69). Low teacher ratings (items with mean ratings < 3.0) were reported for 2 of the 13 items on the satisfaction survey, including (a) *The program did not take much of my time* and (b) *The program did not interfere with my other teaching activities/responsibilities*.

Pre–Post Changes in Outcome Measures

Intervention effects at post test were examined in three domains: (a) prosocial/adaptive behavior (including three outcome measures: ABI, SSRS-SS-Teacher, and the SSRS-SS-Parent), (b) problem/maladaptive behavior (three outcome measures: MBI, SSRS-PB-Teacher, and SSRS-PB-Parent), and (c) academic (four outcome measures: SSRS-AC, AET, WJ III LWI, and ORF). The BH procedure was used to control the FDR at post test at the .05 level within each domain.

Data demonstrated that First Step was successful in improving the behavior and social skills of participating students. First Step participants had significantly higher prosocial and adaptive skills and significantly fewer problem or maladaptive behaviors at post test than comparison group students.

Prosocial/adaptive behavior domain. Results indicated statistically significant positive effects for both SSRS-SS subscales rated by teachers (p < .01; d = .67) and parents (p = .01; d = .33; see Table 2), indicating that First Step participants at post test had significantly improved their social skills beyond those of comparison group students. First Step students also significantly increased their adaptive behaviors as measured by the ABI (p < .01; d = .42) beyond their comparison group peers. The WWC II indicated that First Step students achieved an average percentile ranking that was approximately 13 to 25 percentile points higher than the ranking of the average student in the comparison group on the prosocial/adaptive behavior domain measures.

Problem/maladaptive behavior domain. First Step participants reduced their problem behaviors to a significantly greater degree than the comparison group as perceived by their teachers on the SSRS-PB subscale (p < .01; d = −.38) and the MBI (p = .01; d = −.36). The WWC II indicated that First Step students achieved an average percentile ranking that was approximately 14 to 15 percentile points higher than the ranking of the average student in the comparison group on the problem/maladaptive behavior domain measures. However, although parents reported that problem behaviors declined, there was no statistically significant difference between the groups at post test on the parent-completed SSRS-PB subscale (p = .07).

Academic domain. Significant effects were noted on three out of four academic performance and participation measures. First, results indicated that First Step participants had significantly improved their ability to sustain attention and engagement in academic tasks beyond that of comparison group students, as measured by the AET observations ($p = .02$; $d = .35$). Teachers also perceived that First Step students had significantly greater improvement in their academic competence, as measured on the SSRS-AC subscale ($p = .01$; $d = .19$). The WWC II indicated that First Step students achieved an average percentile ranking that was approximately 8 to 14 percentile points higher than the ranking of the average student in the comparison group on these academic measures.

Although there also was a significant difference between the two groups on the ORF measure ($p = .02$; $d = .11$), the effect size was small. The two groups did not differ significantly on the WJ III LWI subtest measure ($p = .26$).

Relationship of Fidelity and Outcomes

Classroom fidelity measures. HLM regressions were performed to determine whether First Step students whose teachers and coaches implemented the program with higher fidelity (i.e., adherence and quality) achieved better outcomes than students whose intervention was delivered with lower fidelity. HLM regressions were conducted where the dependent variables were the post test outcomes, and the independent variables were the baseline values, a constant, a group indicator, and the product of the group indicator and a fidelity measure.

Average ratings from the IFC (completed by the RA) for each teacher were normalized to have a mean of zero and unit variance for First Step students, and defined as zero for comparison group students. Four of the 10 outcome measures at post test had statistically significant fidelity effects (adjusted for multiple tests) on intervention effectiveness, all of which were on teacher-reported measures (i.e., ABI, SSRS-SS-Teacher, MBI, and SSRS–PB-Teacher). A 1 standard deviation increase in fidelity increased the intervention effect on ABI scores by 57% ($p = .01$) and on SSRS-SS scores by 25% ($p = .04$), and decreased MBI scores by 68% ($p = .01$) and SSRS-PB scores by 46% ($p = .02$).

Although fidelity is an endogenous measure and causality cannot be attributed to these findings, the results suggest that particularly poor implementation (e.g., receiving fidelity ratings 2 SDs below the average score) was associated with almost no effect on teachers' SSRS-PB ratings, and with an adverse effect on teachers' MBI and ABI ratings. Conversely, particularly good implementation was associated with a doubling of the intervention effect on these outcomes at post test.

Classroom dosage measures. HLM regressions were performed on the CMF (completed by the teacher) measures of dosage and student compliance. CMF measures indicated that a higher proportion of intervention days delivered successfully in the classroom was associated with better adaptive behavior (as measured by teachers' ratings on the ABI), more social skills (as measured by teachers' ratings on the SSRS-SS), fewer problem behaviors (as measured by teachers' ratings on the MBI and SSRS-PB), and higher academic engagement (as measured by AET).

A 1 standard deviation increase in classroom dosage increased the intervention effect on ABI scores by 31% ($p < .01$) and SSRS-SS scores by 27% ($p = .00$), decreased MBI scores by 32% ($p = .00$) and SSRS-PB scores by 26% ($p < .01$), and increased AET by 29% ($p < .01$).

homeBase fidelity and dosage measures. HLM regressions on the HMF measures of quality and dosage (i.e., the proportion of parent education sessions delivered times the average fidelity score for the lessons) indicated that higher dosage was associated with better adaptive behavior (as measured by teachers' ratings on the ABI), more social skills (as measured by teachers' ratings on the SSRS-SS), and fewer problem behaviors (as measured by teachers' ratings on the MBI and the SSRS-PB).

A 1 standard deviation increase in homeBase dosage increased the intervention effect on ABI scores by 20% ($p = .01$) and SSRS-SS scores by 21% ($p = .01$), and decreased MBI scores by 19% ($p = .03$) and SSRS-PB scores by 21% ($p = .01$). The homeBase component was not statistically related to beneficial changes in the parent-reported problem behavior and social skills ratings (i.e., SSRS-PB-Parent or SSRS-SS-Parent).

Discussion

The findings from this effectiveness study demonstrate that First Step can be implemented with fidelity with diverse student populations using only the materials and support typically available to those who purchase the intervention (see Note 1). Participating coaches and classroom teachers established positive working relationships in support of their students with behavior problems, and they reported satisfaction with the First Step program. Significant positive effects were documented on multiple measures across multiple domains over a 3-month period. Baseline to post test differences between intervention and comparison group students demonstrated that students who participated in First Step made significantly greater gains in prosocial and adaptive behaviors and reduced their problem and maladaptive behaviors, with effect sizes ranging from −.38 (SSRS-PB-Teacher) to .67 (SSRS-SS-Teacher). The WWC II indicated that First Step students achieved an overall percentile ranking in the prosocial/adaptive behavior domain that was almost 18 percentile points higher than the median comparison group student, and the overall ranking in the problem/maladaptive behavior was about 14 percentile

points higher for First Step students. Furthermore, higher implementation fidelity in the classroom and at home was strongly associated with greater gains across teacher-reported measures of adaptive behavior and social skills and greater reductions in problem behaviors.

With respect to the academic domains, students who participated in First Step once again made significant gains relative to the comparison group in most categories. Effect sizes varied depending on the type of academic measure observed, with a substantially larger effect size noted for academic engagement (.35) relative to the effect sizes for academic competence (.19) and ORF (.11). Such differences in academic outcome measures may indicate that academic engagement is a more proximal outcome of the First Step intervention, whereas the impact of First Step on academic functioning is a distal outcome that requires more time to develop than could be measured in this particular study.

This pattern of findings is similar to those demonstrated in a recent efficacy study of First Step (Walker et al., 2009), and in an independent analysis of data from that study (Woodbridge et al., 2010), with some exceptions. This effectiveness study produced generally lower scores on process measures and significant but smaller effects than the efficacy study. For example, the efficacy study registered a combined adherence score for coaches and teachers of .83 (Walker et al., 2009), compared with a score of .76 for this effectiveness study. Similarly, effect sizes for the prosocial/adaptive behavior domain ranged from .54 to .87 in the efficacy study (Walker et al., 2009) and from .33 to .67 in this effectiveness study. As noted, this pattern of smaller effects is common when contrasting studies in which the developers are and are not directly involved in implementation (e.g., Weisz & Jensen, 2001).

Study Limitations

Several limitations of the First Step effectiveness study should be noted. First, behaviors are the result of ratings on an instrument rather than a direct measure of the behaviors of interest. The limitation of instrument-based behavior ratings may be more acute due to the low raw frequencies of disruptive and aggressive behaviors typically found in school settings. In addition, a few of the measures used in the study, namely, the Alliance and Satisfaction surveys, were created by the research team with untested psychometric properties, which may limit the construct validity and generalizability of findings. However, the behavior rating assessments used (SSRS and SSBD) are popular and highly validated instruments, and all instruments used in the study were highly reliable, as supported by robust Cronbach's alphas; thus, the researchers felt that these measures were appropriate to assess the impact of First Step in this effectiveness study.

Another limitation was that this study did not assess the fidelity of implementation of the homeBase component to the same extent as the classroom assessments. The reasoning behind this decision was to ensure that the one-on-one relationship between the behavior coach and families would not be disrupted by the researcher's presence. Nonetheless, the researchers acknowledge that the lack of observations in the home environment results in a lack of data that could have elucidated the effects of the parent component on a child's behavioral and academic outcomes.

First Step and the SPR Standards for Effectiveness

Standards for establishing the effectiveness (see, Flay et al., 2005) of First Step in improving the behavior and academic skills of young students with behavior problems are satisfied by multiple characteristics of this study, including the following: (a) a strong efficacy foundation, (b) a rigorous experimental design, (c) a diverse study population, (d) off-the-shelf implementation, (e) commercial availability of materials and technical support, (f) high-quality measures, (g) an appropriate analytic approach, and (h) consistently positive and significant effects across multiple domains. However, careful examinations of existing efficacy and effectiveness studies as well as planned future analyses of outcome data are necessary to address several remaining standards.

For example, a critical standard of effectiveness states that analyses of an intervention must identify the population(s) for whom it is effective. Early studies of First Step were conducted primarily in Oregon schools (Walker et al., 1998; Walker, Golly, McLane, & Kimmich, 2005) and showed positive results with these populations. The more recent efficacy study (Walker et al., 2009), conducted with a predominantly Hispanic population, demonstrated significant effects for First Step participants as a whole. Independent analyses of the First Step efficacy study (Woodbridge et al., 2010) explored interaction effects of numerous student-level covariates (e.g., age, gender, race/ethnicity, special education status) and found that effects were robust across student subpopulations. Future analyses of the effectiveness study data will explore mediators and moderators of effects, including student characteristics, to continue to develop the evidence regarding populations for which the program might work more or less effectively.

Another standard to be met to assert an intervention's effectiveness is the maintenance of positive effects over time (Flay et al., 2005). Follow-up data were collected as part of this effectiveness study and will be analyzed and reported in the near future. However, an initial examination of the maintenance of First Step gains as part of the independent analysis of the First Step efficacy study (Woodbridge et al., 2010) showed no statistically significant differences

between intervention and comparison group students 1 year after students completed the program. Therefore, questions remain as to whether the First Step intervention in its current form is sufficient for long-term efficacy or if modifications to the intervention, such as the addition of "booster" sessions, would increase the likelihood of positive effects over time.

To determine a program's effectiveness also calls for measurement of key components of active ingredients in the intervention and comparison groups. Often, the comparison condition entails services that address the same problems that are the focus of the intervention under study. Specification of alternative treatments in the comparison group helps to clarify the degree of difference between groups that produced the demonstrated effects. To date, this standard has not been met in First Step efficacy studies or this effectiveness study.

Beyond these standards for determining the effectiveness of an intervention, Flay et al. (2005) asserted that for an intervention to be ready for broad dissemination, it must provide estimates of program costs that encompass all burdens placed on the organization considering implementation. To date, all necessary materials and resources for acceptable levels of implementation are commercially available (including CD-ROMS of training sessions). As for human resource needs, coaches typically spend 30 to 50 hr over a 3-month period implementing First Step with one student (including training time). Depending on caseload and other duties, coaches can usually accommodate two to four students in the First Step program at one time (H. Walker & A. Golly, personal communication, July 12, 2006).

This comparison of First Step evidence with the SPR standards set forth by Flay et al. (2005) suggested several directions for further research. The evidence base for the short-term effects of First Step is clearly strong, but further evaluation of longer term efficacy and effectiveness is warranted. In addition, attention to describing the comparison condition would clarify the differences between groups on which the evidence of effectiveness is based. Finally, the SPR clearly values and encourages replication studies, even when the evidence is reasonably solid, and particularly when an intervention is implemented in a new context, with new populations, or with other implementation support.

Author's Note

First Step to Success is currently published by Sopris West (www.soprislearning.com).

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by a Social and Behavioral Outcomes to Support Learning research grant (R324B060003), a program of the U.S. Department of Education, Institute of Education Sciences, National Center for Special Education Research.

References

Beard, K. Y., & Sugai, G. M. (2004). First Step to Success: An early intervention for elementary children at risk for antisocial behavior. *Behavioral Disorders, 29*, 396–409.

Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.

Cheney, D., Walker, B., & Blum, C. (2004). *Teacher knowledge and skills survey (Version 2.0)*. Seattle, WA: University of Washington.

Clarke, G. N., Hawkins, W., Murphy, M., Sheeber, L. B., Lewinsohn, P. M., & Seeley, J. R. (1995). Targeted prevention of unipolar depressive disorder in an at-risk sample of high school adolescents: A randomized trial of a group cognitive intervention. *Journal of American Academy of Child & Adolescent Psychiatry, 34*, 312–321.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

DiPerna, J., & Elliott, S. N. (2002). Promoting academic enablers to improve student performance: Considerations for research and practice [Special issue]. *School Psychology Review, 31*, 293–405.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. *Prevention Sciences, 6*, 151–175.

Frey, K. S., Nolen, S. B., Edstrom, L. V., & Hirschstein, M. K. (2005). Effects of a school-based social-emotional competence program: Linking children's goals, attributions, and behavior. *Journal of Applied Developmental Psychology, 26*, 171–200.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research & Practice, 18*, 172–186.

Gresham, F. M., Cook, C. J., Crews, S. D., & Kern, L. (2004). Social skills training for children and youth with emotional and behavioral disorders: Validity considerations and future directions. *Behavioral Disorders, 30*, 19–33.

Gresham, F. M., & Elliott, S. N. (1990). *The social skills rating system (SSRS)*. Circle Pines, MN: American Guidance Service.

Hoagwood, K., Hibbs, E. D., Brent, D., & Jensen, P. S. (1995). Introduction to the special section: Efficacy and effectiveness in studies of child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63*, 683–687.

Hunter, L., Hoagwood, K., Evans, S., Weist, M., Smith, C., Paternite, C., . . . the School Mental Health Alliance. (2005). *Working together to promote academic performance, social and emotional learning, and mental health for all children*. New York, NY: Center for the Advancement of Children's Mental Health at Columbia University.

Lien-Thorne, S., & Kamps, D. (2005). Replication study of the First Step to Success early intervention program. *Behavioral Disorders, 31*, 18–32.

Overton, S., McKenzie, L., King, K., & Osbourne, J. (2002). Replication of the First Step to Success model: A multiple-case study of implementation effectiveness. *Behavioral Disorders, 28*, 40–56.

Payne, C. (2008). *So much reform, so little change*. Cambridge, MA: Harvard Educational Press.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal, 4*, 227–241.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley.

Schalock, R. (1989). Person-environment analysis: Short and long term perspectives. In W. Kiernan & R. Schalock (Eds.), *Economics, industry and disability: A look ahead* (pp. 105–115). Baltimore, MD: Paul H. Brookes.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schoenwald, S., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services, 52*, 1190–1197.

Scott, T. M. (2001). A schoolwide example of positive behavior support. *Journal of Positive Behavior Interventions, 3*, 88–94.

Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology, 45*, 193–223.

Sprague, J., & Perkins, K. (2009). Direct and collateral effects of the First Step to Success program. *Journal of Positive Behavior Interventions, 11*, 208–221.

Supovitz, J. A., & Weinbaum, E. (2008). *The implementation gap: Understanding reform in high schools*. New York, NY: Teachers College Press.

Valentine, J. C., & Cooper, H. (2003). *What works clearinghouse study design and implementation assessment device (Version 1.0)*. Washington, DC: U.S. Department of Education.

Van Buren, S., Brands, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*, 1049–1064.

Walker, H. M., Golly, A. M., McLane, J. Z., & Kimmich, M. (2005). The Oregon First Step to Success replication initiative: Statewide results of an evaluation of the program's impact. *Journal of Emotional and Behavioral Disorders, 13*, 163–172.

Walker, H. M., Kavanagh, K., Stiller, B., Golly, A., Severson, H. H., & Feil, E. G. (1997). *First Step to Success: An early intervention program for antisocial kindergartners*. Longmont, CO: Sopris West.

Walker, H. M., Kavanagh, K., Stiller, B., Golly, A., Severson, H. H., & Feil, E. G. (1998). First Step to Success: An early intervention approach for preventing school antisocial behavior. *Journal of Emotional and Behavioral Disorders, 6*, 66–80.

Walker, H. M., Ramsey, E., & Gresham, F. (2004). *Antisocial behavior in school: Evidence-based practices* (2nd ed.). Belmont, CA: Wadsworth/Thomson Learning.

Walker, H. M., Seeley, J. R., Small, J., Severson, H. H., Graham, B. A., Feil, E. G., . . . Forness, S. R. (2009). A randomized controlled trial of the First Step to Success early intervention: Demonstration of program efficacy outcomes in a diverse, urban school district. *Journal of Emotional and Behavioral Disorders, 17*, 197–212.

Walker, H. M., & Severson, H. H. (1990). *Systematic Screening for Behavior Disorders (SSBD): User's guide and technical manual*. Longmont, CO: Sopris West.

Walker, H. M., Severson, H. H., & Seeley, J. (2007). *Universal, school-based screening for the early detection of academic and behavioral problems contributing to later destructive outcomes* (Paper commissioned by the Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth, and Young Adults). Washington, DC: National Research Council and Institute of Medicine.

Weisz, J. R., & Jensen, A. L. (2001). Efficacy and effectiveness of psychotherapy with children and adolescents. *European Child & Adolescent Psychiatry, 10*, I12–I18.

What Works Clearinghouse. (2008). *Procedures and standards handbook (Version 2.1)*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Woodbridge, M., Sumi, W. C., Thornton, P., Javitz, H., Wagner, M., & Shaver, D. (2010). *National behavior research coordination center: Evaluation results for four behavior interventions*. Menlo Park, CA: SRI International.

Woodcock, R. W., Mather, N., & Schrank, F. A. (2004). *Woodcock-Johnson III diagnostic reading battery*. Itasca, IL: Riverside Publishing.