School Psychology Review,
2014, Volume 43, No. 3, pp. 299–317

Children, Research, and Public Policy

Does First Step to Success Have Long-Term Impacts on Student Behavior? An Analysis of Efficacy Trial Data

Michelle W. Woodbridge, W. Carl Sumi,
Mary M. Wagner, and Harold S. Javitz
SRI International

John R. Seeley, Hill M. Walker, Jason W. Small, Annemieke Golly,
Edward G. Feil, and Herbert H. Severson
Oregon Research Institute

Abstract. First Step to Success (First Step; Walker et al., 1997, 1998) is a secondary-level intervention for students with behavior problems in early elementary school. The purposes of this study were to assess whether effects in student behavior and academics at posttest shown in a recent efficacy trial (Walker et al., 2009) were maintained at follow-up and to examine the relationship of implementation fidelity to outcomes. The findings showed that although First Step's initial impact was significant and positive across all behavior and some academic measures, gains eroded 1 year after the intervention was withdrawn. Results are discussed in the context of students' experience of yearly change in classroom environments, teachers' variable behavioral expectations and perceptions, and the need for intervention maintenance plans to support sustainment of treatment effects.

As school districts face serious budget cuts with concomitant pressure to improve test scores, it is increasingly important that their limited resources be used to support evidence-based programs and interventions that address widespread problems with high social and economic costs such as antisocial behavior exhibited by young children. The rates of aggressive and antisocial behavior among children have increased over the past 50 years and constitute "a major public health problem for society" (Connor, 2004, p. 28). Researchers

The contents of this research report were developed under a $5.9 million grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and one should not assume endorsement by the Federal Government.

Correspondence concerning this article should be addressed to Michelle W. Woodbridge, SRI International, 333 Ravenswood Ave, BS 124, Menlo Park, CA 94025; e-mail: michelle.woodbridge@sri.com

argue that if behavior problems are not corrected early, they can lead to serious behavior disorders that can persist over the life course (Kazdin, 1987; Moffitt, 1993; Webster-Stratton & Taylor, 2001).

When children with emotional and behavioral disorders receive special education services to help them succeed at school, they may still have lower school attendance rates, grades, and graduation rates than students without disabilities or with any other disability classification (Blackorby, Chorost, Garza, & Guzman, 2005; Wagner et al., 2003; Wagner, Newman, Cameto, Garza, & Levine, 2005). In many cases, this scenario of low attendance, grades, and graduation rates leads to a poor transition to young adulthood and adverse life outcomes for youth with emotional and behavioral disorders (Wagner & Davis, 2006), including a 57% likelihood of being arrested within 2 years of leaving high school, employment instability, and risk of entering the adult mental health treatment system (Kessler, Chiu, Demler, & Walters, 2005; Kutash, Duchnowski, & Lynn, 2006; Newman, Wagner, Cameto, & Knokey, 2009; Wagner et al., 2005).

FIRST STEP TO SUCCESS: A MODEL PROGRAM TO PROMOTE POSITIVE BEHAVIOR

First Step to Success (i.e., First Step) is an evidence-based program (Sprague & Perkins, 2009; Walker et al., 2009) for addressing children's behavior problems at school and teaching positive, prosocial behavior. It is a secondary-level intervention (i.e., implemented when children do not respond to primary, school-wide universal prevention strategies) for early elementary students who have moderate to severe behavior problems (Walker et al., 1997, 1998). Grounded in a social–ecological model (Bronfenbrenner, 1979; Schalock, 1989), the fundamental principle behind First Step is that when peers and adults who are central to children's experience learn and systematically apply strategies for eliciting and reinforcing positive behaviors, long-term improvements in children's behavior at school and at home can result. First Step is expected to achieve behavioral improvements by applying three modular components in concert: (a) universal screening for behavior problems; (b) classroom intervention involving the target student, peers, and teacher; and (c) in-home parent education designed to strengthen parenting skills and the home–school relationship.

The First Step screening module involves teachers identifying students in their classrooms who are at risk of, or already exhibiting, internalizing or externalizing behavior problems and evaluating the students using standard measures of antisocial behavior. Screened children who exceed criteria and/or cutoff points are considered appropriate candidates for First Step; one student per classroom can participate in the program at a time.

The First Step classroom intervention, usually implemented over a period of 10 to 12 weeks, involves a trained behavior coach working with the classroom teacher to learn and systematically apply strategies for eliciting and reinforcing the student's positive behaviors (e.g., cooperation, self-regulation). The coach implements the initial 5 program days, modeling for the classroom teacher and peers the techniques and strategies to elicit and support positive behavior. In the classroom, the coach provides feedback and monitors the student's behavior using visual cues (i.e., a green-colored card to indicate on-task behavior and a red-colored card for inappropriate behavior) and tallies points for positive behavior during timed intervals. Each program day has performance criteria that must be met before proceeding to the next program day. If the reward criteria are met, the student earns both a classroom and a home privilege that was prearranged with the teacher and parents (e.g., a brief free-time activity). If the criteria are not met, that program day is repeated or the student is "recycled" to an earlier program day before proceeding.

After Program Day 5, the classroom teacher is responsible for monitoring and responding to behavior according to First Step protocols, with daily coach supervision and support. Gradually, the interval in which a student can earn points and praise is extended, until eventually, the target student must work in blocks of multiple days to earn a single reward of higher magnitude. Thus, the program becomes more demanding, requiring the student to sustain acceptable performance for longer and longer periods to be successful.

The First Step home module, HomeBase, is designed to enable parents and caregivers to build child competencies and skills in areas that affect school adjustment and performance. During six 1-hr weekly HomeBase lessons, the behavior coach works with parents to enhance skills involving (a) communicating and sharing at school, (b) eliciting cooperation, (c) setting limits, (d) problem solving, (e) encouraging friendships, and (f) building confidence. The HomeBase module includes lesson plans, instructional guidelines, and parent–child activities for directly teaching skills to children. The coach begins HomeBase after the participating child has completed classroom Program Day 10 by visiting the parents' home or meeting them at another convenient location. At the end of each session, the coach leaves materials with parents to support their review and practice of each skill with their child daily for 10 to 15 min.

FIRST STEP'S EVIDENCE BASE

Development of the evidence base for First Step spans more than 2 decades and includes positive findings from single-subject designs conducted with students of diverse backgrounds (Beard & Sugai, 2004; Golly, Sprague, Walker, Beard, & Gorham, 2000) and multiple-group design studies conducted primarily in Oregon schools (Golly, Stiller, & Walker, 1998; Walker, Golly, McLane, & Kimmich, 2005; Walker et al., 1998, 2014). First Step also has been studied in conjunction with other behavior strategies (Carter & Horner, 2007, 2009) and interventions in tiered systems of support (Nelson et al., 2009), and it has been adapted to preschool settings (Frey, Faith, Elliot, & Royer, 2006; Gunn, Feil, Seeley, Severson, & Walker, 2006). Randomized controlled studies have further shown positive results on students' behavior and academic engagement at posttest (Walker et al., 1998, 2009). Walker et al. (2009) reported results from a large-scale randomized controlled trial of First Step with 200 first- through third-graders with externalizing behavior problems conducted in a diverse urban school district in the Southwest. Outcome data were collected from teacher and parent surveys, classroom observations, and direct student academic assessments at baseline and postintervention. The results, after variance in the baseline measures was controlled for, indicated that First Step students had significantly greater gains in symptom improvement, functioning, and academic competence, with effect sizes ranging from $d = 0.44$ to 0.87. Details of the study's methodology, which mirror the methods of the current study, can be found in the article by Walker et al. (2009).

STUDY PURPOSE

The purpose of this study was to follow up previous efficacy research to examine sustained effects of First Step. The follow-up study addressed an increasingly salient question, as researchers and policymakers concurred that finding positive effects of an intervention at its conclusion is not a sufficient basis for it to be taken to scale. In that vein, a committee of the Society for Prevention Research (SPR) established standards for identifying effective prevention programs that go beyond measuring effects immediately at postintervention to include showing effects in randomized controlled trials with long-term follow-up research. In fact, the SPR recommends "multiple follow-ups to examine the nature of the time-course of the program effects" (Flay et al., 2005, p. 161).

In line with these standards, this study used follow-up data from the identical sample of early elementary school participants from the First Step efficacy study reported earlier (Walker et al., 2009) to address two questions: (a) What effects on student behavior and academics did First Step achieve and sustain 1 year postintervention? (b) What were the differences in outcomes over time for First Step students exposed to high versus low implementation fidelity?

METHOD

Participants

The study by Walker et al. (2009), from which we report follow-up results here, was the result of a collaboration between the First Step development team at Oregon Research Institute (ORI) and evaluators at SRI International. ORI sought to assess the fidelity and impact of First Step within an ethnically and linguistically diverse school district. Thus, the team recruited 202 first- through third-grade general education teachers and 202 students and their parents from 34 public schools in a large Southwestern urban district that served more than 94,000 students. About 13% of those students received special education services, 57% were Hispanic, and 40% were eligible for free or reduced-price lunches.

Measures

Outcomes

Ten measures were used to address the outcome domains of prosocial/adaptive behavior, problem/maladaptive behavior, and academic performance. ORI researchers collected outcome data for intervention and comparison students at baseline, immediately on completion of First Step (posttest), as well as 1 year after First Step completion (follow-up). Because students had advanced to a new grade by the time the follow-up measures were collected, a different teacher completed these assessments than had completed the pretest and posttest measures; however, in all cases, the same parent or guardian was asked to complete each of the parent-rated measures.

Systematic Screening for Behavior Disorders

In addition to serving as a baseline screening tool, the Systematic Screening for Behavior Disorders (SSBD) Adaptive Behavior Index (ABI; $\alpha = 0.82$) and Maladaptive Behavior Index (MBI; $\alpha = 0.84$) also measured outcomes at posttest and follow-up. The SSBD has excellent psychometric characteristics, is nationally normed, and has been used in a number of research studies for both screening and outcome measures. For example, Lane, Menzies, Oakes, and Kalberg (2012) recently documented how the SSBD could be used to enhance and evaluate the impact of behavioral–academic interventions, and other studies found positive convergent validity between the SSBD and other outcome measures, including the Scale for Assessing Emotional Disturbance (Epstein & Cullinan, 1998) and the Behavioral and Emotional Rating Scale (Epstein & Sharma, 1998). Furthermore, the SSBD indices have documented statistically significant gains produced by First Step (Walker et al., 2009), showing effect sizes of 0.82 (ABI) and 0.87 (MBI). The strong psychometric properties and relative sensitivity of the SSBD indices justify their general use as outcome measures for short-term interventions with general education students in classroom settings.

Academic Engaged Time

Researchers who had been trained to high reliability (minimum interobserver agreement of 0.80) and were blind to group assignments directly observed students in intervention and comparison classrooms to collect measures of academic engaged time (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), as reported by Walker et al. (2009). Using a stopwatch recording procedure, observers documented the proportion of time during two 15-min intervals (collected on different days within 1 week of one another) in which students attended to teacher instructions and academic tasks, made relevant motor responses, and appropriately interacted with other teachers and peers. To minimize the effects of varying classroom contexts to the extent possible, observers collected AET data at the same time of day and during similar classroom activities across all data collection periods. Furthermore, across the waves of data collection, reliability estimates were collected on 33% of the recorded AET observations. Observers were retrained as necessary to minimize drift and ensure adequate reliability of recorded observations. The overall intraclass correlation (ICC) of AET interrater reliability was excellent (ICC[3,1] = 0.80).

Social Skills Rating System

The Social Skills Rating System (SSRS) is a standardized, norm-referenced instrument with strong psychometric properties: content, construct, and concurrent validity has been researched extensively (Nangle, Hansen, Eardley, & Norton, 2009). Teachers completed three subscales of the SSRS–Teacher Version (Gresham & Elliott, 1990) that measured students' social skills (SSRS-SS-T, $\alpha = 0.88$), problem behaviors (SSRS-PB-T, $\alpha = 0.85$), and academic competence (SSRS-AC-T, $\alpha = 0.91$; Gresham & Elliott, 1990). Parents completed the SSRS–Parent Version of the social skills subscale (SSRS-SS-P, $\alpha = 0.88$) and problem behavior subscale (SSRS-PB-P, $\alpha = 0.88$; Gresham & Elliott, 1990).

Woodcock-Johnson III Letter–Word Identification

The Woodcock-Johnson III is a norm-referenced assessment with strong psychometric properties (Cizek, 2001). To assess students' abilities to identify isolated letters and words, researchers administered the Letter–Word Identification (LWI) subtest ($\alpha = 0.91$) from the Woodcock-Johnson III Diagnostic Reading Battery (WJ LWI; Woodcock, Mather, & Schrank, 2004). Students' standard scores, reported here, reflect each student's level of performance compared with the general population of students of his or her same grade level and age.

Oral Reading Fluency

Researchers computed an oral reading fluency (ORF) score based on the average number of words read correctly by a student in 1 min from two different 300- to 400-word first-grade level reading passages previously used in national studies (Fuchs, 2003).

Table 1 shows the mean and standard deviation of each outcome measure at baseline, posttest, and follow-up for intervention and comparison students. Figures 1 and 2 show the mean values and standard errors for the behavior and academic outcome measures, respectively, at the three measurement times.

Fidelity

ORI researchers used the Implementation Fidelity Checklist (IFC, $\alpha = 0.94$; Walker et al., 2009) to document the extent to which the coach and teacher delivered First Step components as intended and with high quality. Researchers observed the implementer in the classroom three times: once for the coach on or around Program Day 5 and three times for the teacher on or around Program Days 10, 17, and 24. Observers rated the implementer on 18 intervention components (e.g., whether the implementer announced the number of points needed for a reward, elicited cooperation from classroom peers, provided positive feedback, and used verbal reminders to prompt the student). For each intervention component, observers rated implementation adherence (*yes* or *no*) and rated quality on a 5-point scale (1, *very poor/not delivered*; 2, *poor*; 3, *okay*; 4, *good*; 5, *excellent*). The ICC assessing interrater reliability for 33% of the fidelity observations was excellent (ICC[3,1] = 0.94). A classroom fidelity score was calculated as the average quality score across the 18 components and 4 observation periods.

Procedures

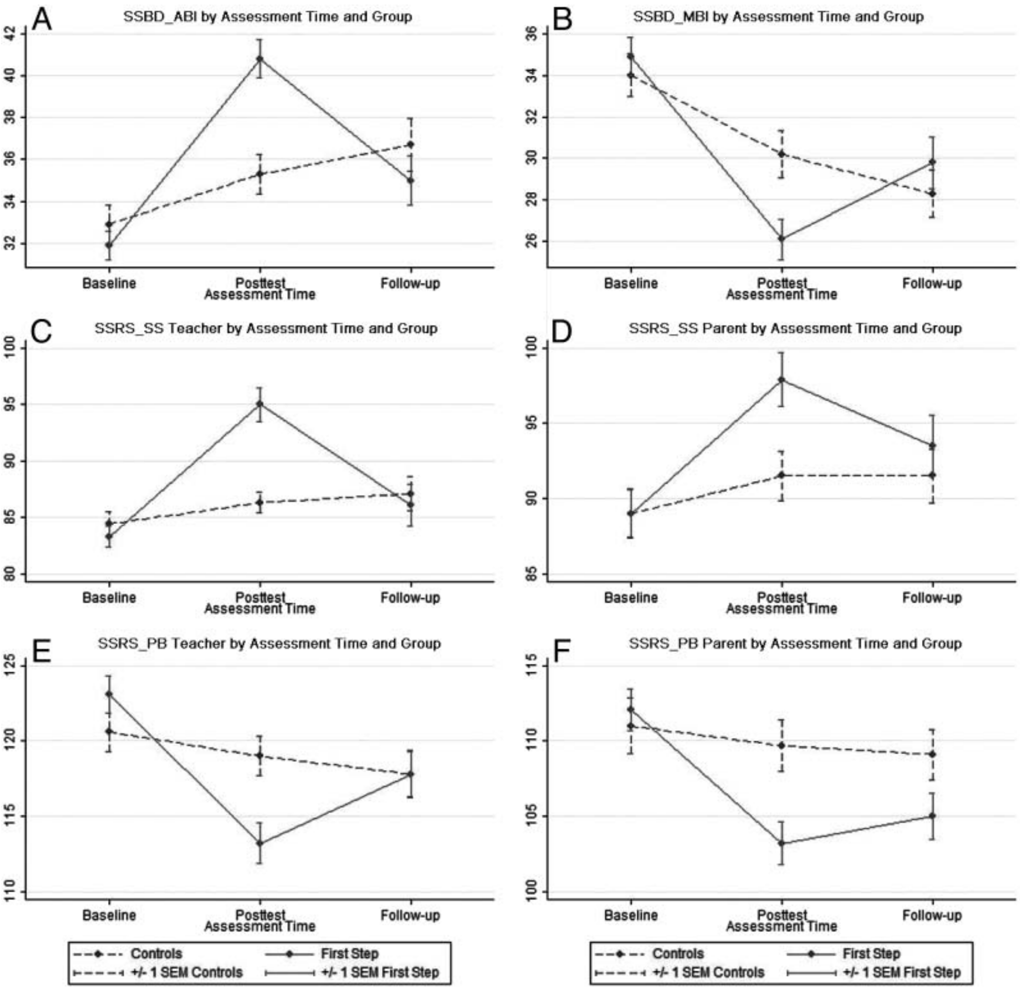
Teachers were randomly assigned to intervention or comparison groups within two cohorts (the 2005–2006 and 2006–2007 school years). Using screening methods described later, teachers identified eligible students for study participation. Comparison-group teachers continued to use their typical instructional and classroom management techniques and to refer students to additional services as available and warranted, with no support other than what the school typically offered. Results of group-equivalence analyses (Woodbridge et al., 2010) indicated that First Step students were not significantly different from comparison-group students on measures of demographic characteristics, school factors, or baseline behavioral or academic measures.

Table 1. Means and Standard Deviations for Baseline, Posttest, and Follow-Up Measures

	Baseline [*M* (*SD*)]		Posttest [*M* (*SD*)]		Follow-Up [*M* (*SD*)]	
	Control	Intervention	Control	Intervention	Control	Intervention
Prosocial/adaptive behavior						
SSBD Adaptive Behavior Index	32.9 (7.8)	31.9 (6.6)	35.3 (7.4)	40.8 (9.1)	36.7 (9.8)	35.0 (10.0)
SSRS Social Skills subscale–Teacher	84.4 (10.1)	83.3 (8.6)	86.3 (8.8)	95.0 (14.4)	87.1 (13.3)	86.1 (15.7)
SSRS Social Skills subscale–Parent	89.0 (14.5)	89.0 (14.8)	91.5 (15.1)	97.9 (15.7)	91.5 (15.8)	93.5 (17.9)
Problem/maladaptive behavior						
SSBD Maladaptive Behavior Index	34.0 (8.6)	34.9 (7.9)	30.2 (9.3)	26.1 (9.3)	28.3 (9.6)	29.8 (10.8)
SSRS Problem Behavior subscale–Teacher	120.6 (11.0)	123.1 (10.3)	119 (10.7)	113.2 (12.6)	117.8 (12.4)	117.8 (13.4)
SSRS Problem Behavior subscale–Parent	111 (15.5)	112.1 (13.7)	109.7 (13.5)	103.2 (13.8)	109.1 (14.3)	105.0 (13.9)
Academic						
SSRS AC subscale–Teacher	88.5 (11.6)	88.6 (10.2)	87.6 (10.9)	91.0 (10.5)	89.3 (11.3)	89.0 (10.9)
Academic Engaged Time	40.6 (18.5)	42.4 (18.5)	48.5 (22.1)	56.9 (19.5)	65.6 (18.8)	65.0 (20.4)
WJ III LWI	97.6 (15.9)	100.3 (12.4)	100.0 (16.0)	101.0 (12.8)	97.1 (15.7)	97.7 (11.9)
Oral reading fluency	47.0 (36.7)	56.1 (41.5)	53.9 (38.3)	64.2 (43.3)	75.9 (42.4)	86.3 (40.3)

Abbreviations: AC, Academic Competence; SSBD, Systematic Screening for Behavior Disorders; SSRS, Social Skills Rating System; WJ III LWI, Woodcock Johnson III Letter–Word Identification subtest.

Figure 1. Behavior Outcomes by Assessment Time and Group (a-f)

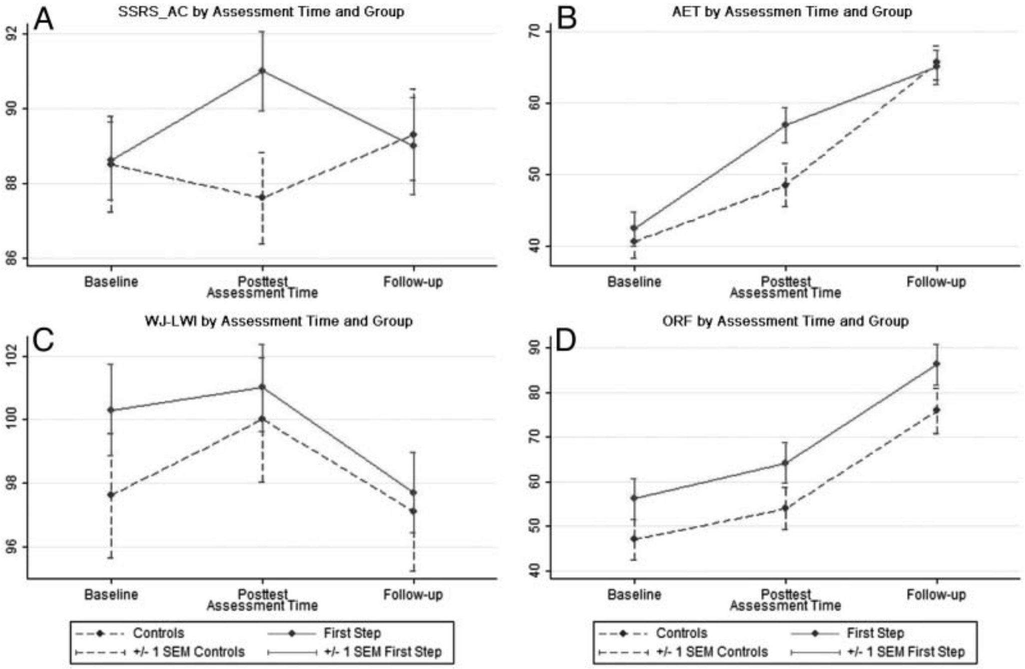


Abbreviations: ABI, Adaptive Behavior Index; MBI, Maladaptive Behavior Index; PB, Problem Behavior subscale; SS, Social Skills subscale; SSBD, Systematic Screening for Behavior Disorders; SSRS, Social Skills Rating System.

ORI's role in the trial involved training behavior coaches and teachers in the intervention protocol, monitoring and supporting implementation fidelity (e.g., by providing consultation to participating teachers), and collecting outcome data. Coaches were in close contact with ORI supervisory staff and were themselves scheduled to regularly undergo fidelity monitoring checks to review their adherence to the First Step implementation protocol.

We determined that there would be minimal risk of contamination in this design from teachers in different groups exchanging First Step behavioral procedures or materials because (a) First Step is a manualized intervention that requires a coach's training and support in the classroom for a prescribed period, (b) no intervention or comparison students were placed in the same classroom, and (c) teachers participating in First Step in the first cohort were not assigned to the comparison condition in the second cohort. Furthermore, through the duration of the study, the ORI team did not report any contaminants that differentially affected classrooms.

Figure 2. Academic Outcomes by Assessment Time and Group (a-d)



Abbreviations: AC, Academic Competence subscale; AET, academic engaged time; ORF, oral reading fluency; SSRS, Social Skills Rating System; WJ-LWI, Woodcock-Johnson III Letter–Word Identification subtest.

After becoming familiar with students for at least 30 days, teachers reviewed descriptions of externalizing behaviors (e.g., displaying aggression, arguing, disturbing others, fighting) and nominated five students in the class who exhibited the highest levels of such behaviors. For the three highest-ranked students, teachers completed brief ratings of students' behaviors. Teachers in both intervention and comparison classrooms used an abbreviated version of the SSBD (Walker & Severson, 1990) composed of the (a) ABI, 12 items (e.g., follows classroom rules, cooperates with peers) rated on a 5-point Likert scale ranging from 1 (*never*) to 5 (*frequently*); (b) MBI, 11 items (e.g., refuses to participate, creates a disturbance) rated on the same 5-point Likert scale; and (c) Critical Events Index (CEI), which indicated how many of 30 high-saliency, low-frequency indicators (e.g., stealing, physical aggression) occurred during the preceding 30 days for each student.

The student with the highest combined score on the ABI, MBI, and CEI was the first to be invited to participate in the study. If two students in a classroom had the same score, the student with the higher raw CEI score was selected first. If the parents of that student did not consent to their child's participation in the study, consent was sought from the parents of the next highest-scoring student. Although the SSBD has an optional Stage 3 that involves classroom and playground observations, it was deemed too labor intensive for the purposes of this study.

Statistical Analysis

Data Imputation

To avoid attrition and potential bias, we imputed missing values using fully conditional specification models (i.e., logistic, polytomous, or linear regression) applied iteratively via Stata's imputation–by–chained equations procedure (Royston, 2004; Van Buren, Brands, Groothuis-Oudshoorn, & Rubin, 2006). Variables used in the imputation–by–chained equations included the baseline, posttest, and follow-up values for the 10 outcome measures; student-specific variables included student age, grade level, gender, race or ethnicity, language, English language learner status, free and reduced-price lunch program status, and special education status; teacher-specific variables included a summary measure of teacher self-reported knowledge and skills in working with students with behavior problems (Cheney, Walker, & Blum, 2004); and school-specific variables included a school mobility score, suspension rate, and summary measure of the presence of school-wide positive behavior support on assessment with the School-wide Evaluation Tool (Horner et al., 2004). These covariates were used in previous analyses of the main effects of First Step (Sumi et al., 2013); their influence on the imputed values was small but not negligible.

The percentage of data that were imputed ranged from 4% to 10% for baseline, from 5% to 9% for posttest, and from 14% to 26% for follow-up measures. Twenty imputations were conducted separately for intervention- and comparison-group students for each missing value. These results were combined to provide estimates of the variability and p values for regression coefficients.

Analysis of Intervention Effects

In the previous First Step efficacy study, Walker et al. (2009) applied multivariate analysis of covariance with baseline values as covariates to examine posttest results. Our analysis used hierarchical linear modeling (HLM) to examine outcomes at all three observation intervals (baseline, posttest, and follow-up) in the same regression. This nested model accounted for uncertainty in the baseline measures and also used centered fidelity checklist (IFC) values as covariates. We analyzed posttest results for reference only because these results were reported by Walker et al. (2009); however it should be noted that effect size estimates in the previous study were larger in absolute value than those in our study. Differences in effect size values ranged from 0.03 on the SSRS-PB-T (effect size of −0.73 in the previous study versus −0.70 in this study) to 0.36 on the SSRS-AC-T (effect size of 0.66 versus 0.30).

We performed HLM regressions with repeated measures on each set of imputed data to estimate intervention effects at posttest and follow-up and to estimate the effects of implementation fidelity. The model included independent terms for time, treatment, and fidelity and variance terms for school, student nested within school, and time nested within student. Each student was associated with baseline, posttest, and follow-up values for the dependent variables (either observed or imputed), as shown in the following model:

$$\begin{aligned} Y_{ijk} = B_0 + B_1T_{ij} + B_2(J_1 + J_2) + B_3T_{ij}(J_1 + J_2) \\ + B_4J_2 + B_5T_{ij}J_2 + B_6T_{ij}(J_1 + J_2) \\ F_{ij} + B_7T_{ij}J_2F_{ij} + e_j + e_{ij} + e_{ijk} \qquad (1) \end{aligned}$$

In this model, i is an index for student within School j; j is an index for school; k is an index for time, with 0 indicating baseline, 1 indicating posttest, and 2 indicating follow-up; Y_{ijk} is the value of the dependent variable for the ith student in the jth school at Time k; J_k is an indicator for Time k; T_{ij} is an indicator for Student i,j being in the intervention group; F_{ij} is the centered IFC fidelity value for student (i,j), which was centered among intervention students and takes values of 0 for comparison students; e_j represents residual variability between schools; e_{ij} represents residual variability between students within schools; e_{ijk} represents residual variability between times within students (also including measurement error); B_0 is the expected mean for comparison students at Time 0; B_1 is the difference between intervention and comparison students at baseline; B_2 is the increase in outcome between baseline and posttest for comparison students; B_3 is the effect of intervention on change in outcome from baseline to posttest; B_4 is the increase in outcome between posttest and follow-up for comparison students; B_5 is the effect of intervention on change in outcome from posttest to follow-up; B_6 is the effect of fidelity on intervention students from baseline to posttest; and B_7 is the effect of fidelity on intervention students from posttest to follow-up. The terms with coefficients B_6 and B_7 were only used in analyses to examine the effect of fidelity.

A simplified version of the full HLM model was used to examine the statistical significance of changes in one of the two groups over time, fit using only comparison-group data or only intervention-group data:

$$Y_{ijk} = B_0 + B_2(J_1 + J_2) + B_4J_2 + e_j + e_{ij} + e_{ijk} \qquad (2)$$

Intervention effects were examined for multiple outcomes in three domains: (a) prosocial/adaptive behavior (ABI, SSRS-SS-T, SSRS-SS-P); (b) problem/maladaptive behavior (SSBD-MBI, SSRS-PB-T, SSRS-PB-P); and (c) academic (SSRS-AC-T, AET, WJ LWI, ORF). Because we performed multiple tests for intervention effects, we applied the Benjamini-Hochberg correction for Type I error rate (Schochet, 2008) within each domain (i.e., for a given test, the reported *p* value was the smallest false discovery rate value for which the corresponding null hypothesis was rejected). Effect sizes are reported using a statistic analogous to Cohen's (1988) *d*, calculated by dividing the treatment indicator coefficient by an estimate of the pooled between-student standard deviation at posttest or follow-up.

RESULTS

Research Question 1: Effects Achieved at Posttest and Sustained at Follow-Up

Posttest Effects

Table 2 summarizes the treatment coefficients at posttest and follow-up using the statistical model described earlier, excluding the fidelity variables. As shown in the *Baseline to Posttest* columns, First Step students achieved significantly greater improvements than comparison students on all behavior measures, with positive treatment coefficients for adaptive behavior and negative treatment coefficients for maladaptive behavior measures (significant effects in a beneficial direction for the intervention group are indicated by an up arrow). First Step students also achieved positive and significant effects for SSRS-AC-T, both before and after adjustment for multiple tests, and for AET before adjustment.

Sustained Effects

As shown in the *Posttest to Follow-Up* columns of Table 2, First Step posttest effects were not sustained through the follow-up measurement point. In fact, First Step students differed significantly from comparison students in the undesired direction on five of the six behavior measures (the exception being SSRS-PB-P), with negative treatment coefficients for adaptive behavior and positive treatment coefficients for maladaptive behavior measures. Treatment effects also were negative and statistically significant for the academic measures of SSRS-AC-T and AET (significant effects in an adverse direction for the intervention group are indicated by a down arrow). As shown in the *Baseline to Follow-Up* columns, the overall change from baseline to follow-up for SSRS-PB-P was the only statistically significant effect after adjustment for multiple comparisons.

To interpret the scoring pattern that resulted in these findings, we present the parameters of the HLM regression that correspond to the change from baseline to posttest, posttest to follow-up, and baseline to follow-up for both groups and their associated standard errors in Table 3. Withdrawal of intervention was associated with First Step students' scores deteriorating significantly on five measures of behavior from posttest to follow-up—decreasing on the SSBD-ABI (–5.7 points, $p < .001$), SSRS-SS-T (-8.9 points, $p < .001$), and SSRS-SS-P (-4.4 points, $p < .01$) and increasing on the SSBD-MBI (3.7 points, $p < .01$) and SSRS-PB-T (4.5 points, $p < .01$). In contrast, the comparison group generally maintained behaviors at posttest levels. In the academic domain, the comparison and First Step groups had similar point decreases on the WJ LWI (-2.9 and -3.3, respectively; $p < .001$) and similar increases on the ORF (22.0 and 22.1, respectively; $p < .001$). On the AET, First Step students showed an increase

Table 2. HLM Results and Effect Sizes for First Step at Postintervention Time Points

	Baseline to Posttest Measurement Period		Posttest to Follow-Up Measurement Period		Baseline to Follow-Up Measurement Period	
Domain/Measure	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Effect Size [Adjusted *p* Value[a]]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Effect Size [Adjusted *p* Value[a]]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Effect Size [Adjusted *p* Value[a]]
Prosocial/adaptive behavior						
SSBD-ABI	6.39 (1.40) [.001]	0.742 [.001] ↑	–7.11 (1.59) [.001]	–0.827 [.001] ↓	–0.73 (1.57) [.645]	–0.084 [.956]
SSRS-SS-T	9.84 (2.10) [.001]	0.803 [.001] ↑	–9.72 (2.25) [.001]	–0.794 [.001] ↓	0.12 (2.24) [.956]	0.010 [.956]
SSRS-SS-P	6.41 (1.98) [.001]	0.407 [.001] ↑	–4.45 (2.22) [.046]	–0.282 [.046] ↓	1.96 (2.24) [.384]	0.124 [.956]
Problem/maladaptive behavior						
SSBD-MBI	–4.98 (1.49) [.001]	–0.531 [.001] ↑	5.60 (1.62) [.001]	0.598 [.001] ↓	0.63 (1.62) [.700]	0.067 [.700]
SSRS-PB-T	–8.26 (1.94) [.001]	–0.697 [.001] ↑	5.78 (2.09) [.006]	0.488 [.007] ↓	–2.48 (2.10) [.239]	0.209 [.359]
SSRS-PB-P	–7.67 (1.93) [.001]	–0.541 [.001] ↑	2.48 (2.05) [.226]	0.175 [.226]	–5.19 (2.08) [.013]	–0.366 [.039] ↑
Academic achievement						
SSRS-AC-T	3.31 (1.30) [.011]	0.303 [.029] ↑	–3.64 (1.37) [.008]	–0.333 [.029] ↓	–0.33 (1.39) [.810]	0.031 [.810]
AET	6.59 (3.27) [.044]	0.334 [.085]	–8.99 (3.49) [.010]	–0.456 [.029] ↓	–2.40 (3.42) [.483]	–0.122 [.810]
WJ III LWI	–1.76 (0.911) [.053]	–0.123 [.085]	–0.388 (1.00) [.698]	–0.027 [.798]	–2.15 (1.01) [.034]	–0.151 [.136]
ORF	1.22 (2.74) [.656]	0.030 [.798]	0.092 (2.97) [.975]	0.002 [.975]	1.31 (2.98) [.660]	0.032 [.810]

Note. An up arrow indicates a significant beneficial effect for the intervention group; a down arrow indicates a significant adverse effect for the intervention group.
Abbreviations: ABI, Adaptive Behavior Index; AC, Academic Competence subscale; AET, academic engaged time; HLM, hierarchical linear modeling; MBI, Maladaptive Behavior Index; ORF, Oral reading fluency; P, parent; PB, Problem Behavior subscale; SS, Social Skills subscale; SSBD, Systematic Screening for Behavior Disorders; SSRS, Social Skills Rating System; T, teacher; WJ III LWI, Woodcock-Johnson III Letter–Word Identification subtest.

[a] *p* Value after applying Benjamini-Hochberg correction for multiple comparisons.

Table 3. HLM Results for Separate Comparison and Intervention Analyses

	Baseline to Posttest Measurement Period		Posttest to Follow-up Measurement Period		Baseline to Follow-up Measurement Period	
Group	Comparison	Intervention	Comparison	Intervention	Comparison	Intervention
Domain/Measure	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]	Treatment Coefficient (*SE*) [Unadjusted *p* Value]
Prosocial/adaptive behavior						
SSBD-ABI	2.46 (1.01) [.015]	8.85 (0.971) [.001]	1.38 (1.17) [.239]	–5.73 (1.10) [.001]	3.84 (1.16) [.001]	3.12 (1.10) [.005]
SSRS-SS-T	1.89 (1.33) [.155]	11.73 (1.62) [.001]	0.837 (1.48) [.566]	–8.88 (1.75) [.001]	2.72 (1.45) [.061]	2.84 (1.75) [.105]
SSRS-SS-P	2.52 (1.39) [.071]	8.93 (1.42) [.001]	0.032 (1.51) [.983]	–4.42 (1.42) [.008]	2.55 (1.52) [.094]	4.51 (1.63) [.006]
Problem/maladaptive behavior						
SSBD-MBI	–3.86 (1.07) [.001]	–8.83 (1.03) [.001]	–1.91 (1.16) [.100]	3.70 (1.16) [.002]	–5.76 (1.17) [.001]	–5.14 (1.14) [.001]
SSRS-PB-T	–1.60 (1.34) [.235]	–9.86 (1.41) [.001]	–1.27 (1.48) [.394]	4.52 (1.51) [.003]	–2.86 (1.47) [.053]	–5.35 (1.52) [.001]
SSRS-PB-P	–1.27 (1.37) [.357]	–8.94 (1.36) [.001]	–0.63 (1.46) [.668]	1.86 (1.50) [.216]	–1.89 (1.44) [.190]	–7.08 (1.51) [.001]
Academic						
SSRS-AC-T	–0.91 (0.97) [.349]	2.40 (0.86) [.005]	1.66 (1.05) [.112]	–1.98 (0.99) [.046]	0.75 (1.05) [.474]	0.42 (0.99) [.674]
AET	7.97 (2.30) [.001]	14.55 (2.33) [.001]	17.06 (2.50) [.001]	8.07 (2.43) [.001]	25.03 (2.44) [.001]	22.63 (2.42) [.001]
WJ III LWI	2.40 (0.70) [.001]	0.63 (0.58) [.276]	–2.92 (0.74) [.001]	–3.31 (0.64) [.001]	–0.52 (0.75) [.488]	–2.68 (0.64) [.001]
ORF	6.87 (2.02) [.001]	8.09 (1.86) [.001]	21.96 (2.19) [.001]	22.05 (2.02) [0.001]	28.83 (2.18) [0.001]	30.14 (2.02) [0.001]

Abbreviation: ABI, Adaptive Behavior Index; AC, Academic Competence subscale; AET, academic engaged time; HLM, hierarchical linear modeling; MBI, Maladaptive Behavior Index; ORF, oral reading fluency; P, parent; PB, Problem Behavior subscale; SS, Social Skills subscale; SSBD, Systematic Screening for Behavior Disorders; SSRS, Social Skills Rating System; T, teacher; WJ III LWI, Woodcock-Johnson III Letter–Word Identification subtest.

of 8.1 points ($p < .001$) but the comparison group showed a larger increase (17.0 points, $p < .001$). These changes in academic measures eliminated the significant differences in SSRS-AC-T and AET between groups that were present at posttest.

We also examined the ICC at the school level because the extent to which outcomes varied by school has potential implications for understanding the generalizability of the intervention. Across all outcome measures and time points, the proportion of variance attributable to school ranged between 0.18 and 0.31, with comparable average ICCs at baseline (0.25), posttest (0.25), and follow-up (0.23).

Research Question 2: Relationships Between Fidelity and Student Outcomes

The mean classroom fidelity score was calculated as 3.73 (i.e., between *okay* and *good* on the 5-point scale), with an *SD* of 0.53 and a range of 2.23 to 4.86. Although overall fidelity ratings were within an acceptable range, we assessed whether First Step students whose intervention was implemented with higher fidelity (i.e., adherence and quality) achieved better outcomes than students whose intervention was delivered with lower fidelity.

Table 4 shows the relationship of implementation fidelity to changes in student outcomes from baseline to posttest and from posttest to follow-up. The only statistically significant effect among the 30 relationships tested was a negative relationship between implementation fidelity and students' academic engagement from posttest to follow-up (a 1–standard deviation increase in fidelity lowered students' academic engagement by almost 4 standard deviations, $p = .016$).

DISCUSSION

Effects of First Step at Follow-Up

The criteria for establishing the effectiveness of an intervention, as set forth by the SPR, include a rigorous demonstration of effects on a representative sample on outcomes measured immediately and at least 6 months after the intervention. Previous research, including multiple randomized trials, showed First Step's positive effects on behavior and academic outcomes at posttest. However, the analyses of follow-up data from a recent efficacy trial reported here show that First Step did not meet the criterion for duration of effect. That is, the significant positive effects of First Step at the conclusion of the prescribed intervention were no longer evident 1 year after intervention on measures of student behavior or most academic outcomes.

The closing of the gaps in behavioral outcomes resulted primarily from First Step students' scores declining with the withdrawal of the intervention and transition to a new teacher/grade level, whereas comparison students' scores remained fairly stable. In contrast, no clear pattern is evident for academic measures. On the measure of academic competence, the intervention group's loss and the comparison group's gain from posttest to follow-up eliminated the gap that initially favored First Step students at posttest. The comparison group's growth in academic engagement also outstripped the posttest advantage of First Step students. On the WJ LWI, the comparison group showed posttest growth but both groups declined by follow-up to levels equal to or below baseline. However, similar increases in ORF maintained the parity of the two groups that was apparent at posttest through to follow-up.

Findings based on parent perspectives are especially intriguing because, unlike teachers, the respondents are consistent across periods. On these measures of behavior, parents of comparison-group students indicated their children showed generally steady (but nonsignificant) improvements in problem behavior and social skills from baseline to follow-up. In contrast, parents of First Step students reported significant behavioral improvements immediately after intervention, including a reduction in problem behaviors at follow-up, but ratings indicated an erosion of gains in social skills from posttest to follow-up.

The lack of sustainment of First Step's effects (and that of some other short-term interventions) should be considered in the theoretical context of the intervention. One theory

Table 4 HLM Coefficients for Impact of Implementation Fidelity on Outcomes at Postintervention Time Points

	Baseline to Posttest Measurement Period		Posttest to Follow-Up Measurement Period		Baseline to Follow-Up Measurement Period	
Domain/Measure	Fidelity Coefficient (*SE*) [Unadjusted *p* Value]	ES per 1-*SD* Fidelity Increase [Adjusted *p* Value[a]]	Fidelity Coefficient (*SE*) [Unadjusted *p* Value]	ES per 1-*SD* Fidelity Increase [Adjusted *p* Value[a]]	Fidelity Coefficient (*SE*) [Unadjusted *p* Value]	ES per 1-*SD* Fidelity Increase [Adjusted *p* Value[a]]
Prosocial/adaptive behavior						
SSBD-ABI	2.08 (1.55) [.181]	0.129 [.866]	–2.25 (2.22) [.313]	–0.139 [.866]	–0.17 (2.07) [.936]	–0.105 [.936]
SSRS-SS-T	–0.491 (2.30) [.831]	–0.021 [.866]	1.54 (3.28) [.639]	0.067 [.866]	1.05 (2.87) [.715]	0.046 [.936]
SSRS-SS-P	–1.00 (2.36) [.673]	–0.033 [.866]	–0.501 (2.96) [.866]	–0.017 [.866]	–1.50 (2.73) [.584]	–0.050 [.936]
Problem/maladaptive behavior						
SSBD-MBI	–1.42 (1.68) [.398]	–0.080 [.710]	2.30 (2.28) [.315]	0.129 [.710]	0.87 (2.06) [.672]	–0.049 [.672]
SSRS-PB-T	–1.58 (2.20) [.473]	–0.070 [.710]	3.38 (3.02) [.264]	0.151 [.710]	1.80 (2.55) [.482]	0.080 [.672]
SSRS-PB-P	0.674 (2.26) [.766]	0.025 [.766]	1.16 (2.72) [.670]	0.043 [.766]	1.83 (2.45) [.454]	0.068 [.672]
Academic						
SSRS-AC-T	1.04 (1.58) [.512]	0.051 [.683]	–1.78 (1.99) [.373]	–0.086 [.662]	–0.74 (1.84) [.810]	–0.036 [.947]
AET	5.50 (3.45) [.111]	0.147 [.444]	–14.65 (4.66) [.002]	–0.393 [.016]	–2.40 (3.38) [.022]	–0.245 [.088]
WJ III LWI	1.12 (1.17) [.340]	0.041 [.662]	–1.04 (1.27) [.414]	–0.038 [.662]	0.082 (1.24) [.947]	0.003 [.947]
ORF	0.733 (3.87) [.835]	0.010 [.835]	–1.78 (3.87) [.645]	–0.023 [.737]	–1.05 (3.78) [.781]	0.003 [.947]

Abbreviations: ABI, Adaptive Behavior Index; AC, Academic Competence subscale; AET, academic engaged time; ES, effect size; HLM, hierarchical linear modeling; MBI, Maladaptive Behavior Index; ORF, oral reading fluency; P, parent; PB, Problem Behavior subscale; SS, Social Skills subscale; SSBD, Systematic Screening for Behavior Disorders; SSRS, Social Skills Rating System; T, teacher; WJ III LWI, Woodcock-Johnson III Letter–Word Identification subtest.

[a] *p* Value after applying Benjamini-Hochberg correction for multiple comparisons.

might posit that the mechanisms of change (e.g., change in teacher behavior) will produce change in child behavior, which is then internalized by the target child and thus sustained through naturally reinforcing contingencies. Unfortunately, long-term findings clearly do not support this prevention theory. Rather, First Step teachers learned effective strategies for eliciting and reinforcing the student's positive behaviors within their current classroom environments, but the program did not show prevention outcomes across years in this study. Students' participation in 30 program days may not have been sufficient to produce the theoretical chain of events resulting in the sustainability of intervention effects within new environments where First Step was not implemented.

Two phenomena may explain both the absence of maintained gains in First Step students' behavior and the greater improvement in comparison students' behavior. First, the pattern might be explained by the change in the behavioral environment of students' classrooms in the year after intervention. That is, by posttest, First Step teachers and classroom peers had learned how to elicit positive behavior from the target students, behavior that was reflected in improvements in students' behavioral assessments. However, the following year's teachers and classroom peers had no such training. In the absence of the reinforcement for appropriate behavior that First Step students had come to expect, their behavior may have reverted to prior patterns of poor behavior.

Second, posttest-to-follow-up improvements in comparison students' behavior may reflect different teacher and peer perceptions and expectations. Disruptive students can suffer reputational bias, where teachers and peers pejoratively judge the student's behavior even after it undergoes improvements in the short term (Hollinger, 1987; Maag, Vasa, Kramer, & Torrey, 1991). The students in the study were initially selected by the implementation-year teachers as those with the most severe externalizing behavior problems in the classroom. As students matriculated from the implementation to the follow-up year, teachers and students at those subsequent grade levels may have had different perceptions and expectations for behavior than those in the previous year's lower grades (i.e., a different composition of students in a class may influence a teacher's perspective on severe behavior). Biases may have caused a negative reputation to persist over time despite any beneficial response the First Step students initially showed. Generalization and maintenance of social skill interventions requires embedding the target students in peer social systems and classroom structures that continue to support the learned behaviors (Farmer, Van Acker, Pearl, & Rodkin, 1999; Maag, 2006).

Each of these contextual factors may account for some degree of variance in the deterioration of effects in First Step outcomes across school years, and cumulatively, these phenomena can be quite powerful. As a selected intervention program, First Step might produce sustained effects if implemented in the context of an effective Tier 1, universal prevention program, but this hypothesis warrants further investigation.

There also are lessons to be learned from other psychosocial programs that have shown impressive long-term outcomes. Programs with long-term effects share particular characteristics that short-term applied interventions like First Step do not. Namely, they (a) span many years at full dosage with sustained implementation and (b) show prevention and intervention effects on distal outcome variables (e.g., drug and alcohol use, school failure and dropout, arrests, assignment to specialized services, health risks such as disease and pregnancy rates) rather than proximal outcomes (e.g., behavioral observations, teacher and parent ratings; Borduin et al., 1995; Conduct Problems Prevention Research Group, 2011; Eddy, Reid, & Fetrow, 2000; Hawkins, Catalano, Kosterman, Abbott, & Hill, 1999).

Relationships Between Fidelity and Effects at Follow-Up

Analyses reported here indicate that students who experienced First Step at higher fidelity also had significantly greater erosion from posttest to follow-up in intervention benefits on a single measure (AET) than students who experienced First Step implemented with lower fidelity. Although a single significant result may not indicate an overarching trend, the AET findings (based on external direct observations of student behavior, not participant perspectives) may be worthy of further investigation and discussion. A possible explanation for this pattern is that students who experienced First Step at high fidelity also experienced the greatest contrast between the intervention classroom and the more typical classroom they entered the following year. The potentially significant disruption from one year to the next in what First Step students had come to expect in the way of teachers' responses to positive and negative behavior and classroom management practices might have prompted students to revert to prior patterns of poor academic engagement. Barkley (2007) attributed this phenomenon to interventionists designing altered environments (i.e., behavioral methods and contingency management systems) to reduce behavior symptomatology that cannot be sustained with natural contingencies alone once the intervention was withdrawn. Among students who experienced First Step at lower fidelity of implementation, less contrast between the implementation-year and second-year classrooms might have prompted relatively less decline in behavior.

Future Directions

As shown by posttest results in this and previous studies, First Step procedures effectively teach students to regard their teachers' and peers' guidance and reinforcement as discriminative stimuli that elicit appropriate behavior; however, these intervention effects erode after the stimuli are withdrawn. Barkley (2007) compared behavioral interventions to medication management programs, which could produce benefits as long as they were in place but would not maintain or generalize once ceased. In fact, Barkley (2007) referred to the expectation of postextinction "bursts" (p. 280) of heightened problematic behavior on the withdrawal of positive reinforcement for their previous occurrences. The data from this study suggest the importance of working with the teachers and classmates of prior First Step students in the year after implementation to reduce the probability of worsening behavior problems and support the enduring effects of the intervention.

Limitations

Although the findings presented here were generated from a study that meets high standards for methodologic rigor, there are some limitations. For one, most behavioral outcome measures used in the study (other than the AET) were second-party reports of students' behavior, not measured through direct observation. Training and implementation involved in First Step may have affected the perceptions and, thus, the ratings of student behavior by participating teachers; still, the behavior scales used in the study have very strong psychometric qualities.

Furthermore, the absence of additional measurement intervals (e.g., 3 or 6 months after implementation, at the beginning of the school year after implementation) limits the ability to gauge the precise duration of First Step's posttest effects. If effects began to deteriorate quickly after posttest, encouraging and enabling the intervention teacher to restate expectations and reinforce positive behaviors could prolong effects. However, if effects were sustained while First Step students remained in their intervention classroom, attention to sustaining gains must shift to subsequent-year teachers. Future evaluations of First Step should include enough posttest measurement points to depict the trajectory in postintervention outcomes.

CONCLUSION

The findings presented here are the latest additions to 2 decades of research and development behind First Step. Developers' efforts continue, with the intent to strengthen and broaden the intervention, to develop further the evidence of its effectiveness, and to establish its readiness for dissemination to diverse populations of students. The First Step story underscores the reality that establishing and disseminating evidence-based practices at scale take a sustained, iterative effort over a considerable period on the part of a research community committed to finding effective, scalable solutions to important problems. Those efforts cannot be sustained without funding sources that share that commitment and understand the importance of continuity in research programs.

REFERENCES

Barkley, R. A. (2007). School interventions for attention deficit hyperactivity disorder: Where to from here? *School Psychology Review, 36*, 279–286.

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.

Blackorby, J., Chorost, M., Garza, N., & Guzman, A. (2005). The academic performance of elementary and middle school students with disabilities. In J. Blackorby, M. Wagner, R. Cameto, E. Davies, P. Levine, L. Newman, . . . C. Sumi (Eds.), *Engagement, academics, social adjustment, and independence: The achievements of elementary and middle school students with disabilities.* Menlo Park, CA: SRI International.

Borduin, C. M., Mann, B. J., Cone, L. T., Henggeler, S. W., Fucci, B. R., Blaske, D. M., & Williams, R. A. (1995). Multisystemic treatment of serious juvenile offenders: Long-term prevention of criminality and violence. *Journal of Consulting and Clinical Psychology, 63*, 569–578.

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

Carter, D., & Horner, R. (2007). Adding functional behavioral assessment to First Step to Success: A case study. *Journal of Positive Behavior Interventions, 9*, 229–238.

Carter, D., & Horner, R. (2009). Adding functional-based behavioral support to First Step to Success: Integrating individualized and manualized practices. *Journal of Positive Behavior Interventions, 11*, 22–34.

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

Cizek, G. C. (2001). Test review of the Woodcock-Johnson III. In B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements yearbook* [Electronic version]. Retrieved from the Buros Center for Testing Web site: http://www.unl.edu/buros

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

Conduct Problems Prevention Research Group. (2011). The effects of the Fast Track Preventive Intervention on the development of conduct disorder across childhood. *Child Development, 82*, 331–345.

Connor, D. F. (2004). Prevalence of aggression, antisocial behaviors, and suicide. In D. F. Connor (Ed.), *Aggression and antisocial behavior in children and adolescents: Research and treatment.* New York: Guilford Press.

Eddy, J. M., Reid, J. B., & Fetrow, R. A. (2000). An elementary school–based prevention program targeting modifiable antecedents of youth delinquency and violence: Linking the Interests of Families and Teachers (LIFT). *Journal of Emotional and Behavioral Disorders, 8*, 165–176.

Epstein, M, & Cullinan, D. (1998). *Manual for the scale for assessing emotional disturbance (SAED).* Austin, TX: Pro-Ed.

Epstein, M., & Sharma, J. (1998). *Manual for the behavioral and emotional rating scale (BERS).* Austin, TX: Pro-ed.

Farmer, T. W., Van Acker, R. M., Pearl, R., & Rodkin, P. C. (1999). Social networks and peer-assessed problem behavior in elementary classrooms: Students with and without disabilities. *Remedial and Special Education, 20*, 244–256.

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, A., Faith, T., Elliot, A., & Royer, B. (2006). A pilot study examining the social validity and effectiveness of a positive behavior support model in Head Start. *School Social Work Journal, 30*, 22–44.

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

Golly, A., Sprague, J., Walker, H. M., Beard, K., & Gorham, G. (2000). The First Step to Success program: An analysis of outcomes with identical twins across multiple baselines. *Behavioral Disorders, 25*, 170–182.

Golly, A. M., Stiller, B., & Walker, H. M. (1998). First Step to Success: Replication and social validation of an early intervention program. *Journal of Emotional and Behavioral Disorders, 6*, 243–250.

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

Gunn, B., Feil, E., Seeley, J., Severson, H., & Walker, H. (2006). Promoting school success: Developing social skills and early literacy in Head Start classrooms. *NHSA Dialog, 9*, 1–11.

Hawkins, J. D., Catalano, R. F., Kosterman, R., Abbott, R., & Hill, K. G. (1999). Preventing adolescent health-risk behaviors by strengthening protection during childhood. *Pediatrics, 153*, 226–234.

Hollinger, J. D. (1987). Social skills for behaviorally disordered children as preparation for mainstreaming: Theory, practice and new directions. *Remedial and Special Education, 8*(4), 17–27.

Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The school-wide evaluation tool (SET) A research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions, 6*, 3–12.

Kazdin, A. E. (1987). *Conduct disorders in childhood and adolescence.* Beverly Hills, CA: Sage.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry, 62*, 617–627.

Kutash, K., Duchnowski, A. J., & Lynn, N. (2006). *School-based mental health: An empirical guide for* *decision-makers.* Tampa, FL: Research and Training Center for Children's Mental Health, Louis de la Parte Florida Mental Health Institute, University of South Florida.

Lane, K., Menzies, H. M., Oakes, W. P., & Kalberg, J. R. (2012). *Systematic screenings of behavior to support instruction: From preschool to high school.* New York: Guilford.

Maag, J. W. (2006). Social skills training for students with emotional and behavior disorders: A review of reviews. *Behavioral Disorders, 32*, 5–17.

Maag, J. W., Vasa, S. F., Kramer, J. J., & Torrey, G. K. (1991). Teachers' perceptions of factors contributing to children's social status. *Psychological Reports, 69*, 831–836.

Moffitt, T. (1993). Adolescent-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*, 674–701.

Nangle, D. W., Hansen, D. J., Eardley, C. A., & Norton, P. J. (2009). *Practitioner's guide to empirically based measures of social skills.* New York: Springer.

Nelson, R., Hurley, K., Synhorst, L., Epstein, M., Stage, S., & Buckley, J. (2009). The child outcomes of a behavior model. *Exceptional Children, 76*, 7–30.

Newman, L., Wagner, M., Cameto, R., & Knokey, A.-M. (2009). *The post-high school outcomes of youth with disabilities up to 4 years after high school. A report from the National Longitudinal Transition Study-2 (NLTS2)* (NCSER 2009–3017). Menlo Park, CA: SRI International.

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

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: Paul Brookes.

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

Sprague, J., & Perkins, K. (2009). Direct and collateral effects of the First Step to Success Program. *Journal of Positive Behavior Interventions, 11*(4), 208–221.

Sumi, W. C., Woodbridge, M. W., Javitz, H., Thornton, S. P., Wagner, M., Rouspil, K., . . . Severson, H. (2013). Assessing the effectiveness of First Step to Success: Are short-term results the first step to long-term behavioral improvements? *Journal of Emotional and Behavioral Disorders, 21*(1), 66–79.

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.

Wagner, M., & Davis, M. (2006). How are we preparing students with emotional disturbances for the transition to young adulthood? Findings from the National Longitudinal Transition Study-2 (NLTS2). *Journal of Emotional and Behavioral Disorders, 14*, 86–96.

Wagner, M., Marder, C., Blackorby, J., Cameto, R., Newman, L., Levine, P., . . . Sumi, C. (2003). *The achievements of youth with disabilities during secondary school.* Menlo Park, CA: SRI International.

Wagner, M., Newman, L., Cameto, R., Garza, N., & Levine, P. (2005). *After high school: A first look at the postschool experiences of youth with disabilities. A report of findings from the National Longitudinal Transition Study (NLTS) and National Longitudinal Transition Study-2 (NLTS2*). Menlo Park, CA: SRI International.

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., 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. R., Feil, E. G., Small, J. W., Golly, A. M., . . . Forness, S. R. (2014). The evidence base of the First Step to Success early intervention for preventing emerging antisocial behavior patterns. In H. M. Walker & F. M. Gresham, *Handbook of evidence-based practices for students having emotional and behavioral disorders* (pp. 518–536). New York: Guilford.

Webster-Stratton, C., & Taylor, T. (2001). Nipping early risk factors in the bud: Preventing substance abuse, delinquency, and violence in adolescence through interventions targeted at young children (0–8 years). *Prevention Science, 2*, 165–192.

Woodbridge, M., Sumi, W. C., Thornton, P., Javitz, H., Wagner, M., & Shaver, D. (2010). *National Behavior Research Coordination Center: Evaluation results for four interventions.* Menlo Park, CA: SRI International.

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

Date Received: November 2, 2012
Date Accepted: April 28, 2014
Associate Editor: Ann Schulte ■

Michelle W. Woodbridge, PhD, is a principal scientist in the Center for Education and Human Services at SRI International. She has more than 20 years of experience in research and evaluation of children's system-of-care services and school-based interventions for children with emotional and behavioral disorders.

W. Carl Sumi, PhD, is a senior education researcher in the Center for Education and Human Services at SRI International. He has worked with children with emotional and behavioral disabilities for more than 20 years in a variety of capacities, from direct services to research and policy development.

Mary M. Wagner, PhD, is a principal scientist in the Center for Education and Human Services at SRI International. She has conducted research for more than 30 years, with a focus on longitudinal studies of the characteristics, experiences, and achievements of children with disabilities and evaluations of interventions serving children and families.

Harold S. Javitz, PhD, is a senior biostatistician and principal scientist in the Center for Health Sciences at SRI International. He has more than 30 years of experience in educational and biostatistical research.

John R. Seeley, PhD, is a senior scientist at the Oregon Research Institute. His current interests include emotional and behavioral disorders in youth, mental health intervention, and research methodology.

Hill M. Walker, PhD, is Director of the Center on Human Development and Co-Director of the Institute on Violence and Destructive Behavior at the University of Oregon and a senior research scientist at the Oregon Research Institute. His research interests include curriculum development and intervention, youth violence prevention, and the development of early screening procedures for detecting students who are at risk of social–behavioral adjustment problems.

Jason W. Small, BA, is a data analyst at the Oregon Research Institute. He assists with data management, analysis, and proposal and manuscript development for multiple projects related to education and mental health.

Annemieke Golly, PhD, is a certified special education teacher and an assistant research scientist at the Oregon Research Institute. She is the coordinator and trainer for the First Step to Success program.

Edward G. Feil, PhD, is an educational psychologist and a senior research scientist at the Oregon Research Institute. His current interests include early screening and intervention for children with behavior problems and using technology to disseminate evidence-based treatments.

Herbert H. Severson, PhD, is a licensed psychologist and senior research scientist at the Oregon Research Institute. He has more than 35 years of experience in intervention and prevention research, and he is the co-developer of the First Step to Success program.