Journal of Emotional and Behavioral Disorders
Volume 17 Number 4
December 2009 197-212

10.1177/1063426609341645
http://jebd.sagepub.com
hosted at
http://online.sagepub.com

A Randomized Controlled Trial of the First Step to Success Early Intervention

Demonstration of Program Efficacy Outcomes in a Diverse, Urban School District

Hill M. Walker
University of Oregon, Eugene
John R. Seeley
Jason Small
Herbert H. Severson
Oregon Research Institute, Eugene
Bethany A. Graham
Albuquerque Public Schools, New Mexico
Edward G. Feil
Oregon Research Institute, Eugene
Loretta Serna
University of New Mexico, Albuquerque
Annemieke M. Golly
Oregon Research Institute, Eugene
Steven R. Forness
University of California, Los Angeles

This article reports on a randomized controlled trial of the First Step to Success early intervention that was conducted over a 4-year period in Albuquerque Public Schools. First Step is a selected intervention for students in Grades 1 through 3 with externalizing behavior problems, and it addresses secondary prevention goals and objectives. It consists of three modular components (screening, school intervention, parent training); lasts approximately 3 months; and is initially set up, delivered, and coordinated by a behavioral coach (e.g., school counselor, behavior specialist, social worker). Project Year 1 of this efficacy trial was devoted to gearing-up activities (e.g., hiring, training, planning, logistical arrangements); Years 2 and 3 each involved implementing First Step with approximately 100 behaviorally at-risk students. Students, teachers, and classrooms were randomly assigned to either intervention or usual care comparison conditions. Year 4 activities focused on conducting long-term, follow-up assessments and implementing sustainability procedures to preserve achieved gains. Pre-post teacher and parent ratings of student behavior and social skills showed moderately robust effect sizes, ranging from .54 to .87, that favored the intervention group. Direct measures of academic performance (oral reading fluency, letter–word identification) were not sensitive to the intervention. The implications and limitations of the study are discussed.

Keywords: *anger/aggression; ecobehavioral assessment; assessment of social behaviors; at-risk populations; antisocial*

Over the past several decades, teachers in elementary classrooms have been challenged by large numbers of behaviorally at-risk children who begin their school careers accompanied by a history of exposure to multiple conditions of risk within family, neighborhood, and community contexts. Many of these children are unresponsive to the necessary demands of the schooling process that require cooperating, sharing, accepting limits, interacting positively with peers and adults, listening to others, self-regulating behavior, focusing attention,

and engaging in academic tasks (Walker, Ramsey, & Gresham, 2004).

As a group, educators report that these students have not been well socialized to the common norms and expectations of schooling and are not prepared to succeed in either an academic or a behavioral sense (Adelman & Taylor, 2006; Nelson, Benner, & Mooney, 2008). As a consequence, they may be unable to take full advantage of the normalizing and protective influences of schooling and will likely not bond with or forge a strong attachment to the schooling experience.

The continuity and stability of these behavior patterns across school years can severely disrupt a student's social-emotional adjustment and academic success (Hawkins, Catalano, Kosterman, Abbott, & Hill, 1999; Reid, Patterson, & Snyder, 2002; Shinn & Walker, in press; Walker et al., 2004). The long-term negative outcomes associated with these maladaptive forms of behavior are serious and have been well documented in past research on antisocial child populations (Lipsey & Derzon, 1998; Loeber & Farrington, 2001; Reid et al., 2002). It is important to move behaviorally at-risk children off this negative trajectory or pathway as soon as possible in their school careers through early, coordinated interventions that are delivered via collaborative partnership arrangements between child behavior experts, parents and caregivers, school staff and peers, and community agencies, as appropriate (Dishion, Stormshak, & Siler, in press; Furlong, Pavelski, & Saxton, 2002).

The emergence of the evidence-based practices movement in our field has raised the stakes regarding which intervention approaches are selected and how they are implemented by practitioners and school-based professionals (Burns & Hoagwood, 2002; Detrich, Keyworth, & States, 2008). Carefully implemented interventions that have been validated as efficacious and/or as effective are viewed by many professionals as necessary to address the challenges presented by the growing at-risk subpopulation served by today's schools (Kutash, Duchnowski, & Lynn, 2006; Van Eck, Evans, & Ulmer, 2007). For special education in particular, Forness has argued that evidence-based practices for children at risk for emotional or behavioral disorders not only should be based on randomized controlled trials but also should be characterized by manualized interventions; diverse samples of participants; multisite collaboration; multimodal treatments; and multiple outcome domains including behavior, impairment, and academics (see Forness, 2005; Forness & Beard, 2007). To date, relatively few studies reported in the educational literature have met this profile.

Although considerable progress has been made in the past decade in the development and dissemination of school-based prevention approaches (Detrich et al., 2008; Greenberg et al., 2003), there still remains an insufficient level of reliable evidence on the efficacy of coordinated early-childhood interventions that address the multiple risk factors and conditions that place antisocial children at risk for school failure and other later, destructive outcomes such as delinquency and substance abuse (Hoagwood, 2003–2004). Hoagwood and colleagues (Hoagwood et al., 2007; Schoenwald & Hoagwood, 2001) have long argued that the adoption, sustainability, and integration of efficacious interventions into the normal practices of applied school and clinical settings remain to be demonstrated on any broad scale. Hoagwood and her colleagues recently reported a review of empirically based school interventions targeted at academic and mental health functioning (see Hoagwood et al., 2007). This comprehensive review illustrates the dearth of proven and promising interventions that are currently available to school-based professionals, researchers, and practitioners.

The First Step to Success Intervention

The First Step to Success program, which is the focus of this study, was initially developed between 1992 and 1996 through a 4-year U.S. Office of Special Education

Authors' Note: This research was conducted under the auspices of a 4-year Behavior Research Center grant to the senior author from the Institute of Education Sciences, U.S. Department of Education, Grant No. H324P040006, Evidence-Based Interventions for Severe Behavior Problems. The contents of this article do not necessarily represent the policy of the Department of Education and do not imply endorsement by the federal government. We are most grateful to the following Albuquerque Public Schools (APS) administrators whose support, advocacy, and assistance made this study possible: Beth Everett, superintendent; Debi Hines, director of special education; and Deborah Duncan, school mental health coordinator. Instrumental to the recruitment process was the involvement of Deborah Duncan, the APS school mental health coordinator. Ms. Duncan served as the key liaison between district administration, cluster leaders, elementary school principals, and the First Step project manager (Graham). She arranged presentations at both district and cluster levels and communicated effectively with the director of special education, district superintendents, and researchers from the Oregon Research Institute. While district and cluster meetings were addressed by both the project manager and Ms. Duncan, individual school presentations were made by the project manager after an invitation was extended by the principal. We wish to express appreciation to Jacquelyn Buckley of the Institute of Education Sciences and Ronnie Detrich of the Wing Institute for their review and feedback on the manuscript. Their comments and recommendations were especially helpful in the revision process.
Correspondence concerning this article should be addressed to Hill M. Walker, 1265 University of Oregon, Eugene, OR 97403-1265; email: hwalker@uoregon.edu.

Programs grant to the senior author that involved a collaborative effort of the Eugene School District, the University of Oregon, the Oregon Social Learning Center, and the Oregon Research Institute. As described by Walker et al. (2008), multiple studies involving differing methodologies over the past decade (randomized controlled trials, quasi-experimental designs, single-subject studies of program outcomes) have been conducted by the First Step program's developers (Golly, Stiller, & Walker, 1998; Walker et al., 1998) and other investigators (Beard & Sugai, 2004; Overton, McKenzie, King, & Osbourne, 2002) to help establish the program's efficacy. Evidence for the First Step program's efficacy is thus informed by the hierarchy standard of evidence (e.g., a mix of single-subject studies, quasi-experimental studies, and randomized controlled trials) rather than by the threshold standard of evidence (e.g., randomized controlled trials only), as described by Detrich et al. (2008) and by Drake, Latimer, Leff, McHugo, and Burns (2004).

The First Step program is a manualized intervention consisting of the three modular components of universal screening, classroom intervention, and parent training; the program is a selected intervention of 3 months' duration designed to address secondary prevention goals and outcomes (Walker et al., 1997). The screening component of First Step is used to identify candidates who meet eligibility criteria for program participation. Classroom intervention and parent training comprise the program intervention component of First Step. During the first 5 days of the program, the behavioral coach, a school professional who works with and coordinates the roles of the target child, parents, teacher, and peers throughout the implementation process, explains and implements the classroom intervention. Typically, this person is a counselor, school psychologist, behavioral specialist, or social worker. On the 6th program day, the teacher takes over implementation of the program with the support, assistance, and oversight of the coach. On the 10th day, First Step is extended to the target student's home setting where the coach trains parents, through six weekly home visits, how to teach their child key school success skills such as communication and sharing, cooperation, problem solving, limit setting, and friendship making. Through instruction, role-playing, cueing, prompting, and feedback, parents learn how to teach and encourage these skills in their child.

First Step requires completion of 30 program days, each with a prescribed set of activities, tasks, and a reward criterion (Walker et al., 1997). For each day that the target student earns at least 80% of the possible daily points, as achieved by exhibiting positive behavior in the classroom, he or she earns either a group activity reward that is shared with the entire class or a prearranged home reward. If the requisite points are not achieved, the program day is "recycled," and the child is given additional opportunities to complete the failed day. The final 10 days of the program aim to maintain the target child's improved behavior without reliance on external rewards. In this phase, the focus shifts to adult praise, intrinsic rewards, and encouragement by teachers, peers, and parents to motivate and sustain the child's improved behavior.

Throughout the First Step program, the child's behavior is carefully monitored by the participating teachers at school and by parents at home. Parents teach school success skills at home, while teachers look for, recognize, and praise the child's positive behavior at school. First Step coaches, and subsequently teachers, use a green card visible to the entire class to signal the target child that his or her behavior is positive and earning points, whereas the red side of the same card is used to signal the opposite. Due to the group-dependent nature of the First Step program's contingencies, peers become supportive of the target child's attempt to display positive behavior. In turn, peer support and involvement, as reflected in increased rates of social bids, invitations, and positive peer-to-peer interactions, along with inclusion of the target child in peer-control activities, helps attenuate the negative reputational bias that peers often hold toward antisocial and disruptive students (Hollinger, 1987).

Walker et al. (2008) recently provided an overview of the research and knowledge base developed to date on the First Step program. Until the current investigation, First Step had been evaluated primarily in suburban and rural school district settings using mainly single-subject and quasi-experimental designs. However, the authors did conduct a small-scale randomized controlled trial that showed relatively positive effects (see Walker et al., 1998). Aside from the limitation of a small sample size of 46 participants, this study involved a waitlist comparison group design that prevented use of a control group in follow-up assessments. Although this collection of studies suggested promising outcomes for First Step, the current study is the first opportunity we had been afforded to evaluate the program's effects under more real-world and complex school district and community conditions.

This article describes a randomized controlled trial of the First Step to Success early intervention program, involving 200 student participants enrolled in Grades 1 through 3, conducted in the Albuquerque, New Mexico, School District. Students in Grades 1 through 3 who

were experiencing behavioral problems of an externalizing nature were the focus of the current study. Each participating student was enrolled in a different general education classroom setting during the 3-month intervention. The purpose of this study was twofold: (a) to conduct a large-scale randomized controlled trial of the First Step program to demonstrate its efficacy and (2) to determine if program effects and outcomes in a diverse, highly urbanized school setting matched those previously obtained in less diverse, suburban and rural settings. The remainder of this article reports on the methods, intervention procedures, measures, and outcomes produced by this 4-year investigation.

Method

Study Design

In the 2005-2006 and 2006-2007 school years, two cohorts of first- through third-grade students, teachers, and general education classrooms from 34 elementary schools of the Albuquerque Public Schools (APS) were participants in this First Step to Success efficacy study. Approximately half of the participating schools were involved in this study during the 2005-2006 school year and the remainder in the 2006-2007 school year.

A cohort design model was used in which waves of intervention and usual care comparison students participated in either of two school years. Random assignment occurred at the classroom level within waves; thus, only 1 student was identified per classroom for study participation. Randomization of identified participants was implemented prior to solicitation of parental consent for their child's participation in the study. In Cohort 1, there were 99 student participants (44 usual care comparison and 55 intervention); in Cohort 2, there were 101 (55 usual care comparison and 46 intervention). Cohort 1 students identified for participation were equally distributed across condition at randomization, but a larger proportion of parents assigned to the usual care condition declined consent for participation. In turn, we randomized a larger proportion of students to the usual care condition in Cohort 2 to achieve a balanced design across conditions.

In Cohort 1, the first wave of students was identified using universal screening procedures conducted in the early fall of the 2005-2006 school year. The identified students, along with their teachers and classrooms, were then randomly assigned to either the First Step intervention or the usual care comparison condition. This procedure was replicated for Wave 2 participants in the late fall and then again for Wave 3 participants in the early spring of the 2005-2006 school year. We followed this same procedure and sequence for identifying and assigning participants in the 2006-2007 school year; however, teachers and classrooms participated across two rather than three waves of data collection due to student testing that delayed recruitment efforts and precluded collection of a third wave of data. To test for potential differences between the two annual recruitment cohorts of students, we included cohort as a between-subjects factor in the MANCOVA models used in our data analyses and tested for Condition × Cohort interactions. No statistically significant interactions with condition were obtained for any of the outcome measures used. Thus, we combined Cohorts 1 and 2 for analysis purposes and respecified the models without including cohort as a between-subjects factor.

Setting

APS is one of the largest and most diverse school districts in the nation, covering more than 1,200 square miles, employing more than 6,000 teachers, and serving nearly 90,000 students. APS ranks as the 17th largest school district nationally. Schools in APS are organized using a cluster system designed to facilitate the development of small learning communities. APS consists of 12 clusters. There are 11 high school clusters into which 84 elementary and 26 middle schools feed, along with 1 alternative school cluster. Seventy-two percent of the APS kindergarten through 12th-grade school population consists of students of color; approximately 57% of the APS student population is of Hispanic origin. APS serves a large, urban community that experiences substantial levels of poverty accompanied by high rates of alcohol and substance abuse.

Recruitment of APS Elementary Schools

The approach used for recruiting APS elementary schools for study participation was to first obtain the support of the district superintendent along with the three APS associate superintendents, the director of special education, and the APS school mental health coordinator. Next, a presentation of the First Step to Success intervention was made to the cluster leader principals. Each principal then took the information back to his or her respective clusters for distribution to elementary principals. Following this step, a presentation was made to elementary staff at individual schools, and teachers were invited to participate in the study.

Participation in the research was contingent on program approval by the principal. Once approval was received, individual teachers were asked to join the study. A minimum of six classrooms per school was requested of participating schools; two classrooms from each first-, second-, and third-grade level. Adherence to this goal was achieved or surpassed in nearly all instances.

Participants and Procedures

General education classroom teachers from at least one elementary school in each school cluster participated in the study. Teachers who consented to participate (n = 260) were randomly assigned to either the intervention or comparison condition and were asked to complete Stages 1 and 2 of a universal problem behavior screener, the *Systematic Screening for Behavior Disorders* (SSBD) procedure (Walker & Severson, 1990). The SSBD uses a multiple-gating approach to detect students in kindergarten through fifth grade who have an elevated risk for school behavior problems. In screening Stage 1, teachers were given descriptions and examples of externalizing behaviors and were asked to nominate and rank order the five students in the class who exhibited the highest levels of externalizing problem behaviors. For the three highest-ranked students identified during the first stage, teachers then completed Stage 2 of the SSBD, which included brief ratings of student adaptive and maladaptive behavior and a checklist of 30 high-intensity, low-frequency, maladaptive behavioral indicators (e.g., critical behavioral events). The student with the highest average ranking across the SSBD Stage 2 measures was targeted for inclusion in the study. Although the SSBD has an optional third assessment stage in which classroom and playground observations are coded and then compared to normative levels for students who meet Stage 2 eligibility criteria, this procedure was judged to be more labor intensive than the study could afford.

Of the 260 consented teachers, 243 (94%) participated in the SSBD screening process, completing Stage 1 and Stage 2 for 723 students across both cohorts. Twenty-seven teachers (10%) dropped out of the study, 17 prior to screening and 10 following screening. An additional 23 classrooms (9%) were dropped from the study because either the top-three teacher-identified externalizing students declined participation (n = 13) or parental consent was not obtained in sufficient time to complete the assessment process prior to the end of the academic year (n = 10). Overall, parental consent was obtained for 210 of the 260 recruited teachers/classrooms (81%). Of the remaining 210 consented students, 107 were randomly assigned to the intervention, and 103 were assigned to the comparison or usual care condition. Ten students (5%) dropped out of the study after parental consent was obtained for participation, thereby reducing the participating sample to 200 students (99 comparison and 101 intervention). Although every effort was made to recruit the first-ranked externalizer in each classroom as identified during universal screening, on occasion parents of the first- or even second-ranked student declined participation. Of the 723 students for whom an SSBD screener was completed, 331 were invited to participate, and 210 parents provided their consent for study participation (63.4%). The CONSORT diagram summarizing the recruitment breakdown is presented in Figure 1.

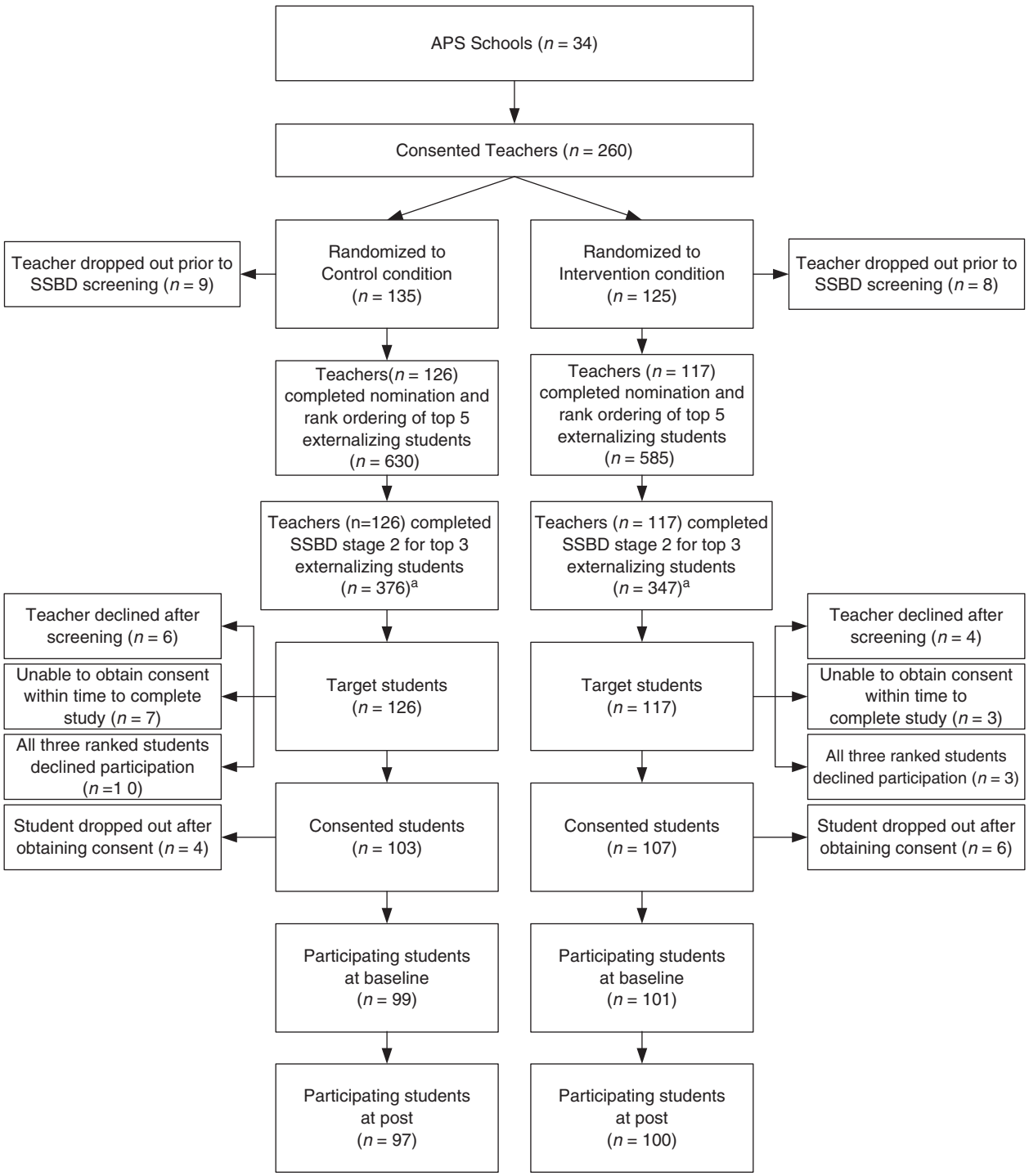
The 200 consented students who participated in the study ranged from 6 to 10 years in age at enrollment (M = 7.2, SD = 1.0) and were predominantly male (73%). Eighty-three students were first graders, 69 students were second graders, and 48 were third graders. Students were predominantly Hispanic (57%) or Caucasian (24.5%), with the remaining racial and ethnic groups representing less than 20% of the sample (4.5% American Indian, 0.5% Asian or Pacific Islander, 7% Black, 3% multiracial, and 3% unknown). Students came primarily from English-speaking households (88.9%). Seventy percent were eligible for free or reduced-price lunches, and roughly 16% were English language learners (ELL).

Systematic Screening for Behavior Disorders Eligibility

In this study, externalizing students met criteria for behavior problems in one of two ways using the SSBD scales: (a) if a student had five or more critical events endorsed on the Critical Events Index (CEI) or (b) if a student had one or more (but fewer than five) critical events endorsed on the CEI, a score of 30 or lower on the Adaptive Behavior Index (ABI), and a score of 35 or higher on the Maladaptive Behavior Index (MBI). Rank ordering of students was based on their CEI, MBI and ABI scores, with the rank order across the three scales averaged for each student. If two students had the same average rank, the one with the higher raw CEI score received the higher rank.

For the First Step efficacy study, 133 of the 200 students (66.5%) met SSBD Stage 2 eligibility criteria. There was no statistically significant difference between study condition for eligibility, as 62.6% of comparison students and 70.3% of intervention students met eligibility criteria, $\chi^2(1, N = 200) = 1.32$. Fifty six (83.6%) of the remaining 67 students who did not meet strict SSBD Stage 2 criteria had elevated CEI scores (between one and four critical events) or met criteria on either the

Figure 1
Schematic Overview of Randomization, Screening, and Consent Procedures



a. Parent decline forms for 6 students were received after the collection of Stage 2 screening data. These students were removed from the sample, and their data were destroyed.

adaptive or maladaptive index. Thirty-eight students had elevated CEI criteria; 10 students met criteria for the MBI, and 8 met criteria for the ABI.

To test whether those who met SSBD Stage 2 criteria (n = 133) differed from those who did not (n = 67) with respect to intervention effects, we included meeting

SSBD Stage 2 criteria (yes/no) as a between-subjects factor in the MANCOVA models used in our data analyses, and we tested for Condition × SSBD Stage 2 Criteria interactions. No statistically significant interactions with condition were obtained for any of the outcome measures examined, indicating that the SSBD eligibility criteria did not moderate the intervention effects. Thus, we respecified the models without including SSBD criteria as a between-subjects factor.

Sample Representativeness

To examine the sample's representativeness, the final group of participating students (n = 200) were compared to all students who were eligible but not selected to participate in the study (n = 523); these students were compared on their baseline demographics and severity of problem behaviors. Participating students were comparable to nonparticipating students on all demographic variables including age, M = 7.2, SD = 1.0, $t(720) = 1.22$; percentage female, 18.7%, $\chi^2(1, N = 723) = 2.97$; percentage Hispanic, 58.5%, $\chi^2(1, N = 723) = 0.14$; percentage Spanish speaking, 13.2%, $\chi^2(1, N = 716) = 0.58$; percentage eligible for free or reduced lunch, 77.4%, $\chi^2(1, N = 452) = 3.32$; and percentage ELL, 21.5%, $\chi^2(1, N = 664) = 2.26$. As expected, there were statistically significant differences ($p < .001$) between the samples selected for participation and the nonparticipating samples on the three SSBD screening measures used to target students with the most severe problem behaviors. Participating students had an average of 6.4 (SD = 3.8) critical events, an average maladaptive score of 34.5 (SD = 8.3), and an average adaptive score of 32.4 (SD = 7.7). In comparison, nonparticipating students averaged 4.4 (SD = 3.3) critical events and scores of 30.8 (SD = 8.4) and 35.3 (SD = 7.7), respectively, on the maladaptive and adaptive scales.

A total of 21 participating students in the sample were receiving services under either an individualized education program or a 504 plan. Twenty-seven students had a behavior support plan in place. There were no statistically significant differences for condition between these two variables.

First Step Coaches

The coaches in this study were drawn from a pool of behavior management specialists and behavior consultants from the APS behavior consultation service team. A total of 6 consultants and 24 behavior management specialists were trained by one of the coauthors.

All coaches attended a 2-day First Step training institute in Albuquerque and were then assigned intervention cases on a randomized basis and worked with the assigned target student, teacher, and parents to implement the First Step intervention over approximately a 3-month period. Coaches were in close contact with First Step supervisory project staff and were scheduled for fidelity monitoring checks regularly to review their adherence to the implementation protocol for the First Step intervention. The First Step trainer also held weekly video conferences with coaches to answer questions and troubleshoot problems.

Coaches reported an average time commitment of 30 hours per case from start to finish. This included time for training; working with the student, teacher, and parents; and any extra calls or time commitments they deemed necessary for program implementation and troubleshooting. Compensation for coaches was $400 for each case completed. The APS First Step to Success project manager provided technical assistance and support on a daily basis as well as attending behavior consultation service team meetings to distribute current information, answer questions, receive feedback, and problem solve solutions to identify problems and implementation challenges.

Outcome Measures

Outcome data were collected with teacher- and parent-reported measures, direct observations, and individual academic performance measures. Baseline data were collected at the beginning of each wave (early fall, late fall, or early spring), and postintervention data were collected upon completion of the First Step intervention (M = 58.0 days, SD = 29.3 days). As part of a larger questionnaire collected prior to and following intervention, teachers and parents completed the *Social Skills Rating System* (SSRS; Gresham & Elliott, 1990), and teachers completed two scales from SSBD (Walker & Severson, 1990). Trained assessors collected direct observation data using the SSBD measure of student academic engaged time (AET), and they also collected academic data using the Letter–Word Identification subtest from the *Woodcock–Johnson III Diagnostic Reading Battery* (WJ-III DRB; Woodcock, Mather, & Schrank, 2004) and a measure of oral reading fluency (Fuchs, 2003). A description of each outcome measure follows.

Social Skills Rating System

The SSRS (Gresham & Elliott, 1990) is a 57-item scale that samples the three domains of social skills, problem behaviors, and academic competence. The 30-item Social Skills subscale (α = .88) assesses the core skills of cooperation, assertion, and self-control as reported by the

teacher's perceived frequency rating on a 3-point scale (*never, sometimes,* or *very often*). The 18-item Problem Behavior subscale (α = .85) assesses the teacher's perceived frequency of internalizing and externalizing problem behaviors that may interfere with social skills performance. The problem behavior items are assessed on a 3-point scale (*never, sometimes,* or *very often*). The 9-item Academic Competence subscale (α = .91) assesses reading and math performance as well as the student's motivation, intellectual functioning, and parental support as estimated by the teacher on a 5-point percentage cluster scale (from the lowest 10% to the highest 10%).

Parent-reported outcomes included the SSRS Social Skills and Problem Behavior scales. The 38-item Social Skills subscale (α = .88) assesses the parent's perceived frequency of the child's development of social competence as it pertains to day-to-day activities and interactions at home. The Problem Behavior subscale (α = .88) has 17 items that measure the parent's perceived frequency of internalizing and externalizing problem behaviors that may interfere with their child's social skills (Gresham & Elliott, 1990). Both SSRS parent subscales are scaled and scored in the same manner as the SSRS teacher-reported measures described above.

Systematic Screening for Behavior Disorders

The SSBD uses a multiple-stage approach to detect students in kindergarten through sixth grade who have an elevated risk for school behavior problems (Walker & Severson, 1990). This universal screening procedure consists of three interrelated, and increasingly intensive, screening stages that cross-validate the results of each other. The SSBD screening stages are (a) nomination and rank ordering according to descriptions and examples of externalizing and internalizing behavioral profiles, (b) teacher ratings of the student's adaptive and maladaptive behavior and completion of a critical events checklist, and (c) behavioral observations of academic engagement in the classroom and social behavior on the playground. Two behavior rating scales from Stage 2 were completed by teachers prior to and following intervention. The ABI, a 12-item scale (α = .88), and the MBI, an 11-item scale (α = .87), assess the student's teacher-related and peer-to-peer adaptive and maladaptive behavioral adjustments based on a 5-point rating scale ranging from *never to frequently*.

The SSBD has excellent psychometric characteristics and is nationally normed. The SSBD procedure has been used in a number of research studies reported in the professional literature (see Severson, Walker, Doolittle, Kratochwill, & Gresham, 2007; Walker & Severson, 1990; Walker, Severson, & Seeley, 2007).

Student Academic Engaged Time

Direct observation data were collected at each data collection time point using the SSBD Stage 3 measure of student AET (Walker & Severson, 1990). AET estimates, via a stopwatch recording procedure, the amount of time a student spends engaged in allocated academic activities. As described by Walker and Severson, AET serves as an important indicator of a student's academic involvement and adjustment to the teacher's classroom expectations for all students. AET is operationalized as follows: (a) attending to the material and task, (b) making appropriate motor responses, (c) asking for assistance at the appropriate time and in an acceptable manner, (d) interacting with the student's teacher and classmates about academic matters, and (e) listening to teacher instructions and direction.

A pool of professionally trained observers, blind to student condition, collected two 15-minute AET observations for each student participant at each time point (baseline, post, and follow-up). Observers were recruited from the University of New Mexico and through advertisements in the local Albuquerque newspaper. These AET assessments were collected on different days within a week of one another ($M = 2.6$ days, $SD = 2.3$ days) and were averaged to compute the percentage of AET for baseline, post, and follow-up time points. For each observation, project staff coded and recorded the classroom structure (circle time, teacher-led discussion, independent seatwork, and cooperative learning) and the classroom activity (literacy, math, art or fine motor, and science) in operation at the time of the observation. Most AET observations were collected during circle time (24.8%), teacher-led discussion (31.6%), and individual seatwork (33.8%) activities during which target students were engaged in literacy-related (68.7%) and math-related (19.1%) instructional activities. To minimize the effect (or effects) of varying classroom contexts on student engagement, every attempt was made to collect postintervention data at the same time of day and during a similar classroom activity and structure as in the preintervention observation. There were no statistically significant differences in classroom structure, $\chi^2(4, N = 787) = 7.23$, or classroom activity type, $\chi^2(4, N = 804) = 6.84$, between baseline and postintervention data collection occasions.

Academic Engaged Time Observer Training and Monitoring Procedures

At the beginning of each school year, observers attended a 2-day AET training session during which they

received explicit instruction in observational procedures and coding techniques. During the first day of training, observer trainers reviewed AET definition examples and nonexamples, reviewed videotaped examples of AET recorded in general education classroom settings, and compared their recordings of these taped sessions with each other and the observer trainer. During the second training day, observers visited an APS school, conducted AET observations, and received feedback from a trained reliability observer.

All observers were required to demonstrate and sustain high reliabilities (minimum .90 interobserver agreement levels) prior to and during data collection periods. Observers were monitored in 20% of conducted observations and retrained as necessary throughout the course of the study in order to minimize drift and ensure adequate reliability of recorded observations.

Across the two cohorts and waves of data collection, reliability estimates were collected on 20% of the recorded AET observations. The intraclass correlation (ICC) assessing interrater reliability for AET observations was excellent, ICC(3, 1) = .99. For Waves 1, 2 and 3 of Cohort 1 and Waves 1 and 2 of Cohort 2, the average AET reliabilities ranged from .95 to .99 across all baseline and postintervention phases.

Woodcock–Johnson III Letter–Word Identification Subtest

To assess students' word identification skills, a trained assessor who was blind to student condition administered the Letter–Word Identification subtest from the WJ-III DRB (Woodcock et al., 2004). The WJ-III DRB is nationally normed and provides raw scores, grade-equivalent scores, age-equivalent scores, percentile ranks, and standard scores. The WJ-III Letter–Word Identification subtest measures the student's reading identification skills in identifying isolated letters and words and has a median reliability of .91 for students ages 5 to 19 (Woodcock et al., 2004).

Oral Reading Fluency

To assess students' reading abilities, a series of standard oral reading fluency passages were administered (Fuchs, 2003). This measure, which has been used in the six U.S. Department of Education–funded Reading and Behavior Centers, includes 300- to 400-word reading passages at a first-grade reading level. Two different passages were administered by a trained assessor who was blind to student condition. The correct words read per minute were calculated for each reading passage and averaged to compute a total score for each assessment time point.

Process Measures

Implementation fidelity, teacher–coach alliance, estimates of student and parent program compliance, and social validity data were collected for all participants assigned to the intervention condition in order to (a) determine the extent to which First Step was implemented as intended, (b) examine perceived satisfaction with the teacher–coach relationship as it pertains to program implementation, (c) measure whether students complied with the program and parents participated in the home-Base component of the program, and (d) assess teacher and parental consumer satisfaction with First Step.

Expert raters collected implementation fidelity data on four occasions during First Step implementation: once for the behavioral coach during the first 5 days of program implementation and then on three other occasions, at the beginning, middle, and end of the teacher phase of the program, for the teacher who implemented the program. The Implementation Fidelity Checklist assesses the extent to which the coach and teacher deliver First Step as intended. The checklist includes 18 First Step implementation components, such as whether the implementer announces the number of points needed for the reward, elicits cooperation from the class, informs the class of the reward, gives points when prompted, provides positive feedback to the target student during the red/green card game, and turns the card to red when inappropriate behavior occurs. For each implementation component, the fidelity checklist assesses (a) whether the component was implemented (yes/no) and (b) the quality of implementation using a 5-point scale with 0 = *very poor,* 0.25 = *poor,* 0.50 = *okay,* 0.75 = *good,* and 1.0 = *excellent* (α = .86). The ICC assessing interrater reliability for implementation fidelity checks was excellent, ICC(3, 1) = .92.

Data from the Implementation Fidelity Checklist were used to calculate both adherence and quality implementation scores for the coach, teacher, and overall classroom. Coach and teacher adherence scores were calculated as the proportion of procedures correctly implemented, and a mean of the coach and teacher adherence scores was computed to estimate the overall classroom adherence score. Mean quality ratings for the coach, teacher, and overall classroom were calculated as well.

To assess alliance, teachers and coaches completed a 10-item rating scale (α = .94) during the postintervention

phase of the study. Alliance items were assessed on a 5-point scale (ranging from *never* to *always*) and measured the respondent's perception of shared goals, communication, trust, and effectiveness of the partnership with respect to implementation.

Data were also collected to assess parental involvement in the homeBase intervention module of the First Step program. A parent compliance measure was computed as the proportion of homework assignments completed, and a dosage measure was calculated as the proportion of intervention units delivered based on the number of 1-hour homeBase sessions (out of six possible) in which the parents participated as described in the First Step manual. Student compliance was measured as the proportion of intervention sessions completed without "recycling" (repeating a program day of the First Step program).

Social validity data were recorded for teacher and parent satisfaction with First Step. The 13-item teacher satisfaction report ($\alpha = .92$) assesses the teacher's perception of the program training and support received, as well as the usability of the program, the teacher's belief about the effectiveness of the program with respect to changes in student behavior and peer interactions, and whether the teacher would use and recommend the program in the future. Satisfaction items were scored on a 5-point scale from *strongly disagree* to *strongly agree*. The 12-item parent satisfaction report ($\alpha = .92$), scored identically to the teacher version, examined the parent's perceptions of the usability, effectiveness, and value of the program based on the impact of the program on the child's behavior in the home setting.

Statistical Analysis

The outcome measures were organized into three domains: problem behavior symptoms, functional social impairment, and academic outcomes. The parent- and teacher-reported Problem Behavior subscales of the SSRS and the MBI from the SSBD comprise the problem behavior symptoms domain (mean intercorrelation = .39). The functional social impairment domain includes the parent- and teacher-reported social skills subscales from the SSRS and the ABI from the SSBD (mean intercorrelation = .39). The academic domain includes teacher-reported academic competence from the SSRS, student AET, the WJ-III Letter–Word Identification subscale, and the measure of oral reading fluency (mean intercorrelation = .37).

MANCOVA models controlling for baseline levels were conducted for each of the three domains to determine the multivariate effect size, followed by univariate ANCOVA models. Because multiple measures were tested for intervention effects, the Benjamini–Hochberg correction for the Type I error rate was applied to the 10 univariate tests (see Schochet, 2008). To conduct an intent-to-treat analysis, missing values of the outcome measures were imputed using the EM (expectation-maximization) method; missing data were less than 4% across all outcome measures. Effect sizes are reported as Cohen's *d* statistic (Cohen, 1988) and were calculated by dividing the difference between the intervention and comparison group adjusted means by the pooled standard deviation at posttest. In addition, to evaluate the practical significance of the intervention effects, we report the What Works Clearinghouse improvement index (Valentine & Cooper, 2003), which can be interpreted as the expected change in percentile rank for an average comparison group student if that student had received the First Step intervention.

Results

Preliminary Analyses: Baseline Equivalency and Attrition

The equivalency of the two study conditions was examined at baseline. The student-level baseline demographic characteristics are reported in Table 1. As can be seen from this table, the two conditions did not significantly differ from each other on any of the demographic characteristics. In addition, no significant differences between conditions occurred on any of the 10 baseline outcome measures.

Four informants provided assessment data at baseline and postintervention: teachers, parents, academic assessors, and behavioral observers. Postintervention data from all four informants were collected on 92% of participating students, with postintervention data from at least three informants being collected on 99% of the sample. Attrition rates by informant were highest for observation data (4.0%), followed by parent (3.0%), academic assessment (1.5%), and teacher data (1.0%). Data were examined for differential attrition rates by condition for each informant type; no significant differences between conditions were found.

Intervention Fidelity, Therapeutic Alliance, Program Compliance, and Satisfaction

The means and standard deviations for the process measures collected for the intervention condition are

Table 1
Student-Level Baseline Demographic Characteristics by Condition

Demographic Characteristic	Total (N = 200)	Comparison (n = 99)	Intervention (n = 101)	p Value
Age $M(SD)$	7.2 (1.0)	7.1 (0.9)	7.2 (1.0)	.317
Percent female	49 (24.5)	28 (28.3)	21 (20.8)	.218
Grade				.653
Percent in 1st grade	83 (41.5)	43 (43.4)	40 (39.6)	
Percent in 2nd grade	69 (34.5)	35 (35.4)	34 (33.7)	
Percent in 3rd grade	48 (24.0)	21 (21.1)	27 (26.7)	
Systematic Screening for Behavior Disorders rank				.189
1st-ranked student	158 (79.0)	73 (73.7)	85 (84.2)	
2nd-ranked student	36 (18.0)	22 (22.2)	14 (13.9)	
3rd-ranked student	6 (3.0)	4 (4.0)	2 (2.0)	
Percent receiving services	11 (5.5)	3 (3.0)	8 (7.9)	.129
Percent Spanish speaking	22 (11.1)	14 (14.1)	8 (8.0)	.167
Percent Hispanic	114 (57.0)	60 (60.6)	54 (53.5)	.308
Percent English language learner	32 (16.4)	17 (17.7)	15 (15.2)	.630
Percent free or reduced lunch eligible	127 (69.8)	61 (66.3)	66 (73.3)	.302

Table 2
Means and Standard Deviations (in parentheses) for First Step to Success Process Measures

	Classroom			homeBase	
	Coach	Teacher	Combined	Parent	Overall
Protocol adherence mean percent	0.84 (0.14)	0.82 (0.15)	0.83 (0.12)	–	0.83 (0.12)
Quality of implementation mean percent	0.85 (0.12)	0.80 (0.11)	0.83 (0.10)	0.76 (0.27)	0.80 (0.13)
Dosage mean percent	–	–	0.89 (0.18)	0.94 (0.17)	0.91 (0.14)
Participant compliance mean percent	–	–	0.94 (0.10)	0.67 (0.39)	0.83 (0.20)
Working alliance mean score	4.53 (0.55)	4.72 (0.43)	–	–	4.62 (0.41)
Program satisfaction mean score	–	3.77 (0.73)	–	4.33 (0.60)	4.10 (0.49)

presented in Table 2. Protocol adherence to First Step implementation was good for both the coach (84%) and teacher (82%) phase of the intervention, with an overall average implementation fidelity percentage of 83%. The quality of implementation averaged 0.83 for classroom implementation and 0.76 for the homeBase components, which indicate mean ratings across intervention components within the good-to-excellent range. With respect to intervention dosage, students received on average 89% of the available classroom program days and 94% of homeBase sessions. Student compliance to the classroom component was also found to be high (mean compliance score = 94%). In contrast, parent compliance to the homeBase homework was moderate (mean compliance score = 67%). Working alliance was rated highly by both coaches (mean score = 4.5 on a 5-point scale) and teachers (mean score = 4.7). Lastly, program satisfaction ratings were quite favorable based on parent report (mean score = 4.3 on a 5-point scale; mean item ratings exceeded 4.0 on all 12 items), whereas teachers reported more moderate satisfaction ratings (mean score = 3.8). Low teacher ratings (mean item ratings < 3.5) were reported for 2 of the 13 satisfaction items: "The program did not take much of my time" (M = 3.0) and "The program did not interfere with my other teaching activities/responsibilities" (M = 3.2).

Pre-Post Changes in Outcome Measures

Symptoms domain. An overall multivariate model was tested for three posttest problem behavior symptom measures, controlling for baseline levels, followed by

Table 3
Means and Standard Deviations for Baseline and Posttest Outcome Measures and ANCOVA Results

	Comparison (n = 99)			Intervention (n = 101)				Effect Size
Domain/Measure	Baseline $M(SD)$	Post $M(SD)$	M_{Adj}	Baseline $M(SD)$	Post $M(SD)$	M_{Adj}	p Value	d
Symptoms								
SSBD-MBI-Teacher	34.0 (8.7)	30.2 (9.3)	30.5	34.9 (8.0)	26.1 (9.4)	25.8	< .001[a]	–.62
SSRS-PB-Teacher	120.9 (11.0)	119.1 (10.8)	119.8	123.1 (10.3)	113.3 (12.6)	112.7	< .001[a]	–.73
SSRS-PB-Parent	111.1 (15.3)	109.5 (13.4)	109.8	111.9 (15.3)	103.3 (13.8)	103.0	< .001[a]	–.69
Functional social impairment								
SSBD-ABI-Teacher	32.9 (7.8)	35.3 (7.4)	35.0	31.9 (6.7)	40.7 (9.0)	40.9	< .001[a]	.82
SSRS-SS-Teacher	84.0 (9.8)	86.3 (8.7)	86.1	83.4 (8.7)	94.9 (14.5)	95.1	< .001[a]	.87
SSRS-SS-Parent	88.8 (14.4)	91.8 (15.1)	91.9	89.0 (14.8)	97.7 (15.6)	97.7	< .001[a]	.54
Academic								
SSRS-AC-Teacher	88.4 (11.6)	87.5 (11.0)	87.6	88.6 (10.2)	91.1 (10.5)	90.9	< .001[a]	.66
Student AET	41.7 (19.2)	48.3 (22.1)	48.6	42.8 (18.6)	56.8 (19.4)	56.5	.002[a]	.44
WJ-III Letter–Word Identification	97.6 (15.7)	100.0 (15.9)	101.3	100.3 (12.5)	101.0 (12.8)	99.7	.010[a]	–.37
Oral reading fluency (words per minute)	47.8 (36.5)	54.5 (38.1)	58.8	56.1 (41.7)	64.2 (43.4)	60.0	.354	.13

Note: M_{Adj} = Posttest mean adjusted for baseline levels; SSBD = *Systematic Screening for Behavior Disorders;* MBI = Maladaptive Behavior Index; SSRS = *Social Skills Rating System;* PB = Problem Behavior subscale; ABI = Adaptive Behavior Index; SS = Social Skills subscale; AC = Academic Competence subscale; AET = academic engaged time; WJ-III = *Woodcock–Johnson III Diagnostic Reading Battery*.
a. Significant after applying the Benjamini–Hochberg correction.

univariate ANCOVA models. The multivariate test was significant in which the intervention students were found to have large overall gains compared to students in the comparison condition, $F(3, 193) = 16.53$, $p < .001$, $\eta^2 = .20$. The intervention group differed significantly from the comparison group ($p < .001$) across all three symptom measures, with effect sizes ranging from $d = .62$ to .73 (see Table 3).

Functional impairment domain. The multivariate test on the three posttest functional impairment measures, controlling for baseline levels, was significant in which intervention students were found to have large overall gains compared to students in the comparison condition, $F(3, 193) = 18.26$, $p < .001$, $\eta^2 = .22$. The intervention group differed significantly from the comparison group ($p < .001$) across all three functioning measures (see Table 3). Effect sizes ($d > .80$) were obtained on teacher reports of adaptive behaviors and social skills, and an effect size of $d = .54$ was obtained on parent-reported social skills.

Academic domain. The multivariate test comparing the two conditions on the four posttest academic measures, controlling for baseline levels, was also significant $F(1, 191) = 8.54$, $p < .001$, $\eta^2 = .15$. The intervention group had significantly greater gains than the comparison group with respect to the SSRS Academic Competence subscale ($d = .66$) and AET ($d = .44$; see Table 3). Unexpectedly, however, the comparison group showed significantly greater improvement on the WJ-III Letter–Word Identification subtest ($d = –.37$) compared to the intervention group. Lastly, the two conditions did not differ significantly from each other with respect to gains in oral reading fluency.

Practical Significance of Intervention Effects

To evaluate the practical significance of the First Step program changes in student behavior, the percentile rank improvement index was calculated for each of the outcome measures in the three domains. With respect to the symptoms domain, the mean improvement index score was +25.1 percentile points, ranging from +23.2 to +26.7. Similarly, the mean improvement index score for the functional impairment domain was +26.8 percentile points, ranging from +20.5 to +30.7. The academic domain had the lowest mean improvement index score and widest range ($M = +8.2$, range = –14.4 to + 24.5). With the exception of the WJ-III Letter–Word Identification subtest, positive gains were obtained across all outcome measures and across all three domains.

Associations Between Process and Outcome Measures

Ancillary analyses were conducted with students assigned to the First Step condition to examine the associations between the process measures and change in outcome measures. Canonical correlation analysis was used to examine the magnitude of association between the set of significant outcome measures and (a) the coach, teacher, and parent quality of implementation measures and (b) the school intervention and homeBase dosage measures. Pre-post change scores were computed for each of the significant outcome measures listed in Table 3; the WJ-III Letter–Word Identification subtest was not included in the analysis given that the effects favored the comparison condition. Because the two annual cohorts differed with respect to means and variances on the quality of implementation measures, coach $t(96) = -8.20$, $p < .001$, Levene's $F = 10.82$, $p = .001$; teacher $t(97) = -9.63$, $p < .001$, Levene's $F = 4.52$, $p = .036$; parent $t(81) = 2.04$, $p = .044$, Levene's $F = 3.50$, $p = .065$, the analyses were conducted separately by cohort. The canonical correlations for the association between change in outcomes and quality of implementation ratings were $R = .67$ and $R = .52$ for Cohorts 1 and 2, respectively. Regarding dosage, the canonical correlations were $R = .60$ and $R = .54$ for Cohorts 1 and 2. The canonical correlations are considered to be within the medium to large effect size range according to Cohen (1988, p. 478). Hence, the quality of implementation and dosage received are considered to have affected the intervention effects that were achieved.

Discussion

With the exception of Beard's research on the First Step program (see Walker et al., 2005), to date the evidence base on First Step has been confined largely to studies of homogeneous, relatively nondiverse student populations concentrated mainly in suburban and rural school districts. The current investigation is the first scaled-up, randomized controlled trial of First Step conducted in a large, diverse, and urban school district. As noted, approximately 72% of the sample of 200 participants were students of color distributed across Hispanic, Black, Native American, Asian, multiracial, and Pacific Islander categories; 24.5% of the students listed themselves as White. The APS context thus provided an opportunity to examine the efficacy of First Step under conditions in which it had not been previously tested.

We are currently conducting 1-year follow-up assessments for Cohort 2 participants and 2-year follow-up assessments for Cohort 1 participants in our APS study. An important feature of the present efficacy trial was that it (a) allowed for evaluation of a procedure for fostering the sustainability of intervention gains achieved during the APS implementation process and (b) assessed the long-term durability of achieved behavioral gains and outcomes produced by the First Step program. Subsequent reports will describe the results of these follow-up assessments and the associated sustainability issues.

Results of this current study were encouraging in that moderate to strong effects were achieved for First Step participants in all three outcome assessment domains and for nearly all of the measures comprising them. Based on prior research and the results of this investigation, First Step appears to hold promise for effective application with students, teachers, and families across a range of school district settings. With a study sample composed of more than 70% students of color, this efficacy trial was a good test of the applicability of First Step to minority children, especially Hispanic children.

The First Step program is a relatively brief intervention of approximately 3 months' duration. For students who come from chaotic, at-risk backgrounds and who have well-developed externalizing behavior patterns, this delivered dosage may be comparatively small. Ideally, such targeted students should be exposed to some form of the First Step intervention across a full school year and hopefully followed up into the next school year to ensure a smooth transition between grades. One of our recommendations for future program adopters would be to implement First Step according to the standard implementation protocol and then to leave a low-cost variation of the program's procedures in effect for the remainder of the school year. Our experience over many case applications suggests that such a low-cost, maintenance procedure may sustain a substantial portion of the gains achieved during full implementation.

Neither the WJ-III subtest assessment nor the oral reading fluency measure proved sensitive to the First Step intervention. In fact, the WJ-III Letter–Word Identification task produced gain scores that significantly favored the usual care comparisons over the intervention participants, although the absolute levels for the two groups on the postassessments were nearly identical. The measures provide direct assessments of academic performance. However, less direct measures of achievement such as AET and the Academic Competence subscale of the SSRS were responsive to the intervention. Thus, it may be that a 3- or 4-month period is insufficient to register significant changes in direct academic performance but adequate for indirect measures. Further, given

the social-behavioral focus of First Step and the fact that the program does not directly teach academic skills, its failure to register on this dimension was not a surprise.

The process measures that were recorded to document the quality and integrity of the First Step implementation process, the quality of working relationships among implementers, whether dosage levels were adequate, and satisfaction indices for parents and teachers showed moderate to strong effects. Based on these measures, it seems safe to say that First Step was generally well implemented, that implementers and the coaches worked well together, that a majority of students received adequate dosage levels, and that there were relatively high levels of satisfaction associated with participation. However, a small proportion of participating First Step teachers reported that the program required too much time and effort. Whether this contributed to a weaker implementation effort by these teachers and/or a lack of significant gains in the academic test results of their students, as described above, are factors that are worth investigating in future research.

Parent participants appeared to have a high level of satisfaction with the First Step program, and they reported at least moderate levels of symptom and functional impairment gains in their children relative to those reported in the usual care comparison condition. This finding underlies the increasingly recognized importance of a parent component in improving school behavior problems (Diamond & Josephson, 2005). Although there were no peer-response measures of outcome included in this study, First Step also may well affect this domain when assessed by changes in sociometric status over time. Such measures may be important, given recent evidence that the influence of classroom peers may represent a critical environmental mediator, above and beyond established genetic risk for disruptive or aggressive behavior (Van Lier et al., 2007).

Working with First Step program originators, Rob Horner and his colleagues at the University of Oregon have conducted a 4-year study of weak and nonresponders to the First Step intervention (see Carter & Horner, 2007; in press). In addition to investigating participating student characteristics and contextual factors around the implementation process, this excellent work also examines the impact of performance feedback provided to First Step teachers regarding the quality of their implementation efforts using sensitive, single-case methodology. Subsequent revisions of the First Step intervention materials will incorporate these research results and provide program modifications based on our collective experience and findings in the current study.

Study Limitations

The limitations of this scaled-up First Step efficacy trial should be noted. First, approximately two thirds of the students in this study met the full SSBD Stage 2 criteria. The reasons that not all of the participating students met these criteria include (a) our strategy of recruiting teachers to participate in the study prior to conducting the SSBD screening, (b) parental decline of SSBD screening, and (c) only 79% of the first-ranked students participated in the study. As a result, the sample included greater variability in the severity of risk of students included in the study compared to a sample that was restricted to only those students meeting full SSBD Stage 2 criteria. However, analyses indicated that meeting full SSBD Stage 2 criteria did not moderate the association between study condition and student outcomes. Hence, the intervention effects can be considered to be comparable in magnitude for those students who met the Stage 2 criteria versus those who did not.

Second, it was not clear from this study what proportion of the sample would ultimately be referred and determined eligible as emotionally or behaviorally disordered under federal and state eligibility criteria for special education. This decision process is influenced by a host of school district fiscal, cultural, political, and other related factors. The emotional and behavioral disorders certification rate approximates 1% nationally but varies significantly from state to state and among school districts within states. Further, the teacher referral rate for this subpopulation does not reach its peak until approximately Grade 9 (see Lloyd, Kauffman, Landrum, & Roe, 1992). It seems clear that the participating members of our sample were experiencing some degree of behavioral adjustment problems as perceived by their general education teachers. However, it was not possible to calibrate their degree of specific risk for future maladaptive outcomes or their likelihood of eligibility for certification under the aegis of the Individuals with Disabilities Education Act or Section 504.

Finally, as numerous catalogues of evidence-based programs and practices have shown, the ultimate standard for judging the efficacy and effectiveness of interventions like First Step is whether (a) they move participants into the typical range and (b) they are sustained in this range across multiple years (Detrich et al., 2008; Forness, 2005; Hoagwood et al., 2007). Not withstanding the positive outcomes of the current efficacy study of the First Step intervention, the program does not, at this point, appear to meet this evaluative standard. The sustainability of program implementation and the concurrent maintenance of prior-achieved and socially

valid intervention effects remains a challenge for research in our field (Detrich et al., 2008). To date, relatively few school interventions appear to have met both of these criteria when used for judging evidence-based programs.

However, several smaller randomized controlled trials of a universal, classroom-wide prevention program were recently conducted in preschool classrooms that feed into the APS school district (see Serna, Nielsen, Lambros, & Forness, 2000; Serna, Nielsen, Mattern, & Forness, 2003). Although not directly related, both First Step and this APS-validated early intervention program could be considered evidence-based examples of primary and secondary interventions within a continuum-of-care system for young children at risk for emotional or behavioral disorders. It may be that selected programs such as First Step have the greatest impact when they are implemented in tandem within a primary prevention classroom or school context.

References

Adelman, H., & Taylor, L. (2006). Mental health in schools and public health. *Public Health Reports*, 121, 294–298.

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

Burns, B., & Hoagwood, K. (2002).*Community treatment for youth: Evidence-based interventions for severe emotional and behavioral disorders*. New York: Oxford University Press.

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

Carter, D., & Horner, R. (in press). Adding function-based behavioral support to First Step to Success: Integrating individualized and manualized practices. *Journal of Positive Behavioral Interventions*.

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

Detrich, R., Keyworth, R., & States, J. (2008). *Advances in evidence-based education: A roadmap to evidence-based education*. Oakland, CA: Wing Institute.

Diamond, G., & Josephson, A. (2005). Family-based treatment research: A 10-year update. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44, 872–877.

Dishion, T. J., Stormshak, E. A., & Siler, C. (in press). An ecological approach to interventions with high-risk students in schools: Using the Family Check-Up to motivate parents' positive behavior support.In M. Shinn, H. Walker, & G. Stoner (Eds.), *An ecological approach to interventions for achievement and behavior in a three-tier model including response to intervention*. Bethesda, MD: National Association of School Psychologists.

Drake, R., Latimer, E., Leff, S., McHugo, G., & Burns, B. (2004). What is evidence? *Child and Adolescent Psychiatric Clinics of North America*, 13, 717–728.

Forness, S. (2005). The pursuit of evidence-based practice in special education for children with emotional or behavioral disorders. *Behavioral Disorders*, 30, 311–330.

Forness, S., & Beard, K. (2007). Strengthening the research base in special education: Evidence-based practice and interdisciplinary collaboration. In J. Crockett, M. Gerber, & T. Landrum (Eds.), *Achieving the radical reform of special education* (pp. 169–188). Mahwah, NJ: Lawrence Erlbaum.

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

Furlong, M. J., Pavelski, R., & Saxton, J. (2002). The prevention of school violence. In S. Brock, P. Lazarus, & S. Jimerson (Eds.), *Best practices in school crisis management* (pp. 131–150). Washington, DC: National Association of School Psychologists.

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

Greenberg, M., Weissberg, R., O'Brien, U., Zins, J., Fredericks, L., & Resnik, H., et al. (2003). Enhancing school-based prevention and child development through coordinated social, emotional and academic learning. *American Psychologist*, 58(6/7), 466–474.

Gresham, F. M., & Elliott, S. (1990). *The Social Skills Rating System (SSRS)*. Circle Pines, MN: American Guidance Service.

Hawkins, D., Catalano, R., Kosterman, R., Abbott, R., & Hill, K. (1999). Preventing adolescent health-risk behaviors by strengthening protection during childhood. *Archives of Pediatrics and Adolescent Medicine*, 153, 226–234.

Hoagwood, K. (2003–2004). Evidence-based practice in child and adolescent mental health: Its meaning, application and limitations. *Emotional and Behavioral Disorders in Youth*, 4(1), 7–8.

Hoagwood, K. E., Olin, S. S., Kerker, B. D., Kratochwill, T. R., Crowe, M., & Saka, N. (2007). Empirically based school interventions targeted at academic and mental health functioning. *Journal of Emotional and Behavioral Disorders*, 15(2), 66–92.

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

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

Lipsey, M., & Derzon, J. (1998). Predictors of violence or serious delinquency in adolescence and early adulthood: A synthesis of longitudinal research. In R. Loeber & D. Farrington, (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 86–105). Thousand Oaks, CA: Sage.

Lloyd, J., Kauffman, J., Landrum, T., & Roe, D. (1992). Why do teachers refer students for special education? An analysis of referral records. *Exceptionality*, 2, 115–126.

Loeber, R., & Farrington, D. (2001). *Child delinquents*. Thousand Oaks, CA: Sage.

Nelson, J. R., Benner, G. J., & Mooney, P. (2008). *Instructional practices for students with behavioral disorders: Strategies for reading, writing, and math*. New York: Guilford.

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.

Reid, J. B., Patterson, G. R., & Snyder, J. J. (Eds.). (2002). *Antisocial behavior in children and adolescents: A developmental analysis*

and the Oregon Model for Intervention. Washington, DC: American Psychological Association.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE No. 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.

Serna, L. A., Nielsen, E., Lambros, K., & Forness, S. (2000). Primary prevention with children at risk for emotional or behavioral disorders: Data on a universal intervention for Head Start classrooms. *Behavioral Disorders*, 26, 70–84.

Serna, L. A., Nielsen, E., Mattern, N., & Forness, S. (2003). Primary mental health orientation in Head Start classrooms: Partial replication with teachers as interveners. *Behavioral Disorders*, 28, 124–129.

Severson, H. H., Walker, H. M., Doolittle, J. H., 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.

Shinn, M., & Walker, H. (in press). *Interventions for achievement and behavior in a three-tier model including response to intervention*. Bethesda, MD: National Association of School Psychologists.

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 Eck, K., Evans, S.W., & Ulmer, L.J. (2007). From evidence-based to best practices: What does it mean? *Report on Emotional and Behavioral Disorders in Youth*, 7, 35–40.

Van Lier, P., Boivin, M., Dionne, G., Vitaro, F., Brendgen, M., & Koot, H., et al. (2007). Kindergarten children's genetic vulnerabilities interact with friends' aggression to promote children's own aggression. *Journal of the American Academy of Child and Adolescent Psychiatry*, 46, 1080–1087.

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(2), 66–80.

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

Walker, H., Seeley, J., Small, J., Golly, A., Severson, H., & Feil, E. (2008). The First Step to Success program for preventing antisocial behavior in young children: Update on past, current and planned research. *Report on Emotional and Behavioral Disorders in Youth*, 8, 17–23.

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., 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.

Walker, H. M., Sprague, J. R., Perkins-Rowe, K. A., Beard-Jordan, K. Y., Seibert, B., & Golly, A. M., et al. (2005). The First Step to Success program: Achieving secondary prevention outcomes for behaviorally at-risk children through early intervention. In M. H. Epstein, K. Kutash, & A. J. Duchnowski (Eds.), *Outcomes for children and youth with emotional and behavioral disorders and their families: Programs and evaluation best practices* (2nd ed., pp. 501–523). Austin, TX: PRO-ED.

Walker, H. M., Stiller, B., Golly, A., Kavanagh, K., Severson, H. H., & Feil, E. G. (1997). *First Step to Success: Helping young children overcome antisocial behavior*.Longmont, CO: Sopris West.

Woodcock, R. W., Mather, N., & Schrank, F. A. (2004) *Woodcock–Johnson III Diagnostic Reading Battery*. Itasca, IL: Riverside Publishing.

Hill M. Walker, PhD, is codirector of the Institute on Violence and Destructive Behavior at the University of Oregon. His research interests include social skills assessment, curriculum development and intervention, longitudinal studies of aggression and antisocial behavior, school safety, youth violence prevention, and the development of early screening procedures for detecting students who are at risk for social-behavioral adjustment problems and/or later school dropout.

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

Jason Small, BA, is a data analyst at the Oregon Research Institute in Eugene. He has assisted with data management and analysis for multiple projects related to education and mental health.

Herbert H. Severson, PhD, is a senior research scientist at the Oregon Research Institute in Eugene and is a licensed psychologist with more than 30 years of experience in intervention and prevention research. He is codeveloper of the First Step to Success program, and he has a proven track record in conducting school-based applied research.

Bethany A. Graham, MA, is a behavior consultant with Albuquerque Public Schools. Her current interests are positive behavior supports, home–school connections, and classroom management.

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

Loretta Serna, PhD, is a full professor of special education at the University of New Mexico. Her current interests include emotional and behavior disorders, social skills, and school-based mental health.

Annemieke M. Golly, PhD, is an associate research scientist at the Oregon Research Institute in Eugene. She is the coordinator and trainer for the First Step to Success program. She conducts training in schoolwide, classroom, and individual behavior management.

Steven R. Forness, EdD, is a professor emeritus at the University of California, Los Angeles, in the Department of Psychiatry and Behavioral Sciences. His research focuses on eligibility of children for special education.