Efficacy Validation of the Revised First Step Program:

A Randomized Controlled Trial

Edward G. Feil, Hill M. Walker, Andy J. Frey, John R. Seeley, Jason W. Small,

Annemieke Golly, Jon Lee, & Steven R. Forness

Acknowledgments

This study was funded by the Institute for Education Sciences #R324A150221 to Oregon Research Institute. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education. This article was made possible, in part, by the support of the Jefferson County, Kentucky, Public Schools. Opinions contained in this report/article reflect those of the author and do not necessarily reflect those of the Jefferson County, Kentucky, Public Schools.

Accepted for publication: January 2020

Exceptional Children: https://journals.sagepub.com/home/ecx

Abstract

Disruptive behavior problems frequently emerge in the preschool years and are associated with numerous, long-term negative outcomes, including comorbid disorders. First Step is a psychosocial early intervention with substantial empirical evidence supporting its efficacy among young children (Walker et al., 2014). The present study reports on a validation study of the revised and updated First Step early intervention, called First Step Next (Walker, Stiller et al. 2015), conducted within four preschool settings. One hundred sixty students at risk for school failure, and their teachers, were randomized to intervention and control conditions. Results indicated coach and teacher adherence to implementing the core components of the program was excellent. Teachers and parents had high satisfaction ratings. For the three First Step Next pro-social domains, Hedges' g effect sizes ranged from .34 to .91. For the problem behavior domain, children who received the First Step Next intervention had significant reductions in teacher and parent-reported problem behavior as compared to children randomized to the control condition. For the problem behavior domain, Hedges' g effect sizes ranged from .33 to .63, again favoring the intervention condition. All of the domains were statistically significant. This study builds on the evidence base supporting the First Step intervention in preschool settings (Feil et al., 2014; 2016; Frey et al., 2015).

Keywords: Behavior disorders, preschool, early intervention, school and home, prevention

Efficacy Validation of the Revised First Step Program: A Randomized Controlled Trial

The onset of disruptive behavior problems, particularly oppositional defiant disorders, usually occurs in the preschool years and often precedes development of later comorbid disorders such as attention deficit hyperactivity disorder (ADHD), anxiety disorders, and depression (Burke et al., 2010; Egger & Arnold, 2006; Gresham, 2015). The delivery of early intervention services to prevent these outcomes thus assumes critical importance. There is good evidence that early interventions for behavior problems are efficacious. A meta-analysis of 36 randomized controlled trials (RCTs) on psychosocial interventions for young children (mean age = 4.7 years) demonstrated a large mean effect size (ES) of 0.8 (Comer et al., 2013). Another more recent meta-analysis of 28 RCTs on such interventions for children across a broader age span (mean age = 8.2 years) demonstrated that outcomes of preschool studies were about 0.4 ES larger than those found for school-age children (Epstein et al., 2015).

The focus of this article is on the First Step to Success early intervention, which is a tier 2 selected program for remediating and preventing externalizing behavior problems at the point of school entry (Walker et al., 1998). We report herein the results of a randomized controlled trial (RCT) of the recently revised and updated version of the intervention called First Step Next (FSN; Walker, Stiller, et al., 2015). Since its publication in 1997, the First Step program has been extensively researched in preschool and K-3 primary grade settings and supported by a series of federal and state-level grants. The accumulated evidence base for First Step is described in Walker et al. (2014).

There have been three prior RCTs supporting the efficacy of the original version of First Step to Success with children in kindergarten through third grade (Sumi et al., 2012; Walker et

al., 1998; Walker et al., 2009). A fourth RCT has been completed with adaptations of the original First Step program for preschoolers (Feil et al., 2014). In Feil et al. 2014, 126 preschool children with disruptive behavior were randomized to either a First Step or usual care condition. Effect sizes in favor of students in the First Step condition ranged from approximately 0.7 to 0.9 on teacher measures of adaptive behavior or social adaptation, and 0.3 to 0.4 on corresponding parent measures. Utilizing the same sample, we then separately examined the affect of First Step on subsamples of preschoolers at risk for comorbid psychiatric disorders. Children at risk for ADHD, for example, did particularly well (ESs ranged from 0.6 to 1.2) not only on the same outcome measures noted previously but also on measures specific to ADHD (Feil et al., 2016). Children in a subsample at risk for comorbid autism spectrum disorder (ASD) also did well but were slightly more variable in their outcomes especially in regard to ASD-specific measures (Frey et al., 2015). Children at risk for comorbid anxiety disorders also did relatively well on general outcome measures, but their anxiety failed to show significant improvement (Seeley et al., 2018).

Recently, the second author led a year-long effort to merge and standardize the original version of First Step (K-3nd grade) with the adapted preschool version into a single unified program serving the preschool through grade 2 age range call First Step Next (FSN). Major goals for the revision process included making the merged program more user-friendly for implementers, especially parents, and increasing the program's efficacy by adding new components and updating existing ones (Walker, Stiller et al., 2015). The revision process that resulted in the updated version of First Step preserved the core elements of the original program that seem to account for its prior efficacy demonstrations (i.e., direct instruction in school success skills, group and individual contingencies, peer and home support, school and home

reward activities, and a dense schedule of positive feedback and descriptive praise). A number of the original procedures remained unchanged, were revised or were only slightly updated. For example, a minor modification was the expansion of the coach phase from 5 to 7 program days pending the focus child's progress and the teacher's judgment. Specific additions included: (1) Super Student Skill Lessons, (2) different and more robust maintenance options and troubleshooting procedures, (3) a more formal debriefing component with the focus child and the coach with input from the teacher as appropriate, (4) a student safety and management plan procedure for dealing with the escalating, out of control student with whom general education teachers are usually not trained to cope and (5) additional supplemental materials (parent and teacher workbooks, coloring books, behavioral skill charts, and new demonstration videos). The Super Student Skill Lessons teach mastery of discrete social-emotional skills and academic enabling skills as follows: 1) Follow directions, 2) Be cool (anger management), 3) Be a team player (be in the right place, do the right thing, look around at your classmates for guidance), 4) Mistakes are okay, 5) Ask for help the right way, 6) Do your best work, and 7) Be safe. Although the classroom component of FSN remains relatively unchanged from the original, the home component focus shifted from parent engagement and support (i.e., school-home intervention) to only parent engagement (i.e., school intervention with parent involvement). Overall, the FSN revision team's goal was to make the program more streamlined, less complex, and easier to implement with integrity. A full description of the revision process is described in detail in a recent article (Walker et al., 2018). As well, a process evaluation was conducted on the revision (Feil et al., 2020).

Given that these modifications were made, the authors initiated this randomized evaluation to confirm that the changes did not substantially reduce the overall efficacy of the

intervention. Standards for developing an evidence-based practice demand that the practice produces positive efficacy outcomes, is capable of being successfully replicated, and also demonstrates that the practice can be scaled up thus leading to large-scale applications (Bacon et al., 2011; Feldstein & Glasgow, 2008; Flay et al., 2005; Wandersman et al., 2008). With its larger sample, the current study allows for a more systematic examination of these outcomes. In addition to a larger sample, the unit of randomization in this study was at a site (i.e., building) level rather than at the classroom level to better inform potential mediating factors that might also influence outcomes.

Method

Study Participants

Project staff recruited state- and federally-funded preschool programs in four states to participate in this study: Illinois, Indiana, Kentucky, and Oregon. After receiving IRB approval, we obtained consent to conduct the study from the program directors located in one county in Illinois; one county in Indiana; two counties in Kentucky; and two counties in Oregon. Project staff recruited teachers to participate in the study using a brief presentation. Across three cohorts, we invited 185 teachers from 51 programs for study participation. In total, the recruited teachers from 181 of 185 classrooms participated (98%) in the screening phase of the study (see Figure 1).

Prior to screening, project staff distributed a waiver of consent letter to teachers and asked them to give a copy of the letter to the parents of each student in their classroom. This letter explained the proposed study and described steps for declining participation in the class-wide screening process. If parents did not want their child to participate in screening, they returned the consent form either in-person or by mail via a pre-paid postcard to the teacher. Parents had two

weeks to return the card before screening began.

Participating teachers completed an abbreviated version of the *Systematic Screening for Behavior Disorders* (SSBD), a multi-stage screening procedure (Walker et al., 2014). During screening stage 1, we asked teachers to nominate and rank-order five children in their classroom based on the students' externalizing behavior. We gave teachers a detailed description of externalizing behavior problems to inform this initial screening stage. During stage 2, teachers completed three rating scales for each child previously identified during stage 1. These were the Adaptive Behavior Index (ABI), Maladaptive Behavior Index (MBI), and Aggressive Behavior Scale (ABS).

The 181 teachers who participated in the class-wide screening procedure provided behavioral rating scale data for 866 students. Participating teachers contributed screening data for at least five students in 153 of 181 classrooms (84.5%). On average, participating teachers completed stage 2 rating scales for 4.8 students in each classroom (*SD* = 0.8). We converted ABI, MBI, and ABS raw scores to severity scores corresponding to 1 *SD*, 1.5 *SDs*, and 2 *SDs* respectively from the normative mean (Feil et al., 1998). Severity scores ranged from 0 (within 1 *SD* of mean) to 3 (2+ *SDs* from mean) for each scale. We then summed the three individual severity scores to compute an overall severity ranking from 0 to 9 for each of the nominated students within each classroom. A child had to have elevated severity on at least one scale (e.g., elevated behavior of at least one *SD* above the mean) to meet minimum eligibility requirements. In seven classrooms, none of the students (n = 27) met minimum eligibility requirements. These classrooms did not participate in the parent recruitment process.

Project staff rank-ordered students according to severity within the remaining 174 classrooms and invited parents of the highest ranked child in each classroom to participate in the study. If the

parents of the highest-ranked child declined, project staff contacted the parents of the next highest ranked child in the classroom. Project staff repeated this procedure until obtaining parent consent for one eligible child in each classroom or until the families of all eligible children had declined participation. After screening, one program with three classrooms discontinued participation in the project; teachers from eight classrooms either withdrew or were non-responsive after screening. For three classrooms, we were unable to obtain parent consent for any eligible children. Thus, 160 classrooms from 50 recruited Head Start and preschool programs were eligible for randomization.

The preschool program was our unit of randomization. After screening and collection of baseline data, we randomly assigned 25 programs containing 77 classrooms to the intervention condition and assigned 25 programs containing 83 classrooms to the usual-care or control group condition. The average number of classrooms per program was comparable across conditions (t[48]= -.60, p = .550). The average cluster size was 3.1 classrooms per program (SD = 1.4) in the intervention condition and 3.3 classrooms per program (SD = 1.4) in the control condition.

As reported in Table 1, the age of participating children averaged four years. Nearly two-thirds of participating children were male (67%). The majority of students were either Caucasian (48%) or African American (36%). Just over 15% of participating children were Hispanic. Table 2 summarizes teacher and classroom characteristics. Almost all participating teachers were female (99%) and the majority were Caucasian (74%). Just over one-fifth of participating teachers were African American (21%). Teachers reported having taught for an average of 12.7 years (SD = 9.6). Most teachers had earned either a bachelor's degree (41%) or a master's degree (41%). A much smaller percentage reported having earned a high school diploma (4%) or an associate's degree (14%).

Participating parents, as reported in Table 3, had a mean age of 32 years. The majority of participating parents were female (88%) and just over half were Caucasian (53%). Most were employed (74%) but over half of the sample (55%) lived below the federal poverty level. Only 13% of participating parents reported holding a four-year degree. An examination of baseline equivalence across conditions and cohorts is discussed in the Results section.

Usual-Care Control Condition

In programs randomized to the usual-care condition, participating teachers were offered a 4-hour training session in classroom management and the principles of positive behavior support. During the training, teachers discussed their experiences with positive behavior support and learned strategies for promoting a positive classroom environment such as praise of appropriate behavior (Golly, 2006; Sprague & Golly, 2013). The workshop was designed to provide teachers in the usual-care condition with some degree of intervention support. However, the training was more generic in nature (e.g., did not provide specific intervention strategies) than the FSN training provided to teachers in programs randomized to the experimental condition. Teachers in the usual-care control group were eligible to receive training and implementation support in FSN during the following academic year.

Experimental Condition

Teachers in programs randomized to the experimental condition received a daylong workshop training session in the FSN intervention (Walker, Stiller, et al., 2015) and in the universal principles of classroom management (Golly, 2006; Sprague & Golly, 2013). During the first half of the workshop, teachers learned how to (a) develop and communicate behavioral expectations; (b) implement strategies to teach the expectations; (c) how to positively reinforce and manage expectations; and (d) to organize effective classroom environments (e.g., quiet-time

areas) and routines (e.g., transitions). During the second half of the workshop, teachers learned about FSN (described). Following training, a behavioral coach provided direct support to the focus child through initial implementation of FSN within the classroom and provided teachers with one-on-one consultation and supervision as needed in the teacher's classroom during instructional periods.

First Step Next

As noted earlier, FSN is a Tier 2 early intervention program for pre-K through grade 2 children that targets social skills and academic enablers central to promoting school success (Walker, Marquez, et al., 2015). The program includes three major tasks: social skills instruction, the green-card game, and home-school connections. During the social skills instruction task, a coach helps the target child in the classroom to master a set of "Super Student Skills" through delivery of one-on-one behavioral lessons that target social-emotional and academic skills such *as following directions, being safe, dong your best work, asking directions the right way, and being a team player*. During the green-card game, the coach or teacher – depending on the program phase – uses a laminated card with a green side and a red side to provide feedback to the target child and classmates (e.g., the green side) for complying with the teacher's expectations and non-verbal corrective feedback (e.g., the red side) when the child does not follow the teacher's expectations. At the outset of the program, the target children are taught that when the green side is visible they should continue with what they are doing but if the red side is visible, they should "stop, think, and get back on track" (Walker et al., 2018). For the home-school connection component of the program, parents receive a parent workbook focused on promoting positive parenting strategies that reinforce the skills the child is learning in the classroom and receive daily feedback in the form of a note or phone call from the FSN coach.

In general, a trained coach delivers the first 5-7 days of the program (e.g., coach phase) including delivery of behavioral lessons, implementation of the green-card game, and daily notes or phone calls to parents. Between days 8-10, the coach transitions control and management of the intervention to the teacher who begins full implementation of the intervention (e.g., teacher phase) from day 11 onward. During the teacher phase, the teacher supervises playing of the green-card game and, as needed, reviews the "Super Student Skills" with the target child individually.

FSN Implementation

As noted previously, a coach initially implements FSN. For this project, FSN coaches were employees of Oregon Research Institute, the University of Louisville, Head Start or an Educational Service Agency. In total, 21 coaches participated across the three cohorts of the program. The majority of coaches had a bachelor's degree or higher (76%). Coaches attended a two-day training session. During FSN training the coaches role-played (a) holding consent meetings with parents, (b) delivering "Super Student Skills" lessons to the target child, (c) introducing the program to all students in the classroom, and (d) implementing the first day of the program in the classroom. Additionally, coaches learned problem-solving strategies and how to use the daily summary chart and timing device. Research staff closely monitored coaches during initial implementation and, throughout implementation, program staff conducted frequent fidelity checks to ensure program implementation quality. To troubleshoot cases and minimize drift in program implementation, coaches attended weekly meetings with lead implementers at each site.

Data Collection Procedures

Project staff collected baseline data from teachers and parents prior to FSN randomization, training, and implementation. Staff mailed or hand-delivered questionnaire packets to participants. We provided participants with two options for returning packets. They could mail the packets back to us using a postage-paid envelope or project staff would pick up the packets from participants. We distributed post-intervention questionnaire packets using the same procedures. We distributed post-intervention questionnaires packets to participants after completion of the FSN intervention. The two conditions did not differ significantly on the average number of days between the collection of baseline and post intervention data ($t[152] = 1.31$, $p = .192$). We collected post questionnaire packets from participants randomized to intervention an average of 104 days ($SD = 28.5$) after collection of baseline data. For participants randomized to the control group, we collected post packets an average of 111 days ($SD = 32.3$) after baseline collection. Parents and teachers received $50 for the questionnaire packet they returned (i.e., screening, baseline, and post data packets). Spanish-speaking parents had the option to complete questionnaires in Spanish if they wanted. Six parents (3.8%) completed Spanish versions of the questionnaires.

Outcome Measures

Social Skills Improvement System rating scales (SSiS)

We collected the teacher-reported and parent-reported SSiS social skills and problem behavior scales as the primary outcome measures for this study (Gresham & Elliott, 2007). The SSiS social skills scale assesses behaviors that encourage positive interactions and minimize negative interactions with adults and peers in the classroom or home setting. The SSiS problem behavior scale assesses behaviors that impede an adaptive classroom adjustment (Gresham & Elliott, 2007). For social skills, both versions have 46 items. For problem behavior, the teacher

version has 30 items and the parent version has 33 items. Items across both scales are reported on a 4-point frequency scale (i.e., never, seldom, often, almost always). Coefficient alpha for this sample was high across all scales. For the social skills scale, coefficient alpha was .93 and .95 for the teacher-reported and parent-reported versions of the scale, respectively. For the problem behavior scale, coefficient alpha was .90 for teacher report and .93 for parent report. We converted raw scale scores to standard scores using gender-specific normative data from the SSiS manual.

Systematic Screening for Behavior Disorders scales (SSBD)

We included three stage 2 SSBD (Walker et al., 2014) subscales as secondary outcome measures in this study: Adaptive Behavior Index (ABI), Maladaptive Behavior Index (MBI), and Aggressive Behavior Scale (ABS; Feil & Becker, 1993; Feil et al., 1998, Walker et al., 2014). For the ABI, MBI, and ABS teachers rate the target child's behavior on a 5-point frequency scale ranging from *never* to *frequently*. For each scale, we computed a raw total score. The ABS, consisting of nine items, assesses the frequency of aggressive behavior ($\alpha = .77$). The ABI (8 items; $\alpha = .71$) and MBI (9 items; $\alpha = .78$) measure adaptive and maladaptive behavior, respectively. Although the SSBD was developed as a screening measure, other research studies with preschool children have demonstrated the ABS, ABI, and MBI are sensitive to target child, behavioral change (Gunn et al, 2006; Serna et al., 2000; Sumi et al., 2013; Walker et al., 2009). In the SSBD normative sample (N= 4,463), alpha levels were .94 and .92 for the ABI and MBI, respectively.

We grouped the SSiS and SSBD outcome measures into a pro-social behavior domain and a problem behavior domain to facilitate interpretation and discussion. The ABI and SSiS teacher-reported and parent-reported social skills scales make up the pro-social behavior domain.

The ABS, MBI, and SSiS teacher-reported and parent-reported problem behavior scales comprise the problem behavior domain. For the pro-social domain, the mean intercorrelation is .30 and, for the problem behavior domain, it is .42.

Relational Aggression

The relational aggression scale is a 6-item, teacher-reported subscale from the Preschool Social Behavior Scale – Teacher Form (Crick et al., 1997). This scale measures a child's relational aggression toward peers. For example, teachers indicate the extent to which the target child excludes other children from play groups, verbally threatens to exclude other children, or encourages other children not to play or be friends with a child in the classroom. The items are rated on a 5-point frequency scales. Raw total scores range from 6 to 30 with higher scores indicating higher levels of relational aggression. Coefficient alpha for the 6-item scale is high ($\alpha = .94$).

Child-Teacher Conflict

The 12-item child-teacher conflict scale ($\alpha = .89$) is one of three subscales from the Student-Teacher Relationship Scale (Pianta, 2001). The child-teacher conflict subscale assesses the extent to which the teacher perceives his or her relationship with the target child to be negative and defined by conflict. Items, rated on a 5-point scale, range from 12 to 60 with higher scores indicating higher levels of child-teacher conflict. According to Pianta, higher scores on the scale identify situations where the teacher struggles with the child and perceives the student's behavior as unpredictable.

Process Measures

Project staff collected a range of process measures either on or from participants in programs randomized to the intervention condition. Specifically, we collected fidelity data,

compliance data, alliance data, and satisfaction data. Each measure is described in greater detail subsequently.

Implementation Fidelity Checklist (IFC)

The IFC is an abbreviated version adapted from Walker et al. (2009). The 12-item IFC checklist assesses implementation tasks pertaining to general implementation (3 items), use of the green and red card (4 items), delivery of points and feedback (3 items), peer involvement (1 item), and school-home connections (1 item). For each item, the fidelity checklist assesses (a) delivery (e.g., adherence using a dichotomous "yes" or "no" scale) and (b) quality of delivery using a five-point scale (i.e., 0 = very poor, .25 = poor, .50 = okay, .75 = good, to 1.0 = excellent (α = .87). Observers collected data on one occasion during the coach phase and twice during the teacher phase. Inter-rater reliability collected on 24% of the fidelity checks conducted was excellent (ICC[3,1] = .96). We calculated coach, teacher, and overall classroom fidelity scores to assess adherence and implementation quality. Adherence scores are the proportion of critical program features implemented by the coach and teacher. As a measure of overall classroom adherence, we calculated a mean coach and teacher adherence score. Similarly, we calculated average quality ratings for teachers and coaches and combined them as a measure of overall classroom implementation quality.

Classroom Program Monitoring Form (CMF)

We collected coach- and teacher-completed CMF data to track the target child's compliance with daily goals during FSN implementation (Walker et al., 2009). On the CMF, the coach or teacher records the number of daily points possible, the number needed, the number the child earned, and if the focus child met criterion or a recycle day was necessary (i.e., in the recycling procedure, the program day was repeated if the child did not meet the daily reward

criterion). We calculated dosage as the proportion of program days the child completed successfully and compliance as the proportion of successful to total program days. Scores ranged from 0 to 1 for dosage and compliance.

Alliance survey

The coach and teacher each completed a measure of alliance (Walker et al., 2009) to assess their partnership as it related to program implementation. Coefficient alpha for this scale is excellent for the 10-item coach- ($\alpha = .92$) and 12-item parent versions ($\alpha = .91$) and good for the 10-item teacher-version ($\alpha = .81$). The survey evaluates aspects of the teacher or parent relationship with the coach (and vice versa). The teacher and coach rate each item on a 5-point frequency scale ranging from *never* to *always.* For each respondent, we calculated a mean total alliance score, ranging from 0 to 5 with higher scores indicating higher mean ratings of alliance.

Satisfaction survey

At the end of program implementation, teachers and parents also provided satisfaction data. The satisfaction measures assess perceptions of support, usability, and effectiveness and have been used in prior research (Sumi et al., 2012; Walker et al., 2009). Teachers reported their satisfaction with FSN on a 13-item measure ($\alpha = .92$), rated on a 5-point Likert-type scale from *strongly disagree* to *strongly agree* for each item. The 12-item parent satisfaction report ($\alpha = .92$) is scaled the same way. We calculated a mean total satisfaction score, ranging from 0 to 5 with higher scores indicating higher levels of satisfaction.

Analysis

For each outcome, we fit two-level random intercept regression models in Mplus 7.0 (Muthèn & Muthèn, 1998–2010). Each model included a level-1 covariate – the baseline value of the outcome of interest. The level-2 model included a dichotomous predictor indicating

intervention condition (1 = First Step Next, 0 = Control). To account for missing data in the models, we used the robust maximum likelihood (RML) estimator.

We report Hedges' *g* as a measure of effect size. Hedges' *g* is calculated by taking the difference between the mean outcome of each group and dividing it by the pooled within-group standard deviation (Hedges, 2007). Effect sizes of .2 and .5 are considered small and medium, respectively; whereas an effect size of .8 or higher is considered large. We applied the Benjamini-Hochberg correction to adjust for multiple comparisons (B-H; Benjamini & Hochberg, 1995). The B-H correction is applied by ranking outcomes in ascending order within domain by p-values and then applying a cutoff for each. For the three outcomes in the prosocial behavior domain, the rank-ordered effects are considered significant at a .05 level if *p*-values are below .017, .033, and .05 for each respective outcome. For the four outcomes in the problem behavior domain, rank-ordered effects are considered significant at a .05 level if *p*-values are less than .013, .025, .038, and .05.

Results

Baseline Equivalence

We examined the equivalence of the intervention and control conditions on child, teacher, and parent demographics at baseline and on the outcome measures at baseline. Child baseline and demographic characteristics are summarized in Table 1. Participating parents, children, and teachers in programs randomized to the FSN intervention condition did not differ significantly from those in programs randomized to the control condition on any of the demographic variables summarized in Tables 1 through 3. In terms of screening characteristics, the percent of 1st ranked children (76% vs. 79%) was comparable for children in the intervention and control conditions, respectively. Also, mean scores on the three SSBD screening measures were comparable for

these participants. For teachers, the number of years teaching was comparable across groups as were teacher-reported education levels. Teachers reported similar baseline levels of motivation to change their behavior and nearly identical levels of belief in their ability to manage classroom behavior (M = 59.6 vs. 59.8). A slightly higher percentage of Head Start classrooms and full-day classrooms were in programs randomized to the control condition; though neither of these differences were statistically significant. Participating parents were also comparable across conditions with similar percentages of parents in the intervention and control conditions reporting they held a bachelor's degree or higher (15% vs. 11%), were currently employed (77% vs. 71%), and were living below the federal poverty level (56% vs. 53%) based on reported annual household income.

Table 4 details the equivalence of the outcome measure at baseline. For the two conditions, there were no statistically significant differences in mean baseline scores for 9 of 10 outcomes. Parent-reported baseline scores on their child's level of problem behavior did differ significantly (p = .047) with parents in programs randomized to the control condition reporting slightly higher baseline scores (M = 120) than parents in programs randomized to the intervention condition (M = 114).

Attrition and Missing Data

Project staff collected baseline packets from all 160 participating teachers and baseline packets from 158 parents (99%). Post-intervention attrition rates were low. We collected post-intervention data from 96% of teachers and 94% of parents. Scale-level baseline missing data for teacher-reported outcomes ranged from 0% to 3%. For parent-reported baseline outcomes, scale-level missing data rates ranged from 3% to 5%. The percent of missing scale-level data on teacher-reported, post-intervention outcomes was 4%; whereas the percent of missing scale-level

data for parent-reported outcomes at post-intervention ranged from 6% to 11%. To test the assumption that data were missing completely at random (MCAR), we examined missing data patterns and Little's MCAR test. Little's MCAR test was non-significant (χ^2 = 194.11, *n* = 160, *p* = .545), suggesting the data are MCAR.

Fidelity, Program Compliance, Alliance and Satisfaction

Coach and teacher adherence to implementing the core components of the program was excellent. For coaches the average proportion of core components implemented was .99 (*SD* = .02). For teachers, the average proportion of core components implemented was .98 (*SD* = .04). Implementation quality varied by phase with higher quality implementation occurring when coaches were implementing (*M[SD]* = .93[.06]) and slightly lower quality implementation occurring during the teacher phase of the program (*M[SD]* = .84[.15]). Students received 78% of the requisite program days on average (range = 27% - 100%). On average, student compliance was excellent (M[SD] = .99[.02]). Both coaches (*M* = 4.62 on a 5-point scale) and teachers (*M* = 4.94) reported high levels of alliance with one another. Teachers and parents also reported favorable satisfaction ratings. Teachers, reporting on a 5-point scale, averaged mean satisfaction ratings of 4.36 (*SD* = .61) and parents had mean satisfaction ratings of 4.28 (*SD* = .52).

Posttest Differences on Outcome Measures

As can be seen in Table 5, the intervention and control groups differed significantly on the parent- and teacher-reported outcomes in the pro-social behavior domain. Parents and teachers reported statistically significant improvement at posttest in the prosocial functioning of children receiving First Step Next. For the three pro-social domains, Hedges' *g* effect sizes ranged from .34 to .91. For the problem behavior domain, children who received the First Step Next intervention had significant reductions in teacher and parent-reported problem behavior as

compared to children in programs randomized to the control condition. For the problem behavior domain, Hedges' *g* effect sizes ranged from .33 to .63. As noted earlier, for outcomes in the prosocial domain to be considered statistically significant at the .05 level using the B-H correction, the three rank-ordered outcomes must have *p*-values less than .017, .033, and .05, respectively. For outcomes in the problem behavior domain, the four rank-ordered outcomes must have *p*-values less than .013, .025, .038, and .05, respectively. After applying the aforementioned B-H criteria, the three outcomes in the prosocial domain and four outcomes in the problem behavior domain remained statistically significant at the .05 level.

Discussion

This research on the First Step program's recent revision, FSN, replicates the significant effects shown in previous RCTs (Feil et al., 2014; Sumi et al., 2012; Walker et al., 1998; Walker et al., 2009). As such, FSN remains an evidence-based approach to altering the trajectory of early onset disruptive behavior disorders, as well as subsequent comorbid disorders such as attention deficit hyperactivity disorder (ADHD), anxiety disorders, and ASD (Burke et al., 2010; Egger & Arnold, 2006; Frey et al., 2015; Gresham, 2015). Further, process data indicate that, similar to the original First Step to Success version, FSN can be implemented with fidelity and results in high satisfaction ratings from teachers and parents.

Noteworthy strengths of this study include high internal and external validity, multiple indicators to assess the main outcomes, and little missing data. With regard to internal validity, the randomization resulted in baseline equivalency, and attrition across conditions was also equal. Thus, all plausible threats to internal validity were controlled. External validity is solid because the intervention was successfully implemented in multiple preschool programs. The main effects were consistent across several indicators of prosocial behavior and problem

behavior. Across both domains, effect sizes for four of the seven measures were in the medium to large range, and results for all seven were statistically significant. Finally, missing data was minimal, with 150 of 160 parents completing baseline and post-test packets. The reader should note limitations include a lack of observational data, direct measure of pre-academic behaviors (e.g., early literacy skills) and maintenance within the school year or following year in Kindergarten.

The magnitude of effects for pro-social behaviors were slightly higher than were those for problem behavior, and teacher-reported effects were greater than parent-reported effects. This is consistent with Comer et al.'s (2013) meta-analysis of RCTs on psychosocial interventions for young children, which demonstrated a large mean effect size (i.e., .8). Results are also relatively similar to those produced by Feil, et al. (2014), where effect sizes in favor of First Step ranged from approximately 0.7 to 0.9 on teacher measures of behavior or social adaptation, and 0.3 to 0.4 on corresponding parent measures.

With regard to process data, coaches and teachers reported having strong alliances with the other, and satisfaction was high across items and raters. Also similar to previous First Step RCTs (See Sumi et al., 2012; Walker et al., 1998; Walker et al., 2009) overall adherence to core components and quality of implementation scores were high, with implementation quality being lower during the teacher phase than the coach phase.

As the first RCT since the First Step program was revised in 2014-2015, this study adds to its accumulated literature base by providing empirical support that the revised program, which unified the preschool and elementary versions and streamlined components to improve usability and satisfaction, retains its effectiveness in the preschool population (Feil et al., 2014; Walker et al., 2014). The effect sizes from this study are particularly interesting in light of the program

revisions for several reasons. First, it demonstrates, at least with the regard to its application with preschoolers, that standardizing the early elementary and preschool components into a unified program was successful. Second, the addition of the "super student skills" was considered a substantial content addition to the program and this shift in content focus did not seem to reduce effect sizes in comparison to previous studies. Third, the effect sizes from parent-reported measures remained in the small effect-size range indicating the reduced dosage of the intervention with parents may not have had noticeable effects.

Although this study has shown some robust results, there are three important issues that remain unaddressed. First, the lack of direct observational and academic performance measures would provide more convincing evidence of FSN's overall efficacy. Second, although the pre-post design of this study demonstrated short term benefits, their sustained effects later in the preschool year as well as into kindergarten would be a much better test of the program. Third, a cost-effectiveness analysis of the program has been needed to assist potential adopters, and to this end, a cost-analysis was recently completed (Frey et al., 2019).

There are some important additional areas needing examination in future research. First, research demonstrating these effects with a community-implementation sample (i.e., school-based as compared to a research-based implementation) would increase external validity, and be a significant resource to behavioral and educational providers. It might also be interesting to conduct subsample analyses to examine the FSN effects on students with risk status for ADHD, ASD, or comorbid anxiety disorders; and therefore, replicate previous findings in this regard (Frey et al., 2015; Feil et al., 2016; Seeley et al., 2018). Additionally, it is important to examine (a) longer-term behavioral and academic outcomes using longitudinal tracking methods, (b) the trajectory of behavioral and academic outcomes over time and (c) the parallel trajectory of

behavioral and academic outcomes in relation to one another. It would be particularly interesting to evaluate the effect of the intervention on academic achievement tests, special education status, exclusionary discipline, office disciplinary referrals, and attendance using archival school records (Walker et al., 1991). We are currently collecting long-term data on FSN outcomes and plan to present these results in subsequent articles. Finally, it is important to investigate more thoroughly contextual factors (e.g., classroom climate) as well as mediators and moderators, that impact intervention effects to guide future program applications.

References

Bacon, A., Walker, H. M., Schwartz, A. A., O'Hara, D. M., Calkins, C., & Wehmeyer, M. L. (2011). Lessons learned in scaling up effective practices: Implications for promoting self-determination within developmental disabilities. *Exceptionality*, *19*(1), 46-60. https://doi.org/10.1080/09362835.2011.537233https://doi.org/10.1080/09362835.2011.537233

Benjamini, Y., & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57,* 289-300.

Burke, J. D., Waldman, I., & Lahey, B. B. (2010). Predictive validity of childhood oppositional defiant disorder and conduct disorder: Implications for the DSM-V. *Journal of Abnormal Psychology*, *119*, 739-751. https://doi.org/10.1037/a0019708 https://doi.org/10.1037/a0019708

Comer, J. S., Chow, C., Chan, P. T., Cooper-Vince, C., & Wilson, L. A. S. (2013). Psychosocial treatment efficacy for disruptive behavior problems in very young children: A meta-analytic examination. *Journal of the American Academy of Child and Adolescent Psychiatry, 52*(1), 26-36. https://doi.org/10.1016/j.jaac.2012.10.001

Crick, N. R., Casas, J. F., & Mosher, M. (1997). Relational and overt aggression in preschool. *Developmental Psychology, 33*, 579-588. https://doi.org/10.1037/0012-1649.33.4.579

Egger, H. L., & Angold A. (2006). Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry, 47*, 313-337. https://doi.org/10.1111/j.1469-7610.2006.01618.x

Epstein, R. A., Fonnesbeck, C., Potter, S., Rizzone, K. H., & McPheeters, M. (2015). Psychosocial interventions for child disruptive behaviors: A meta-analysis. *Pediatrics, 136*(5), 947-960. https://doi.org/10.1542/peds.2015-2577

Feil, E. G. & Becker, W. C. (1993). Investigation of a multiple-gated screening system for preschool behavior problems. *Behavioral Disorders*, *19*(1), 44-53.

Feil, E. G., Frey, A., Walker, H. M., Small, J. W., Seeley, J. R., Golly, A., & Forness, S. R. (2014). The efficacy of a home-school intervention for preschoolers with challenging behaviors A randomized controlled trial of Preschool First Step to Success. *Journal of Early Intervention*, *36*(3), 151-170. https://doi.org/10.1177/1053815114566090

Feil, E. G., Small, J. W., Seeley, J. R., Walker, H. M., Golly, A., Frey, A., & Forness, S. R. (2016). Early intervention for preschoolers at risk for Attention-Deficit/Hyperactivity Disorder: Preschool First Step to Success. *Behavioral Disorders, 41*, 95-106. https://doi.org/10.17988/0198-7429-41.2.95

Feil, E. G., Small, J. W., Walker, H. M., Frey, A. J. Crosby, S. D, Lee, J. Seeley, J. R. Golly, A. Forness, S. (2020). The revision of First Step to Success: A process evaluation of FIRST STEP Next. Manuscript submitted for publication.

Feil, E. G., Severson, H. H., & Walker, H. M. (1998). Screening for emotional and behavioral delays: The Early Screening Project. *Journal of Early Intervention*, *21*(3), 252-266. https://doi.org/10.1177/105381519802100306

Feldstein, A. C., & Glasgow, R. E. (2008). A Practical, Robust Implementation and Sustainability Model (PRISM) for integrating research findings into practice. *The Joint Commission Journal on Quality and Patient Safety, 34,* 228-243. https://doi.org/10.1016/s1553-7250(08)34030-6

Flay, B., Biglan, A. Boruch, R., Castro, F., Gottfredson, D., Kellam, S., Moscicki, E. K., Schinke, S., Valentine, J. C. & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151-175. https://doi.org/10.1007/s11121-005-5553-y

Frey, A., Small, J. W., Feil, E. G., Seeley, J. R., Walker, H. M., & Forness, S. R. (2015). First Step to Success: Applications to preschoolers at risk of developing autism spectrum disorders. *Education and Training in Autism and Developmental Disabilities, 50,* 397-407. https://doi.org/10.17988/0198-7429-41.2.95

Frey, A. J., Kuklinski, M. R., Bills, K., Small, J. W., Forness, S. R., Walker, H. M., Feil., E. G., & Seeley, J. R. (2019). Comprehensive cost analysis of First Step Next for preschoolers with Disruptive Behavior Disorder: Using real-world intervention data to estimate costs at scale. *Prevention Science*, 1-14. https://doi.org/10.1007/s11121-019-01035-z

Golly, A. (2006). *Five universal principles of positive behavior support*. Verona, Wisconsin: Attainment Company, Inc.

Gresham, F. M. (2015). *Disruptive behavior disorders*. Guilford.

Gresham, F., & Elliott, S. N. (2007). *Social skills improvement system (SSIS) rating scales*. San Antonio, TX: Pearson Education Inc.

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

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*(2), 151–179.

Muthèn, L. K., & Muthèn, B. O. (1998-2010). *Mplus user's guide, sixth edition.* In. Los Angeles, CA: Psychological Assessment Resources, Inc.

Pianta, R. (2001). *Student-teacher relationship scale. Professional manual.* Lutz, FL: Psychological Assessment Resources.

Seeley, J. R., Small, J. W., Feil, E. G., Frey, A. J., Walker, H. M., Golly, A., & Forness, S. R. (2018). Effects of the first step to success intervention on preschoolers with disruptive behavior and comorbid anxiety problems. *School Mental Health*, *10*(3), 243-253. https://doi.org/10.1007/s12310-017-9226-3

Serna, L. A., Lambros, K. M., Nielsen, E., & Forness, S. R. (2000). Primary prevention with children at risk for emotional and behavioral disorders: Behavioral profiles and clinical implications of a primary prevention program. *Behavioral Disorders, 25*, 137-141. https://doi.org/10.1177/019874290002600107

Sprague, J., Golly A. (2013). *BEST Behavior. Building positive behavior support in schools. (2nd ed.)* Longmont, CO: Sopris Learning.

Sumi, W. C., Woodbridge, M. W., Javitz, H. S., Thornton, S. P., Wagner, M., Rouspil, K., Yu, J. W., Seeley, J. R., Walker, H. M., Golly, A., Small, J. W., Feil, E. G., & Severson, H. 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), 66-78. https://doi.org/10.1177/1063426611429571

Walker, H.M., Block-Pedego, A., Todis, B., & Severson, H. (1991, revised edition, 2014). *School Archival records search (SARS): Users guide and technical manual*. Pacific NW Publishing.

Walker, H. M., Feil, E. G., Frey, A. J., Small, J., Golly, A., Crosby, S, Lee, J., Forness, S., Sprick, M., Coughlin, C., & Stiller, B. (2018). First Step Next: An updated version of the

First Step to Success early intervention program. *Perspectives on Early Childhood in Psychology and Education*, *3(1)*, 89-110. ISBN: 978-1-935625-28-5

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. https://doi.org/10.1177/106342669800600201

Walker, H. M., Marquez, B., Yeaton, P., Pennefeather, J., Forness, S. R., & Vincent, C. G. (2015). Teacher judgement in assessing students' social behavior within a response-to-intervention framework: Utilizing what teachers know. *Education and Treatment of Children, 38*(3), 363-382. https://doi.org/10.1353/etc.2015.0019

Walker, H. M., Seeley, J. R., Small, J. W., Severson, H. H., Graham, B., Feil, E. G., Serna, L., Golly, A. M., & Forness, S. R. (2009). A randomized controlled trial of the First Step to Success early intervention: Demonstration of program efficacy outcomes within a diverse, urban school district. *Journal of Emotional and Behavioral Disorders, 17*(4), 197-212. https://doi.org/10.1177/1063426609341645

Walker, H. M., Severson, H., Seeley, J., Feil, E. G., Small, J., Golly, A., Frey, A. J., Lee, J., Sumi, C., Woodbridge, M., Wagner, M., & Forness, S. (2014). The evidence base of the First Step to Success early intervention for preventing emerging antisocial behavior patterns. In H. Walker & F. Gresham (Eds.), *Handbook of evidence-based practices for students having emotional and behavioral disorders* (pp. 518-536). Guilford, Inc.

Walker, H. M., Stiller, B., Coughlin, C., Golly, A., Sprick, M., & Feil, E. G. (2015). *First Step Next: Coach's Gui*de. Pacific Northwest Publishing.

Wandersman, A., Duffy, J., Flaspohler, P., Noonan, R., Lubell, K. Stillman, L., Blachman, M., Dunville, R., & Saul, J. (2008). Bridging the gap between prevention research and practice: The interactive systems framework for dissemination and implementation. *American Journal of Community Psychology, 41*(3-4), 171-181. https://doi.org/10.1007/s10464-008-9174-z

Table 1. Baseline equivalence of child demographic characteristics and screening measures.

	Total (n = 160)	Control (n = 77)	Intervention (n = 83)	Test statistic	*p*-value
Demographic characteristic					
Age *M(SD)*	4.1 (0.3)	4.1 (0.3)	4.1 (0.4)	-0.54	.588
Female *n* (%)	53 (33.1)	27 (35.1)	26 (31.3)	0.25	.616
Hispanic *n* (%)	25 (15.6)	10 (13.0)	15 (18.1)	0.78	.376
African American *n* (%)	58 (36.3)	25 (32.5)	33 (39.8)	0.92	.338
Caucasian *n* (%)	76 (47.5)	39 (50.6)	37 (44.6)	0.59	.442
Screening characteristics					
SSBD-ABS *M(SD)*	22.5 (6.3)	22.3 (6.2)	22.6 (6.5)	-0.37	.715
SSBD-ABI *M(SD)*	21.6 (4.1)	21.4 (3.9)	21.8 (4.4)	-0.52	.603
SSBD-MBI *M(SD)*	31.2 (6.3)	31.3 (6.5)	31.1 (6.2)	0.20	.842
1st ranked student *n* (%)	124 (77.5)	61 (79.2)	63 (75.9)	0.25	.616
2nd ranked student *n* (%)	29 (18.1)	14 (18.2)	15 (18.1)	0.01	.986

Note. Reported test statistics are *t* for continuous and χ^2 for dichotomous measures.

Table 2. Baseline equivalence of teacher and classroom characteristics.

	Total (n = 160)	Control (n = 77)	Intervention (n = 83)	Test statistic	*p*-value
Teacher characteristics					
Years teaching	12.7 (9.6)	12.0 (10.2)	13.4 (9.0)	-0.94	.350
Years at current school	6.5 (7.3)	6.6 (7.9)	6.4 (6.8)	0.21	.838
Percent Lead Teacher	158 (98.8)	76 (98.7)	82 (98.8)	0.01	.957
Percent Female	156 (97.5)	73 (94.8)	83 (100.0)	4.42	.051
Percent Hispanic	3 (1.9)	2 (2.6)	1 (1.2)	0.42	.609
Percent African American	33 (20.8)	18 (23.7)	15 (18.1)	0.69	.407
Percent Caucasian	118 (73.8)	55 (71.4)	63 (75.9)	0.41	.520
Education Level				5.36	.147
Percent HS diploma	7 (4.4)	2 (2.6)	5 (6.0)		
Percent AS/AA degree	22 (13.8)	14 (18.4)	8 (9.6)		
Percent BS/BA degree	65 (40.9)	34 (44.7)	31 (37.3)		
Percent MS/MA degree	65 (40.9)	26 (34.2)	39 (47.0)		
Motivation to change	102.6 (11.5)	101.1 (11.5)	104.0 (11.4)	-1.58	.116
Classroom management efficacy	59.7 (7.9)	59.8 (8.2)	59.6 (7.6)	0.17	.866
Classroom characteristics					
Number of personnel	2.1 (2.0)	2.3 (2.4)	2.0 (1.4)	0.84	.404
Percent Head Start	91 (58.3)	47 (63.5)	44 (53.7)	1.55	.213
Percent Full-day	95 (60.1)	49 (64.5)	46 (56.1)	1.15	.283

Note. Reported test statistics are *t* for continuous and χ^2 for dichotomous measures.

Table 3. Baseline equivalence of parent demographic characteristics.

	Total (n = 160)	Control (n = 77)	Intervention (n = 83)	Test statistic	*p*-value
Demographic characteristic					
Age *M(SD)*	32.3 (8.4)	32.1 (8.1)	32.5 (8.8)	-0.29	.775
Percent Female	139 (88.0)	66 (88.0)	73 (88.0)	0.00	.993
Percent Hispanic	21 (14.6)	7 (10.3)	14 (18.4)	1.90	.168
Percent African American	55 (34.8)	25 (33.3)	30 (36.1)	0.14	.711
Percent Caucasian	84 (53.2)	44 (58.7)	40 (48.2)	1.74	.188
Percent with BA/BS degree	20 (12.8)	8 (10.8)	12 (14.6)	0.51	.476
Percent currently employed	116 (74.4)	52 (71.2)	64 (77.1)	0.70	.402
Percent below federal poverty level	81 (54.7)	37 (52.9)	44 (56.4)	0.19	.665

Note. Reported test statistics are *t* for continuous and χ^2 for dichotomous measures.

Table 4. Baseline equivalence of child outcome measures.

Domain/Measure	Total (n = 160)	Control (n = 77)	Intervention (n = 83)	Test statistic	*p*-value
Pro-social behavior					
SSBD-ABI	21.6 (4.1)	21.4 (3.9)	21.8 (4.4)	-0.52	.603
SSiS-SS-Teacher	80.3 (10.1)	81.4 (9.9)	79.2 (10.3)	1.37	.173
SSiS-SS-Parent	92.3 (13.4)	90.5 (15.3)	93.8 (11.3)	-1.58	.117
Problem behavior					
SSBD-MBI	31.2 (6.3)	31.3 (6.5)	31.1 (6.2)	0.20	.842
SSBD-ABS	22.5 (6.3)	22.3 (6.2)	22.6 (6.5)	-0.37	.715
SSiS-PB-Teacher	125.5 (13.7)	125.1 (13.9)	125.8 (13.6)	-0.35	.730
SSiS-PB-Parent	117.0 (17.4)	120.0 (18.8)	114.4 (15.7)	1.99	.047
Relational Aggression	13.0 (6.8)	12.4 (6.5)	13.7 (7.0)	-1.21	.227
Child-Teacher Conflict	32.9 (10.6)	31.4 (10.4)	34.2 (10.7)	-1.65	.102

Note: SS = Social skills; PB = Problem behavior; ABI = Adaptive Behavior Index; MBI = Maladaptive Behavior Index; ABS = Aggressive Behavior Scale. Reported test statistics are *t* for continuous and χ^2 for dichotomous measures.

Table 5. Baseline and post-intervention means and standard deviation for outcome measures by condition and regression results.

Domain / measure	Control (n = 77)			Intervention (n = 83)			Condition effect		
	Baseline	Post-Intervention		Baseline	Post-Intervention				
	M(SD)	*M(SD)*	M_{Adj}	*M(SD)*	*M(SD)*	M_{Adj}	ω	*p-value*	*Hedges' g*
Pro-social behavior									
SSiS-SS-Parent	90.5 (15.3)	92.9 (15.8)	94.3	93.8 (11.3)	100.2 (11.0)	98.9	2.25	.024	0.34
SSiS-SS-Teacher	81.4 (9.9)	80.6 (12.9)	79.9	79.2 (10.3)	91.1 (13.2)	91.7	7.29	< .001	0.91
SSBD-ABI	21.4 (3.9)	23.7 (5.1)	23.8	21.8 (4.4)	27.9 (6.2)	27.9	4.60	< .001	0.73
Problem behavior									
SSiS-PB-Parent	120.0 (18.8)	118.1 (20.2)	116.0	114.4 (15.7)	107.9 (14.3)	110.0	-2.70	.007	0.34
SSiS-PB-Teacher	125.1 (13.9)	124.8 (15.6)	125.2	125.8 (13.6)	115.4 (17.0)	115.0	-4.80	< .001	0.63
SSBD-MBI	31.3 (6.5)	27.9 (7.0)	27.8	31.1 (6.2)	23.5 (7.8)	23.5	-4.39	< .001	0.59
SSBD-ABS	22.3 (6.2)	18.4 (5.9)	18.6	22.6 (6.5)	16.6 (7.1)	16.4	-2.37	.018	0.33
Relational Aggression	12.4 (6.5)	12.3 (6.4)	12.5	13.7 (7.0)	10.4 (5.4)	10.1	-2.78	.005	0.41
Child-Teacher conflict	31.4 (10.4)	31.4 (10.4)	32.6	34.2 (10.7)	25.1 (10.0)	24.4	-6.52	< .001	0.80

Note: SS = Social skills; PB = Problem behavior; ABI = Adaptive Behavior Index; MBI = Maladaptive Behavior Index; ABS = Aggressive Behavior Scale.

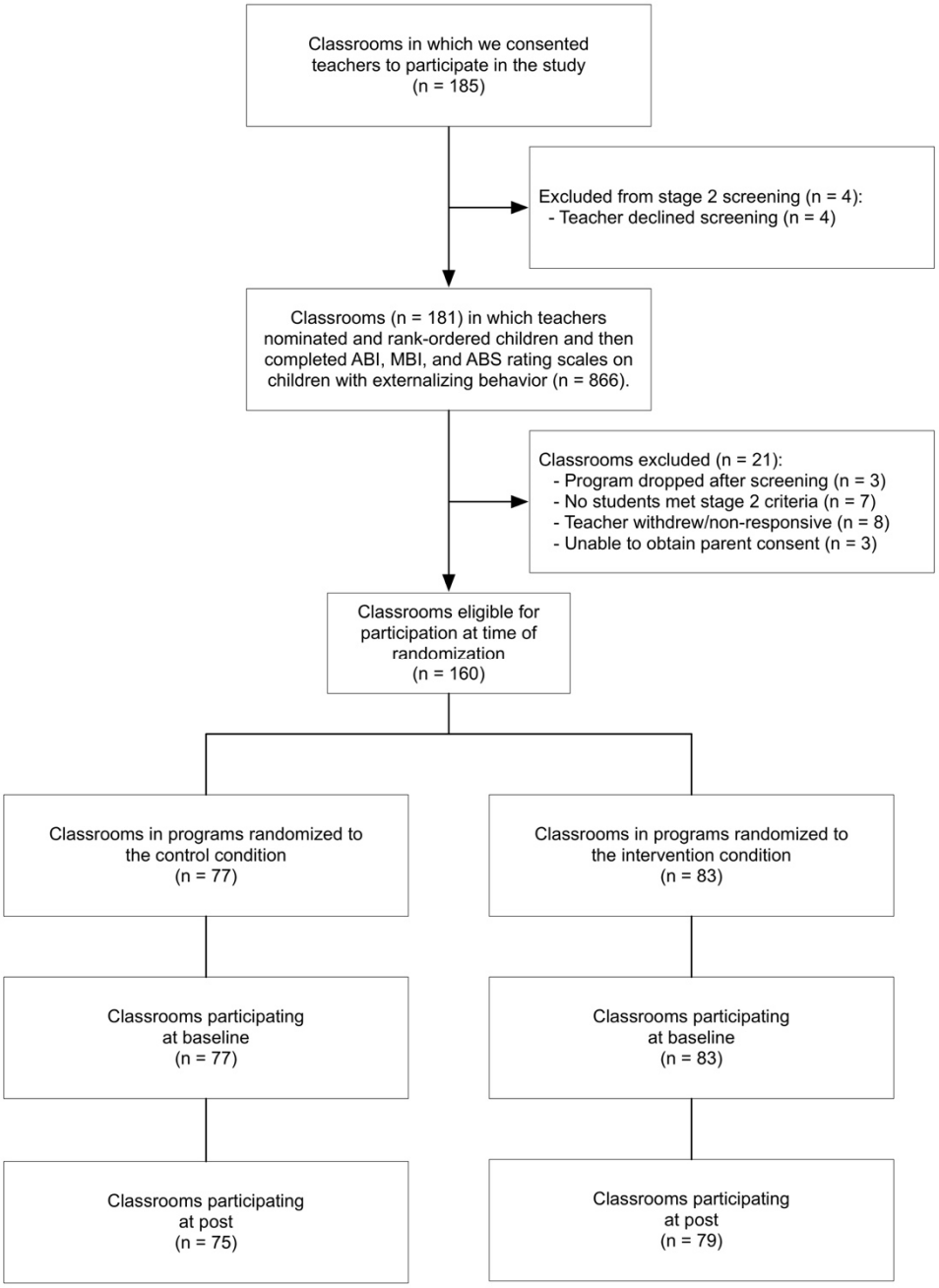

Figure 1. CONSORT diagram.