Hammill Institute on Disabilities

Remedial and Special Education
2023, Vol. 44(1) 60–72

DOI: 10.1177/07419325211068145
rase.sagepub.com
SAGE

First Step Next: A Best-Evidence Synthesis of Replication Randomized Controlled Trials From 2009 to 2021

Andy J. Frey, PhD[1], Jason W. Small, MPP[2], Hill M. Walker, PhD[2,3], Brandon Mitchell, MSW[1], John R. Seeley, PhD[2,3], Edward G. Feil, PhD[2], Jon Lee, PhD[4], and Steven R. Forness, EdD[5]

Abstract

Early intervention efforts have been effective in reducing disruptive behaviors and the probability of poor developmental outcomes. Early interventions include common practice elements to improve social functioning and decrease problem behaviors that disrupt the teaching-learning process. The First Step Next intervention has been well validated across early childhood and early elementary settings. In the present article, we provide a best-evidence synthesis of five randomized controlled replication studies between 2009 and 2021. Collectively, these studies show the intervention has resulted in small to large effect sizes and statistically significant improvements, compared with students randomized to control conditions, on multiple indicators of prosocial and problem behavior. The current synthesis focuses on a range of outcomes across elementary and preschool populations, as well as implementation fidelity and social validity ratings, to help school professionals in determining whether First Step Next might be useful in their efforts to support students at risk of school failure.

Keywords

effect sizes, behavior disorders, preschool, early intervention, replication studies

Addressing the needs of young children with early challenging behaviors is an important task for educators who strive to maintain positive learning environments and enhance school success for all children in preschool and primary grades. Kaminski and Claussen (2017) noted the results of several randomized controlled trials (RCTs) support the use of targeted interventions for supporting children with problem behaviors (PBs). Furthermore, replication is a critical tenet of scientific research and is necessary to establish a practice as evidence-based (Flay et al., 2005; Gottfredson et al., 2015; Makel et al., 2016). School practitioners rely on this evidence to evaluate for whom and under what conditions early intervention programs are appropriate given their resources and needs. Metrics such as statistical significance, effect sizes (ESs), improvement indices, and reliable change indices (Jacobson & Truax, 1991) provide evidence of statistical, practical, and clinical impact (Thompson, 2002). As well, measures of satisfaction provide evidence of a program's social validity and fidelity data provide evidence of a program's ease of implementation. The purpose of this review is to provide information about the impact, implementation, and social validity for the First Step Next (FSN) (see Note 1) early intervention program across a range of settings and student populations.

FSN is an early intervention for students in pre-K through Grade 2 and is implemented by general education teachers in collaboration with a coach, who is often a provider of special education services. FSN is a targeted intervention for students having moderately severe, behavioral challenges primarily of an externalizing nature. Cost analyses have estimated the FSN coach invests approximately 45 hr of time (for each additional student) to implement FSN (Frey, Kuklinski, et al., 2019; Kuklinski et al., 2021); coordination and collaboration activities take 2 to 3 months to

[1]University of Louisville, KY, USA
[2]Oregon Research Institute, Eugene, USA
[3]University of Oregon, Eugene, USA
[4]Northern Arizona University, Flagstaff, USA
[5]University of California, Los Angeles, USA

Corresponding Author:
Andy J. Frey, Kent School of Social Work, University of Louisville, Louisville, KY 40292, USA.
Email: afrey@louisville.edu

Associate Editor: Wendy Oakes

implement—depending on case difficulty. FSN contains the following elements: social skills (SS) instruction, green card game, and home–school connections; it is implemented across four implementation phases (i.e., preparation, coach, teacher, and maintenance phase).

The SS instruction element has seven super student skills, taught by the FSN coach in a one-to-one instructional situation outside the classroom. During SS instruction, a coach helps the target child improve positive personal relationships, develop problem-solving skills, and improve self-regulation through delivery of *super student skills* (e.g., following directions, being safe). While teaching the super student skills, the coach introduces the student to a color-coded card that functions as a tool for the teacher to provide subtle but direct, non-verbal feedback that encourages continued use of the super student skills or encourages the child to stop, think, and change their behavior when it is not consistent with the super student skills. For example, if during one-on-one instruction, the student walks around the room after being asked to stay in his or her seat, the coach would show the student the red side of the card. As soon as the student returns to his or her seat, the coach would turn the card to green and acknowledge the student's effort.

After the student understands how the green card works to provide feedback, the green card game begins and is played in the classroom with all students. Each day, prior to playing the green card game, an adult (coach or teacher) introduces it to the students. Specifically, they let the students know the entire class is going to play the green card game and if they win, there will be a reward activity. The target student is introduced as a special helper, and the adult says the student is going to work hard to keep the card on green and solicits input from the class on what they can do to help the student keep the card on green and help the class win the game. Once the game begins, the focus student can earn points for keeping the card on green by displaying the super student skills. The focus student earns points on a fixed interval schedule and the class "wins" the game and earns a reward activity if 80% of the point opportunities are earned. Initially lasting for only 20 min, the game's duration gradually increases over the course of the intervention until it covers the entire school day. The coach runs the game for approximately 10 days, and then gradually turns responsibility for the game over to the teacher. The teacher phase lasts for 10 days followed by a maintenance phase (also 10 days) in which the coach is involved on an as needed basis.

The FSN home–school connection element includes an initial meeting between the parent(s), teacher, and coach. During the coach phase, the coach also meets with the child's parent (or caregiver) to explain the parent's role in the home–school connection component. The parent is also asked to sign and return a daily home–school note and engage the child in a 5- to 10-min reinforcement activity after school on days the child succeeds in the green card game. There is also a super student skills resource for parents, which provides strategies for how parents can teach and reinforce the use of super student skills in the home setting. The purpose of this review is to provide information about the impact, implementation, and social validity of the FSN early intervention program across a range of settings and student populations.

Methodological Rigor

There is an existing 20-plus year history of empirical research on FSN, including single-subject, quasi-experimental, and experimental (i.e., RCT) study designs that are described in the work of Walker, Severson, Feil et al., (2014). Consistent with Slavin's (1995) recommendations for best-evidence syntheses, we identified for inclusion in this review all methodologically rigorous studies of FSN conducted since publication of the initial small-scale RCT of the program, which consisted of two cohorts of kindergarteners ($n = 46$) conducted within a school district in Eugene, Oregon (Walker et al., 1998). Thus, we report findings herein from five replication RCTs of FSN, each of which extended the findings of the original study, as well as data from four published subsample analyses. The subsample analyses focused on young children included in the larger study samples who were at risk of comorbid attention deficit hyperactivity disorder (ADHD), comorbid autism spectrum disorder (ASD), and comorbid anxiety disorders.

Following, we describe the common methods that characterized all five RCT replications. Next, we report the ES estimates, improvement indices, and statistical significance, as well as fidelity of implementation data and satisfaction data from teachers and parents, within and across studies. ES estimates and improvement indices empirically estimate the magnitude of effect or relative strength of an intervention as well as the degree of improvement one can expect from its implementation (U.S. Department of Education, n.d.). As such, they are generally recommended as a means of assessing the overall relative impact of an intervention as opposed to *p* values, which only document the existence of an intervention effect (Sullivan & Feinn, 2012).

These analyses also allow descriptive comparisons of FSN effects across student populations (preschool vs elementary), settings (parent vs. teacher reports of student progress at home or in school), efficacy versus effectiveness studies (investigator controlled vs. end user controlled), and type of disorder (ADHD, ASD, and anxiety).

Common Methods

In this section, the shared methods across the five RCT replication studies are described. Methods include screening procedures, interventionist training and supervision, fidelity monitoring, social validity, and outcome measures.

Screening Procedures

Screening procedures were comparable across the five primary studies. Specifically, all studies utilized a multiple-gating, universal screening approach (Walker, Severson, & Feil 2014; Walker & Severson, 1990). At screening Stage 1, teachers nominated and rank-ordered five students in their classroom based on descriptions and examples of externalizing behavioral profiles. Then, at Stage 2, teachers completed brief behavior rating scales on a subset of students identified during Stage 1. Two versions of multiple-gating screening procedures were used across these studies to accommodate developmental differences between the samples: These were (a) the Early Screening Project (ESP; Walker et al., 1995) for preschoolers and (b) the Systematic Screening for Behavior Disorders (SSBD; Walker, Severson, & Feil, 2014) for elementary students.

Interventionist Training and Supervision

The FSN program developers and research staff trained project assistants and school district staff to implement the FSN intervention across the reported studies. Training required a 1- or 2-day workshop that covered content about students with challenging behavior and provided detailed coverage of the FSN intervention procedures and implementation guidelines as well as teacher support materials. Workshops were didactic, consisting of but not limited to lectures, discussion, video demonstrations of key implementation tasks, role-playing, skills practice, and feedback sessions. Weekly supervision was maintained between experienced FSN coaches and teachers who jointly implemented the intervention. The coach, who was a research staff member, provided supervision and trouble-shooting assistance to interventionists for 1 hr weekly during FSN implementation.

Fidelity Monitoring

Across studies, expert raters collected implementation fidelity data on three to four occasions. The first round of fidelity monitoring was for the coach, during the initial days of program implementation (i.e., coach phase). The remaining fidelity checks (two to three depending on the study) were for the teacher implementing the program. Fidelity was monitored across studies using an implementation fidelity checklist (see the See Figure 1; Walker et al., 2009). The fidelity checklist examined implementation components, such as whether the implementer announced the number of points needed for the reward, elicited cooperation from the class, informed the class of the reward, delivered points and praise when prompted by the coach, provided positive feedback to the focus student during the green card game, and turned the card from green to red when inappropriate behavior occurred. For each implementation element, the fidelity checklist assessed (a) adherence (e.g., whether the coach or teacher implements a component) on a dichotomous scale (yes/no) and (b) the quality of implementation using a 5-point rating scale.

Social Validity

For each of the studies, teachers and parents randomized to an FSN condition submitted social validity data via completion of a satisfaction survey. Parents completed a 12-item satisfaction measure ($\alpha = .93$). Teachers completed a 13-item satisfaction measure ($\alpha = .91$). Satisfaction items for both sets of informants were reported on a 5-point Likert-type scale from *strongly disagree* (1) to *strongly agree* (5). Satisfaction items assessed usability, support, and program effectiveness. Higher scores indicated higher levels of satisfaction.

Outcome Measures

Baseline and post-intervention data were collected using teacher- and parent-reported measures. The Stage 2 SSBD measures included the teacher-completed Adaptive Behavior Index (SSBD-ABI) and the Maladaptive Behavior Index (SSBD-MBI; Walker, Severson, & Feil, 2014). The Social Skills Improvement System Rating Scale (SSiS-RS; Gresham & Elliot, 1990; Gresham et al., 2011) was used to assess prosocial and PB in all five RCT studies (see Note 2). The SSiS-RS provided a valid measure of peer-to-peer and teacher-related SS as well as a measure of the teacher's perception of important SS as they relate to successful school adjustment. The SSiS-RS contains a series of subscales completed by teachers (SSiS-TR) and parents (SSiS-PR) across two domains (SS [SSiS-SS] and PBs [SSiS-PB]. Finally, academic engaged time (AET; see Note 3) was observed and recorded as (a) the level of attention to the teacher and engagement with the academic material and task, (b) appropriate movement throughout the classroom and in response to teacher expectations, (c) appropriate requests, (d) appropriate interaction with peers and teachers, and (e) the ability to listen and follow directions. Trained project staff observed and recorded the focus child's AET for a minimum of three 20-min observations during structured classroom instruction led by the teacher. During the AET observations, the total amount of time that a student engaged in behavior consistent with the above definition of academic engagement was tracked using a stopwatch-type timing device turned on and off by the observer in response to the child's engagement. The sum of these values was divided by the total time for the three observation periods (typically 60 min) and multiplied by 100 to compute the average percentage of time the observed student was engaged. Average inter-observer agreement coefficients on this measure across

FIRST STEP Next

CLASS Fidelity Checklist

Student ID: ☐☐☐☐ Observer ID: ☐☐☐ Today's Date: ☐☐ / ☐☐ / 20☐☐ (Month / Day / Year)

Evaluation being completed for (MARK ONE): ○ Coach ○ Teacher Reliability: ○ Yes ○ No

Tchr/Coach Name: _____ Time of Game/Observation: ○ Beginning ○ Middle ○ End

Time of Day: ○ A.M. ○ P.M. ○ All Day Time of Implementation: _____ Program Day/Recycle Day: ☐☐

INSTRUCTIONS: The purpose of this form is to record your impressions of the extent to which the coach and/or teacher adheres to guidelines for correctly implementing the First Step Program. The judgments you are asked to make are inferential in nature but should be based upon the session during which you directly observe the First Step program being implemented by the teacher and/or coach. Please read each statement and make your rating on the 5 point scale provided (by filling in the appropriate circle for each statement) as to the degree or extent to which you consider it to be true. A rating of 1 means the statement is not at all true; a rating of 3 means the statement is somewhat true; and a rating of 5 means the statement is very true. See example below.

Part 1. Schedule of Implementation	**Component Implementation**		**Quality of Implementation**				
N.O.*	**Yes**	**No**	**Very Poor**	**Poor**	**Okay**	**Good**	**Excellent**
○ 1. Is the program being implemented daily?	○	○	①	②	③	④	⑤
○ 2. Is the Daily Checklist and Point Summary form followed (i.e., the implementer adheres to the correct session length, point interval, and daily point goal)?	○	○	①	②	③	④	⑤
○ 3. Is the Progress Monitoring Log completed after each session?	○	○	①	②	③	④	⑤

Part 2. Use of Green and Red Card	**Component Implementation**		**Quality of Implementation**				
N.O.*	**Yes**	**No**	**Very Poor**	**Poor**	**Okay**	**Good**	**Excellent**
○ 4. Is the Green and Red Card displayed so that it's visible to the student?	○	○	①	②	③	④	⑤
○ 5. Is the Green and Red Card used appropriately (i.e., implementer displays the green side when the student is behaving appropriately and displays the red side to signal inappropriate behavior)?	○	○	①	②	③	④	⑤
○ 6. Is negative attention minimized when the card is on the red side?	○	○	①	②	③	④	⑤
○ 7. If the student responds to the red side of the card by behaving appropriately, is the card turned back to green and is positive feedback delivered to the student?	○	○	①	②	③	④	⑤

*N.O. = Not Observed

Page 1 of 2 ☐☐☐☐

Figure 1. Implementation fidelity checklist.

Table 1. Child Demographic Characteristics and Screening Measures for First Step Next RCTs.

	Setting and grade				
	Kindergarten			Preschool	
Grades 1-3	RCT 1 (Walker et al., 2009) $N = 200$	RCT 2 (Sumi et al., 2013) $N = 286$	RCT 3 (Frey et al., in press) $N = 379$	RCT 4 (Feil et al., 2014) $N = 126$	RCT 5 (Feil et al., 2021) $N = 160$
Demographic characteristics					
Age, *M* (*SD*)	7.2 (1.0)	7.9 (1.0)	6.8 (1.2)	4.1 (0.4)	4.1 (0.3)
Percentage male	51	77	74	65	67
Percentage Black	7	24	52	18	36
Percentage White	24	45	37	72	48
Percentage Hispanic	57	27	4	5	16
Percentage special education	10	15	23	N/A	N/A
Percentage free/ reduced-price lunch	70	73	71	N/A	N/A
Screening measures					
Percentage ranked first on	79	60	68	75	78
SSBD/ESP ABI, *M* (*SD*)	NR	NR	NR	22.3 (5.8)	22.5 (6.3)
SSBD/ESP ABS, *M* (*SD*)	32.4 (7.2)	34.0	29.3 (6.5)	21.8 (4.8)	21.6 (4.1)
SSBD/ESP MBI, *M* (*SD*)	34.5 (8.3)	31.8	37.6 (7.3)	30.9 (6.1)	31.2 (6.3)

Note. For the SSBD externalizing dimension, students with an ABI score < 30 or an MBI score > 35, are at "extreme risk." For the ESP, an ABI < 21, an MBI > 25, or an ABS > 18 places a student in the "extreme risk" category on the respective scale (Walker, Severson, & Feil, 2014). RCT = randomized controlled trial; SSBD = Systematic Screening for Behavior Disorders; ESP = Early Screening Project; ABS = Aggressive Behavior Scale; NR = not reported; ABI = Adaptive Behavior Index; MBI = Maladaptive Behavior Index.

several published studies have ranged from .95 (Walker et al., 1994) to .98 (Quinn et al., 1995). Observers were trained to recognize examples of AET through multiple stages of video recordings and during in vivo, independent observations until 80% agreement was achieved.

Analytic Strategy

In this best-evidence synthesis, we report two objective indicators of impact for each of the five replication studies: ESs and the What Works Clearinghouse (WWC, 2020) improvement index. We have calculated a Hedges' *g* ES for each outcome or, when available, summarized previously reported ESs. Walker et al. (2009) and Sumi et al. (2013) reported Cohen's *d*, whereas all other studies reported Hedges' *g*. The two ESs are equivalent with larger samples (e.g., >50) and, therefore, can be interpreted as comparable measures for the studies reported haerein (Lakens, 2013). An ES of 0.2 is considered a "small" ES, whereas 0.5 and 0.8 represent a "medium" and "large" ES, respectively. The WWC improvement index represents the expected change in percentile rank for an average student in the control condition if the student had received the intervention. It captures the hypothetical improvement from the 50th percentile (e.g., no effect at all) that a control student would have made based on the difference between the mean values of a given outcome for the two conditions (WWC, 2020). Although the WWC has not adopted specific guidelines for determining the minimum percentile improvement necessary to be considered "meaningful," Durlak (2009) suggests a percentile improvement of 10% or more (e.g., ES ≥ 0.25) might be considered a reasonable (and "welcome") improvement on educational outcomes (p. 924).

Study Descriptions and Findings

Herein, we provide the narrative descriptions of each RCT study setting and sample, along with any deviations from the common methods. In addition, results are described for the five replication RCTs, as well as four subsample analyses. Table 1 provides demographic information for students from each of the studies and Tables 2 and 3, respectively, contain ES estimates for the RCT studies included along with ESs for each subsample analysis. Estimates are presented for school and parent ratings across each of the study outcome measures (i.e., prosocial, PB, AET) and for elementary versus preschool populations.

Randomized Controlled Trial 1. FSN Elementary Efficacy Study

The first large-scale replication study of the FSN intervention included 200 first- through third-grade students from 34 elementary schools within the Albuquerque Public

Table 2. Effect Sizes for First Step Next Randomized Controlled Trials.

	Elementary						Preschool			
	Walker et al. (2009)[a] – FSN efficiency replication		Sumi et al. (2013)[b] – FSN effectiveness		Frey et al. (2021)[c] – FSN & homeBASE comparative efficacy		Feil et al. (2014)[d] – FSN preschool efficacy		Feil et al. (2021)[e] – FSN preschool efficacy validation	
Measure	ES	WWCII	ES	WWCII	ES	WWCII	ES	WWCII	ES	WWCII
Prosocial behavior										
SSBD-ABI	0.82	+29.4	0.42	+16.3	0.51	+19.5	0.88	+31.1	0.73	+26.7
SSiS-SS-TR	0.87	+30.8	0.67	+24.7	0.75	+27.3	0.77	+27.9	0.91	+31.9
SSiS-SS-PR	0.54	+20.5	0.33	+12.8	0.24	+9.5	0.29	+11.4	0.34	+13.3
Problem Behavior										
SSBD-MBI	0.62	+23.2	0.36	+13.9	0.50	+19.1	0.71	+26.1	0.59	+22.2
SSiS-PB-TR	0.73	+26.7	0.38	+14.9	0.64	+23.9	0.79	+28.5	0.63	+23.6
SSiS-PB-PR	0.69	+25.5	0.21	+8.4	0.21	+8.3	0.45	+18.8	0.34	+13.3
AET	0.44	+17.0	0.35	+13.5	0.37	+14.4	NR	NR	NR	NR

Note. ES = effect size (Cohen's *d* for Walker et al. (2009) and Sumi et al. (2013); Hedges' *g* for all other studies; WWCII = What Works Clearinghouse Improvement Index; SSBD = Systematic Screening for Behavior Disorders; ABI = Adaptive Behavior Index; SSiS = Social Skills Improvement System; SS = social skills; TR = teacher report; PR = parent report; MBI = Maladaptive Behavior Index; PB = problem behavior; AET = academic engaged time; NR = not reported.
[a]*N* = 200. [b]*N* = 286. [c]*N* = 379. [d]*N* = 126. [e]*N* = 160.

Table 3. Effect Sizes for First Step Next Subsample Analyses.

	RCT 1[a] – Elementary		RCT 4[b] – Preschool					
	Seeley et al. (2009)[c] – ADHD		Frey et al. (2015)[d] – ASD		Feil et al. (2016)[e] – ADHD		Seeley et al. (2018)[f] – Comorbid anxiety disorder	
Measure	ES	WWCII	ES	WWCII	ES	WWCII	ES	WWCII
Prosocial behavior								
SSBD-ABI	0.80	+28.8	0.80	+28.8	1.16	+37.7	0.85	+30.2
SSiS-SS-TR	1.01	+34.4	0.80	+28.8	1.16	+37.7	0.60	+22.6
SSiS-SS-PR	0.31	+12.2	0.77	+26.7	0.65	+24.2	0.78	+28.2
Problem Behavior								
SSBD-MBI	0.96	+33.1	1.12	+36.9	1.09	+36.2	0.70	+25.8
SSBD-ABS	NR	NR	0.98	+33.6	1.14	+37.3	0.79	+28.5
SSiS-PB-TR	NR	NR	0.85	+34.8	1.05	+35.3	0.83	+29.6
SSiS-PB-PR	0.60	+22.6	0.87	+36.9	0.77	+27.9	0.93	+32.5
AET	0.76	+27.6	NR	NR	NR	NR	NR	NR

Note. RCT = randomized controlled trail; ADHD = attention-deficit/hyperactivity disorder; ASD = autism spectrum disorder; ES = effect size; WWCII = What Works Clearinghouse Improvement Index; SSBD = Systematic Screening for Behavior Disorders; ABI = Adaptive Behavior Index; SSiS = Social Skills Improvement System; SS = social skills; TR = teacher report; PR = parent report; MBI = maladaptive behavior index; ABS = Aggressive Behavior Scale; NR = not reported; PB = problem behavior; AET = academic engaged time.
[a]Walker et al. (2009). [b]Feil et al. (2014). [c]*N* = 42. [d]*N* = 34. [e]*N* = 45. [f]*N* = 38.

Schools (Walker et al., 2009). Randomization to conditions occurred prior to baseline assessment and was at the classroom level (i.e., one focus student per classroom). As noted in Table 1, students averaged 7.2 (*SD* = 1.0) years of age and were majority male (51%). Participating students were Black (7%), White (24%), and Hispanic (57%). Ten percent received special education services and 70% were eligible for free or reduced-price lunches.

As shown in Table 2, *Cohen's d* ESs ranged from 0.44 (moderately small) to 0.87 (large) and improvement indices ranged from +17.0 to +30.8 across prosocial, PB, and AET domains. Within the prosocial domain, *Cohen's d* ESs were medium to large, ranging from 0.54 (*improvement index* = +20.5) on the parent-reported SSiS-PB subscale to 0.87 (*improvement index* = +30.8) on the teacher-reported SSiS-PB subscale. Within the PB domain, ESs ranged from

0.62 (medium) on the SSBD-MBI to 0.73 (medium) on the teacher-reported SSiS-PB subscale, with corresponding improvement indices of +23.2 and +26.7. The ES for AET (0.44) was a moderately small-sized effect with an improvement index of +17.0. Across the teacher-reported outcomes, ESs ranged from 0.62 (medium; *improvement index* = +23.2) on the SSBD-MBI to 0.87 (large; *improvement index* = +30.8) on the SSiS-SS compared with parent-reported estimates of 0.54 on the SSiS-SS (medium; *improvement index* = +20.5) and 0.69 on the SSiS-PB (medium; *improvement index* = +25.5). All differences between intervention and control conditions were significant (Walker et al., 2009). Adherence to implementation was 84% during the coach phase and 82% during the teacher phase. Implementation quality averaged .85 (*SD* = 0.12) and 0.80 (*SD* = 0.11) for the coach and teacher phases, respectively. The average of teacher and parent satisfaction scores was 3.8 (*SD* = 0.7) and 4.3 (*SD* = 0.6), respectively.

A subsample analysis was also conducted as part of the Walker et al. (2009) efficacy study and is described below. Specifically, the efficacy of the intervention for students with or at risk of ADHD was examined by Seeley et al. (2009).

Attention-Deficit/Hyperactive Disorder

The ADHD subsample was identified from the RCT above using the clinical cutoff for teacher report on ADHD symptomatology (Seeley et al., 2009). Specifically, participants were considered at risk of developing ADHD if they presented six or more symptoms on the 18-item version of the Connors' ADHD/*Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM-IV*; American Psychiatric Association, 1994) scale (CASD-T; Conners, 1999). Forty-two students (21%) met the criteria for an ADHD diagnosis, with 23 students from the intervention group (22.8%) and 19 from the control group (19.2%). In addition to measures previously described, teachers completed the Achenbach Teacher Report Form (TRF/6-18) *DSM-IV* oriented subscale for oppositional defiant disorder (ODD; Achenbach & Rescorla, 2001).

As shown in Table 3, ESs across prosocial, PB, and academic engagement domains ranged from 0.31 (small; *improvement index* = +12.2) on the parent-reported SSiS-SS to 1.01 (large; *improvement index* = +34.4) on the teacher-reported SSiS-SS. Similarly, within the prosocial domain, *Cohen's d* ranged from small (0.31) on the SSiS-PR to large (1.01) on the SSiS-TR, with corresponding improvement indices of +12.2 and +34.4, respectively. Within the PB domain, ESs ranged from 0.60 (medium; *improvement index* = +22.6) on the SSiS-PR to 0.96 (large; *improvement index* = +31.1) on the SSBD-MBI. Seeley et al. (2009) also reported on measures indicative of ADHD symptomatology (not reported in Table 3 for all teacher-reported measures). For example, ESs of 1.32 (large), 0.82 (large), and 0.74 (medium) were observed for the SSiS-hyperactive subscale, the SSiS-Inattentive subscale, and the TRF/ODD subscale, respectively. All differences between intervention and control conditions were significant for the teacher-reported measures and none were significant for the parent-reported measures (Seeley et al., 2009).

Randomized Controlled Trial 2. FSN Effectiveness Study

The national effectiveness trial of FSN by Sumi et al. (2013) was an independent replication study in that a different research group (from Stanford Research International) conducted the study with minimal support from the program developers. It involved 48 elementary schools across five U.S. states (Oregon, West Virginia, Illinois, California, and Florida). Over the course of this 4-year study, 24 schools were randomly assigned to either the experimental (n = 24) or control (n = 24) condition after collection of the baseline assessments. Two hundred eighty-six teacher–student–parent triads participated. Student participants averaged 7.9 (*SD* = 1.0) years of age and were majority male (77%). Students in this diverse sample were Black (24%), White (45%), and Hispanic (27%). Fifteen percent were receiving special education services and 73% percent were eligible for free or reduced-price lunches. Coaches, who were school employees for this effectiveness study, represented a wide range of staff roles (special education teachers, counselors, social workers, psychologists, and graduate students) and educational levels. Throughout the project, teachers and coaches were encouraged to take advantage of technical assistance made available to staff from any school that implemented the FSN program. Technical assistance involved emails or conference calls with the FSN trainer to problem-solve and trouble-shoot implementation problems.

As shown in Table 2, ESs across prosocial, PB, and academic engagement domains ranged from 0.21 (small; *improvement index* = +8.4) on the parent-reported SSiS-PB subscale to 0.67 (medium; *improvement index* = +24.7) on the teacher-reported SSiS-SS subscale. Within the prosocial domain, *Cohen's d* ESs were all in the small to medium range, from 0.33 (*improvement index* = +12.8) on the SSiS-PR to 0.67 (*improvement index* = +24.7) on the SSiS-TR. Within the PB domain, ESs were all in the small range. A small ES (0.35; *improvement index* = +13.5) was also observed for AET. Across the prosocial and PB domains, ESs for teacher-reported measures ranged from 0.38 (small; *improvement index* = +14.9) on the SSiS-PB to 0.67 (medium; *improvement index* = +24.7) on the SSiS-SS compared with parent-reported measures of 0.21 (SSiS-PB) and 0.33 (SSiS-SS). Improvement indices for the parent-reported measures ranged from +8.4 to +12.8. All

differences between intervention and control conditions were significant (Sumi et al., 2013).

Adherence to implementation was 76% for coaches and teachers, and implementation quality was in the good to excellent range ($M = 0.78$, $SD = 0.15$) for overall classroom implementation. The average of teacher and parent satisfaction scores was 3.54 ($SD = 0.69$) and 4.21 ($SD = 0.62$), respectively.

Randomized Controlled Trial 3. FSN and homeBase Comparative Efficacy Study

The FSN and homeBase comparative efficacy study examined the impact of FSN and homeBase, a recently developed intervention targeting parents of young children with early onset behavior problems (Frey et al., 2015; Frey, Small, et al., 2019). Over this 5-year study, 379 triads participated (Frey et al., in press). Randomization to four conditions (FSN only, homeBase only, FSN plus homeBase, and control) occurred after collection of baseline assessments at the student level. Participating students averaged 6.8 ($SD = 1.2$) years of age and were majority male (74%). Participating students were primarily Black (52%) and White (37%). Twenty-three percent received special education services and 71% were eligible for free or reduced-price lunches.

Coaches, who were research staff, spanned a wide range of professional backgrounds (e.g., special education teachers, counselors, social workers, graduate students) and educational levels. In addition to being trained in FSN implementation procedures during a 2-day training session, coaches practiced implementing the intervention with one to three students who were not part of the study before implementing it with students who were actually part of the study.

As show in Table 2, ESs across prosocial, PB, and academic engagement domains ranged from 0.21 (small; *improvement index* = +8.3) on the parent-reported SSiS-PB subscale to 0.75 (medium; *improvement index* = +27.3) on the teacher-reported SSiS-SS subscale. Within the prosocial domain, *Hedges*' *g* ESs were in the small to medium range, ranging from 0.24 (*improvement index* = +9.5) on the SSiS-PR to 0.75 (medium; *improvement index* = +27.3) on the SSiS-TR. Within the PB domain, ESs varied from 0.21 (small; *improvement index* = +8.3) on the SSiS-PR to 0.64 (medium; *improvement index* = +23.9) on the SSiS-TR. The ES for academic engagement (0.37) was small (*improvement index* = +14.4). Across the prosocial and PB domains, ESs for teacher-reported measures ranged from 0.50 (medium; *improvement index* = +19.1) on the SSBD-MBI to 0.75 (medium; *improvement index* = +27.3 on the SSiS-SS, compared with parent-reported measures ranging from 0.21 on the SSiS-PB (small; *improvement index* = +8.3) to 0.24 on the SSiS-SS (small; *improvement index* = +9.5). The differences between intervention and control conditions were significant for 4 of the 5 teacher outcome measures and AET; differences for the two parent-reported measures were not significant.

Adherence to implementation was high during the coach phase (M [SD] = 0.99 [0.04]) and teacher phase (M [SD] = 0.98 [0.07]). Quality of implementation was also high during both phases. Average quality ratings during the coach and teacher phases were 0.96 ($SD = 0.04$) and 0.90 ($SD = 0.09$), respectively. Parents randomized to the FSN only condition reported slightly higher mean levels of overall satisfaction (M [SD] = 4.10 [0.61]) than did teachers (3.94 [0.68]).

Randomized Controlled Trial 4. FSN Preschool Efficacy Study

The initial FSN efficacy trial for preschoolers was conducted in Oregon, Indiana, and Kentucky (Feil et al., 2014). The study included 32 Head Start and preschool programs across two counties in Oregon and 31 Head Start and preschool programs across two counties in Kentucky and Indiana. One teacher–student–parent triad participated from each of 126 classrooms. Randomization to condition occurred prior to baseline assessment and was at the classroom level. Students averaged 4.1 ($SD = 0.40$) years of age and were majority male (65%). Participating students were predominantly White (72%), Black (18%), and Hispanic (5%).

As shown in Table 2, ESs across prosocial and PB domains ranged from 0.29 (small; *improvement index* = +11.4) on the parent-reported SSiS-PB subscale to 0.88 (large; *improvement index* = +31.1) on the teacher-reported SSBD-ABI. Within the prosocial domain, *Hedges*' *g* ESs ranged from small (0.29; *improvement index* = +11.4) on the SSiS-PR to large (0.88; *improvement index* = +31.1) on the teacher-reported SSBD-ABI. Within the PB domain, ESs ranged from 0.45 (medium; *improvement index* = +18.8) on the SSiS-PR to 0.79 (medium; *improvement index* = +28.5) on the SSiS-TR. Across the prosocial and PB domains, ESs for teacher-reported measures ranged from 0.71 (medium; *improvement index* = +26.1) on the SSBD-MBI to 0.88 (large; *improvement index* = +31.1) on the SSBD-ABI compared with parent-reported measures of 0.29 (SSiS-SS; small; *improvement index* = +11.4) and 0.45 (SSiS-PB; medium; *improvement index* = +18.8). All of the differences between intervention and control conditions were significant (Feil et al., 2014).

Three subsample analyses were also conducted as part of this RCT. Specifically, analyses examining the efficacy of FSN on young students with or at risk of developing ASD and ADHD were completed, as well an analysis of preschoolers with comorbid externalizing and internalizing disorders.

Adherence to core protocol components of the program was excellent during both coach (95%) and teacher (95%) phases. The quality of implementation was excellent during the coach phase (M [SD] = 0.92 [0.06] and good during the teacher phase (M [SD] = 0.78 [0.15]). Mean teacher- and

parent-reported satisfaction scores were 4.36 (*SD* = 0.54) and 4.34 (*SD* = 0.65), respectively.

ASD (preschool). This ASD subsample analysis (Frey et al., 2015) included children who were also identified as at risk of developing comorbid ASD as a result of being 2 *SD*s above the mean on the parent report of the Early Childhood Inventory–4 (ECI; Gadow & Sprafkin, 2000). The ECI is a screening tool comprised of diagnostic criteria specified in the *DSM-IV*. Thirty-four of 126 participants from the larger RCT (27%) met the ASD criteria, and students were evenly divided in the intervention group and control groups. Participating students averaged 4.1 (*SD* = 0.04) years of age and were majority male (76%).

Table 3 displays ESs across prosocial and PB domains that ranged from 0.77 (medium; *improvement index* = +26.7) on the parent-reported SSiS-SS to 1.12 (large; *improvement index* = +36.9) on the teacher-reported SSBD-MBI. Within the prosocial domain, *Hedges*' *g* ESs ranged from medium (0.77) on the SSiS-PR to large (0.80) on the SSiS-TR, with corresponding improvement indices of +26.7 and +28.8, respectively. Within the PB domain, *ES*s ranged from 0.85 (large; *improvement index* = +34.8) on the SSiS-TR to 1.12 (large; *improvement index* = +36.9) on the SSBD-MBI. In addition, Frey et al. (2015) reported on measures indicative of ASD symptomatology (not shown in table). For example, medium ESs of 0.48 and 0.52 were obtained for the teacher- and parent-reported SSiS-Communication subscales, respectively. Furthermore, the ESs for the teacher-reported SSiS-Empathy subscale was 0.29 (small), while the parent-reported SSiS-empathy subscale was 0.57 (medium). Finally, the SSiS-parent- and teacher-reported SSiS-ASD subscale ESs were in the medium range (0.65 and 0.77, respectively).

ADHD (preschool). The next subsample analysis included preschoolers who were also identified as at risk of developing comorbid ADHD (Feil et al., 2016). This sample included participating students whose parent and teacher-reported baseline scores exceeded one standard deviation above the mean on the Conners' ADHD Scales (CADS; Conners, 1999). Applying this cutoff, 45 (35%) participants met inclusion criteria; 19 of the 45 had been randomized to the control condition and 26 had been randomized to the FSN condition.

As shown in Table 3, ESs across prosocial and PB domains ranged from 0.65 (medium; *improvement index* = +24.2) on the parent-reported SSiS-SS to 1.16 (large; *improvement index* = +37.7) on both the teacher-reported SSBD-ABI and the teacher-reported SSiS-SS. Within the prosocial domain, *Hedges*' *g ES*s ranged from medium (0.65) on the SSiS-PR to large (1.16) on the teacher-reported SSBD-ABI and the teacher-reported SSiS-SS. Within the PB domain, ESs ranged from 0.77 (medium; *improvement index* = +27.9) on the SSiS-PR to 1.14 (large; *improvement index* = +37.3) on the SSBD-Aggressive Behavior Scale (ABS). In addition, Feil et al. (2016) reported on measures indicative of ADHD symptomatology, respectively (not shown in Table 3). For example, large ESs of 1.40, 0.92, and 1.10 were observed for the SSiS-cooperation, self-control, and hyperactive subscales, respectively.

Comorbid anxiety disorders (preschool). The final subsample analysis included children who were identified as also having behavior problems with comorbid anxiety (Seeley et al., 2018). As in the ASD preschool subsample analysis above, Seeley et al. used the ECI-4 (Gadow & Sprafkin, 2000) to identify the subsample. For this analysis, the *DSM-IV* criteria for separation anxiety disorder, generalized anxiety disorder (GAD), social phobia, and specific phobia, obsessive-compulsive disorder (OCD), and posttraumatic stress disorder (PTSD) identified participants at risk of comorbid anxiety problems. Thirty-eight (30%) of the 126 study participants met inclusion criteria; 19 of the 38 were randomized to the control condition and 19 were randomized to the FS condition.

As shown in Table 3, ESs across prosocial and PB domains ranged from 0.60 (medium; *improvement index* = +22.6) on the parent-reported SSiS-SS to 0.93 (large; *improvement index* = +32.5) on the parent-reported SSiS-PB. Within the prosocial domain, *Hedges*' *g* ESs ranged from medium (0.60) on the SSiS-TR to large (0.85) on the teacher-reported SSBD-ABI. Within the PB domain, ESs ranged from 0.70 (medium; *improvement index* = +25.8) on the SSBD-MBI to 0.93 (large; *improvement index* = +32.5) on the SSiS-PR. In addition, Seeley et al. (2018) reported on measures indicative of comorbid internalizing symptomatology (not shown in table). For example, ESs of 0.23 (small) and 0.42 (medium) were obtained for the parent- and teacher-reported SSiS-internalizing subscale, respectively.

Randomized Controlled Trial 5. Preschool FSN Efficacy Validation Study

Feil et al. (2021) conducted an efficacy validation with the FSN intervention in preschool settings. Participating preschools spanned programs in Illinois, Indiana, Kentucky, and Oregon. During screening, teachers also completed the ABS (Walker, Severson, & Feil, 2014) from the revised SSBD. One hundred sixty teacher–student–parent triads from 50 Head Start and preschool programs participated. Randomization to condition occurred prior to baseline assessment and was at the program level. Students averaged 4.1 (*SD* = 0.3) years of age and were majority male (67%). Student participants comprised a diverse sample and were predominantly White (48%), Black (36%), and Hispanic (16%).

Table 2 contains ESs across prosocial and PB domains ranging from 0.34 (small; *improvement index* = +13.3) on the parent-reported SSiS-SS and parent-reported SSiS-PB to 0.91 (large; *improvement index* = +31.9) on the teacher-reported SSiS-SS. Within the prosocial domain, *Hedges*' *g* ESs ranged from small (0.34) on the SSiS-PR to large (0.91) on the SSiS-TR. Within the PB domain, ESs ranged from 0.34 (small) on the parent-reported SSiS to 0.63 (medium; *improvement index* = +23.6) on the SSiS-TR. Across the prosocial and PB domains, ESs for teacher-reported measures ranged from 0.59 (medium; *improvement index* = +22.2) on the SSBD-MBI to 0.91 (large) on the SSiS-SS compared with parent-reported measures of 0.34 (small) on the SSIS-SS and SSiS-PB. All differences between intervention and control conditions were significant.

Mean implementation adherence was 0.99 ($SD = 0.02$) for coaches and 0.98 ($SD = 0.04$) for teachers. Implementation quality was slightly higher during the coach phase ($M = 0.93$, $SD = 0.06$) than the teacher phase ($M = 0.84$, $SD = 0.15$). Mean satisfaction ratings were similar for teachers ($M = 4.36$, $SD = 0.61$) and parents ($M = 4.28$, $SD = 0.52$).

Discussion and Implications of Findings

This best-evidence synthesis of the FSN intervention contributes to the available literature by directly comparing studies with similar methodological rigor, permitting conclusions regarding the overall weight of the evidence (Slavin, 1995). As such, this review allows for public-school professionals in pre-K and elementary school settings to determine whether FSN might be appropriate for their students' needs.

The best-evidence synthesis, along with subsample analyses, resulted in small to large ESs, very positive improvement indices, and statistically significant behavioral gains compared with students randomized to control conditions. Results also demonstrate the intervention has been implemented with acceptable fidelity and received positive satisfaction ratings from parents and teachers. These results replicated overall outcomes documented for the original FSN study (Walker et al., 1998); they also provide consistent replication of efficacy and effectiveness outcomes across elementary (Frey et al., in press; Sumi et al., 2013; Walker et al., 2009) and preschool (Feil et al., 2014, 2021) populations. In the original small-scale study (Walker et al., 1998), Hedges' *g* ESs were 1.17 (large) and 0.93 (large) for teacher-reported prosocial and PB, respectively. In addition, the ES for direct observations of AET was 0.97 (large). All differences between intervention and control conditions were statistically significant. Overall, those who adopt FSN can be confident that FSN will have a substantial impact on student PBs and SS, so long as it is implemented with adherence fidelity of at least 75%.

In each of the five RCTs, the impact of the FSN intervention was more robust in the school than in the home setting, as indicated by teacher versus parent reports on the SSiS measures. Specifically, teacher reports of prosocial behavior were in the medium to large range across studies and teacher reports of PB were in the medium range for all of the efficacy studies and in the small range for the effectiveness study. Parent-reported measures of prosocial behavior generated small ESs for all of the studies except Walker et al. (2009), which was a medium ES. Thus, programs that choose to adopt FSN should be aware that its impact is likely to be more robust in the school versus the home setting.

The reduced impact in the home setting, as well as teacher reports regarding the time it took to implement the program, identified through item-level analyses of the satisfaction survey, were the important factors in updating and revising the intervention by the program authors and ACORA Publishing Company staff in 2015. This effort resulted in a revised intervention, called FSN, which was more streamlined, less complex, and more focused on academic skills (Walker, Severson, Feil, et al., 2014). The revised FSN program also had a more effective behavior management component than original. A full description of the revision process is described in detail in Walker et al. (2018).

When FSN was delivered by school personnel (as in the effectiveness study) rather than by research staff (i.e., efficacy studies), its impact was less robust, as expected. Specifically, when implemented by school staff (Sumi et al., 2013), the ESs were all in the small range, except for the teacher-reported ES for prosocial behavior, which was medium. However, all the efficacy studies reported herein, including the one with a higher risk sample (Frey et al., in press), had multiple ESs in the medium and large ranges. Therefore, when FSN is adopted and implemented by endogenous providers (i.e., school personnel in real-world settings), expectations with regard to impact should be more closely aligned with the findings from Sumi et al. (2013).

This best-evidence synthesis also provides some verification that the FSN intervention has had a similar impact on preschoolers and elementary age students. Although the ESs for Sumi et al. (2013) and Frey et al. (in press) are smaller than for the other reported studies, it is important to note Sumi et al. (2013) conducted an effectiveness trial and Frey et al.'s (in press) study contained a higher risk (e.g., tertiary rather than secondary level) sample of students. When examining effects across children in preschool versus elementary settings, the ESs for Walker et al. (2009) are similar to those of Feil et al. (2014) and Feil et al. (2021);

specifically, the majority are medium or large. When examining the measures within the prosocial behavior and PB domains across studies, the SSiS-TR seems consistently more sensitive to change than the SSBD-ABI and SSBD-MBI screening subscales.

The synthesis of the subsample populations suggests the FSN intervention can be expected to have similar effects with students with or at risk of having ASD or ADHD. For example, the ESs for students with ADHD in the work of Seeley et al. (2009) are similar to those in the full sample (Walker et al., 2009); furthermore, the ESs for teacher-reported SS and three of the four measures specific to ADHD symptomatology (SSiS-hyperactive subscale, the SSiS-inattentive subscale) were all in the large range. The other three subsample analyses were from the Feil et al. (2014) study. In these analyses, the ESs for the subsamples were of similar magnitude as the full sample for students with or at risk of developing comorbid ASD or comorbid anxiety disorders. The ESs for the general measures of prosocial and PB, however, as well as the measures of ADHD symptomatology (SSiS-cooperation, self-control, and hyperactive subscales) were all in the large range.

Finally, it is also important to note the ESs for FSN in this best-evidence synthesis are comparable to, if not slightly larger than, those reported in Comer et al.'s (2013) meta-analysis of 36 RCTs on psychosocial interventions for preschoolers with disruptive behavior (mean age of 4.7 years) and Epstein et al.'s (2015) meta-analysis of 28 RCTs for young elementary-age children (mean age of 8.2 years). FSN contains a number of best practice elements currently considered to be of critical importance in interventions for young children at risk of emotional and behavioral disorders (Sutherland et al., 2019; Zaheer et al., 2019). Best practice recommendations are also provided for professionals to use in supporting participating teachers and parents (Mitchell et al., 2019; State et al., 2019).

Limitations

There are some limitations to this best-evidence synthesis worth noting. Despite the relative consistency of the results across studies that met our inclusion criteria, only one (Sumi et al., 2013) was an *independent* replication of FSN; even that study was not completely independent in that one of the intervention co-authors provided the training and ongoing technical assistance. Lack of independent replication has been noted as a major short-coming in current special education research (Cook, 2014). Also important to note is the absence of a follow-up study of the effects of FSN over time. To address this, within-year (e.g., end-of-year) and 1-year follow-up outcomes on the most recent preschool sample (Feil et al., 2021) are currently being analyzed and a multi-year follow-up study with the same sample is currently being conducted, including collection of long-term follow-up data at third grade and archival school records.

Another aspect of this best-evidence synthesis that makes the ESs challenging to compare is the 2015 FSN intervention update. In addition, there are challenges to interpreting the improvement index. Specifically, unlike ESs, improvement indices are less commonly used, and not intuitive. Next, while this best-evidence synthesis effectively compares multiple replication studies to the original RCT, it does not compare the RCT findings to the many FSN evaluations that do not employ a randomized controlled design. A summary of all of the FSN evaluation literature is provided in the work of Walker et al. (2018); we believe that the body of evidence for those studies demonstrates similar outcomes. Finally, social validity data were limited to satisfaction surveys; the data would have been richer and more informative if collected via qualitative methods.

Conclusion

This synthesis of the FSN intervention research shows positive effects across a range of settings and across diverse populations of young children with challenging behaviors. General educators are faced with increasing demands for accessing inclusionary practices for children who cause disruptions in their classrooms. Consequently, empirically supported interventions, such as FSN, that focus on positive behavioral support and social-emotional learning would seem to have a much higher acceptability by professionals and end-users going forward. We hope this synthesis will facilitate achievement of this priority.

Acknowledgments

The authors wish to especially recognize the crucial role that Annemieke Golly, PhD, played in the conduct of these studies. Annemieke was the lead trainer of FS coaches and in providing follow-up trouble-shooting and technical assistance services to them. Without her skills and efforts, none of these studies would have been possible.

Authors' Note

Hill Walker and Edward Feil are three of the authors of the original First Step intervention and the revised First Step Next program.

Declaration of Conflicting Interests

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

Funding

This research has been funded by grants from the Institute for Education Sciences #R324A150179 and #R324A150221, U.S. Department of Education #R324A090237 and #H324C040047 and

The National Institute of Child Health and Development #R01HD055334.

ORCID iD

Andy J. Frey https://orcid.org/0000-0002-7493-5728

Notes

1. The original version was called First Step to Success. In 2015, the program was updated and renamed First Step Next (Walker et al., 2015). In this manuscript, we refer to it as First Step Next throughout for consistency.
2. The 1990 version was used in the work of Walker et al. (2009) and the 2008 version was used in all other studies.
3. Academic engaged time observations were not collected for the two studies in early childhood settings (Feil et al., 2014, 2021).

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Research Center for Children, Youth, & Families, University of Vermont.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

Comer, J. S., Chow, C., Chan, P. T., Cooper-Vince, C., & Wilson, L. A. (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

Conners, C. K. (1999). *Conners attention deficit/hyperactivity disorder diagnostic scale*. Multi-Health Systems.

Cook, B. G. (2014). A call for examining replication and bias in special education research. *Remedial & Special Education*, *35*(4), 233–246. https://doi.org/10.1177/0741932514528995

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*, 917–928. https://doi.org/10.1093/jpepsy/jsp004

Epstein, R. A., Fonnesbeck, C., Potter, S., Rizzone, K. H., & McPheeters, M. (2015). Psychosocial interventions for child disruptive behaviors: A meta-analysis. *Journal of the American Academy of Pediatrics*, *148*(4), 2015–2577. https://doi.org/10.1542/peds.2015-2577

Feil, E. G., Frey, A. J., Walker, H. M., Seeley, J., Golly, A., & Small, J. (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*, 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*(2), 95–106. https://doi.org/10.17988/0198-7429-41.2.95

Feil, E. G., Walker, H. W., Frey, A. J., Seeley, J. R., Small, J. W., Golly, A., Lee, J., Crosby, S., & Forness, S. R. (2021). Efficacy validation of the revised First Step program: A randomized controlled trial. *Exceptional Children*, *87*(2), 183–198. https://doi.org/10.1177/0014402920924848

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., Mościcki, E. K., Schinke, S., Valentine, J. C., & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175.

Frey, A. J., Kuklinski, M., Bills, K., Small, J., Forness, S., Walker, H., Feil, E., & Seeley, J. (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*, *20*, 1219–1232. https://doi.org/10.1007/s11121-019-01035-z

Frey, A. J., Small, J. W., Forness, S., Feil, E. G. A., Seeley, J., & Walker, H. M. (2015). First Step to Success: Applications with preschoolers with autism spectrum disorders. *Education and Training in Autism and Developmental Disabilities*, *50*, 397–407.

Frey, A. J., Small, J. W., Lee, J., Crosby, S., Seeley, J. R., Forness, S., & Walker, H. M. (2019). homeBase: Participation, engagement, alliance, and social validity of a motivational parenting intervention. *Children & Schools*. Advance online publication. http://doi.org/10.1093/cs/cdz016

Frey, A. J., Small, J. W., Seeley, J. R., Walker Feil, E. G. H. M., Lee, J., Cohen Lissman, D., Crosby, S., & Forness, S. R. (in press). First Step Next and homeBase: A comparative efficacy study of children with disruptive behavior requiring intensive support. *Exceptional Children*. https://doi.org/10.1177/00144029211062588

Gadow K. D., & Sprafkin, J. (2000) *Early Childhood Symptom Inventory-4 Screening Manual*. Checkmate Plus.

Gottfredson, D. C., Cook, T. D., & Gardner, F. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*, 893–926. https://doi.org/10.1007/s11121-015-0555-x

Gresham, F. M., & Elliot, D. S. (1990). *The social skills rating system (SSRS)*. American Guidance Services.

Gresham, F. M., Vance, M., Elliott, S., & Cook, C. (2011). Comparability of the social skills rating system to the social skills improvement system: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly*, *26*(1), 27–44. https://doi.org/10.1037/a0022662

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Kaminski, J. W., & Claussen, A. H. (2017). Evidence base update for psychosocial treatments for disruptive behaviors in children. *Journal of Clinical Child & Adolescent Psychology*, *46*(4), 477–499. https://doi.org/10.1080/15374416.2017.1310044

Kuklinski, M., Small, J. W., Frey, A. J., Bills, K., & Forness, S. R. (2021). Comprehensive cost effectiveness of a school- and home-based interventions for elementary school students with disruptive behavior disorders [Manuscript submitted for publication]. https://ies.ed.gov/ncee/wwc/PermissionsAndDisclaimers

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863.

Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research. *Remedial & Special Education*, *37*(4), 205–212. https://doi.org/10.1177/0741932516646083

Mitchell, B. S., Kern, L., & Conroy, M. A. (2019). Supporting students with emotional or behavioral disorders: State of the field. *Behavioral Disorders*, *44*(2), 70–84. https://doi.org/10.1177/0198742918816518

Quinn, M., Mathur, S., & Rutherford, R. (1995). Early identification of antisocial boys: A multi-method approach. *Education and Treatment of Children*, *18*, 272–281. https://www.jstor.org/stable/42899413

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: A Multidisciplinary Research and Practice Journal*, *10*(3), 243–253. https://doi.org/10.1007/s12310-017-9226-3

Seeley, J. R., Small, J. W., Walker, H. M., Feil, E. G., Severson, H. H., Golly, A. M., & Forness, S. R. (2009). Efficacy of the First Step to Success intervention for students with attention-deficit/hyperactivity disorder. *School Mental Health*, *1*(1), 37–48. https://doi.org/10.1007/s12310-008-9003-4

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, *48*(1), 9–18. https://doi.org/10.1016/0895-4356(94)00097-A

State, T. M., Simonsen, B., Hirn, R. G., & Wills, H. (2019). Bridging the research-to-practice gap through effective professional development for teachers working with students with emotional and behavioral disorders. *Behavioral Disorders*, *44*(2), 107–116. https://doi.org/10.1177/0198742918816447

Sullivan, G. M., & Feinn, R. (2012). Using effect size- or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Sumi, W. C., Woodbridge, M. W., Javitz, H. S., Thornton, S. P., Wagner, M., Rouspil, K., & 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

Sutherland, K. S., Conroy, M. A., McLeod, B. D., Kunemund, R., & McKnight, K. (2019). Common practice elements for improving social, emotional, and behavioral outcomes of young elementary school students. *Journal of Emotional & Behavioral Disorders*, *27*(2), 76–85. https://doi.org/10.1177/1063426618784009

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider. *Journal of Counseling & Development*, *80*, 64–71.

U.S. Department of Education. (n.d.). Institute for Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/PermissionsAndDisclaimers

Walker, H. M., Feil, E. G., Frey, A. J., Small, J., Seeley, J. R., 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*, 9–38. https://doi.org/10.1177/106342669800600201

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 & Behavioral Disorders*, *6*(2), 66–81. https://doi.org/10.1177/019874290402900407

Walker, H. M., Seeley, J. R., Small, J., Severson, H. H., Graham, B. A., 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 in 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. H. (1990). *Systematic screening for behavior disorders (SSBD)* [User's guide and technical manual]. Sopris West.

Walker, H. M., Severson, H. H., & Feil, E. G. (1995). *Early Screening Project: A proven child-find process*. Sopris West.

Walker, H. M., Severson, H. H., & Feil, E. G. (2014). *Systematic screening for behavior disorders (SSBD): User's guide and technical manual* (Revised 2nd ed.). Pacific Northwest Publishing.

Walker, H., Severson, H. H., Nicholson, F., Kehle, T., Jenson, W., & Clark, E. (1994). Replication of the systematic screening for behavior disorders procedure for the identification of at-risk children. *Journal of Emotional and Behavioral Disorders*, *2*, 66–77. https://doi.org/10.1177/106342669400200201

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 Press.

Walker, H. M., Stiller, B., Coughlin, C., Golly, A., Sprick, M., & Feil, E. (2015). *First Step Next*. Pacific Northwest Publishing.

What Works Clearinghouse. (2020). *What Works Clearinghouse procedures handbook, version 4.1*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf

Zaheer, I., Maggin, D., McDaniel, S., McIntosh, K., Rodriguez, B. J., & Fogt, J. B. (2019). Implementation of promising practices that support students with emotional and behavioral disorders. *Behavioral Disorders*, *44*(2), 117–128. https://doi.org/10.1177/0198742918821331