Hammill Institute on Disabilities

Remedial and Special Education
2017, Vol. 38(3) 131–144

DOI: 10.1177/0741932516653477
rase.sagepub.com
SAGE

A Meta-Analysis of Educator Training to Improve Implementation of Interventions for Students With Disabilities

Matthew E. Brock, PhD[1] and Erik W. Carter, PhD[2]

Abstract

Teachers and paraprofessionals need effective training to improve their implementation of interventions for students with disabilities. Reviews of the single-case design literature have identified some features associated with effective training for these educators, but the group-design literature has received little attention. This meta-analysis systematically reviews group-design studies testing the efficacy of training to improve implementation of interventions for students with disabilities. The mean effect size of educator training on implementation fidelity was $g = 1.08$, and results from meta-regression analysis suggest training that involves a combination of two specific training strategies (i.e., modeling and performance feedback) was associated with improved implementation fidelity. Increased duration of training was not associated with larger effects. Considered alongside findings from the single-case design literature, these results suggest that *how* educators are trained is a more important consideration than the number of hours they spend in training.

Keywords

meta-analysis, teacher training, paraprofessional training, modeling, performance feedback, coaching

Bridging the research to practice gap has been an enduring challenge in the field of special education. Researchers assert and educators agree that educational practices supported by rigorous scientific evidence can and should be used to improve outcomes for students with disabilities (e.g., Cook & Cook, 2013; Cook, Smith, & Tankersley, 2012). Furthermore, the law requires that teachers implement evidence-based practices (No Child Left Behind [NCLB] Act, 2002). However, translating research into practice remains a substantial challenge. For example, educators report a lack of confidence in implementing evidence-based practices (e.g., Brock, Huber, Carter, Juarez, & Warren, 2014), and researchers continue to express concerns about practitioner implementation (e.g., Cook & Cook). The research evidence supporting the efficacy of a given educational practice consists of experimental studies designed to determine whether implementation of a specific intervention protocol is effective for improving specific student outcomes. Findings from these studies can only be generalized to situations where practitioners implement the intervention with fidelity (i.e., as described in the research study). When practitioners do not follow this protocol with fidelity, the intervention is no longer supported by research evidence (Cook & Odom, 2013). Effective training is needed that enables pre-service teachers, in-service teachers, and paraprofessionals to better implement evidence-based practices to improve outcomes for students with disabilities.

One frequently highlighted element of effective training for these educators is performance feedback. Performance feedback involves observing an educator to collect data on implementation of a teaching strategy, and then sharing data with him or her to improve future performance. Performance feedback was identified as an effective strategy across four recent narrative reviews and meta-analyses of the single-case design literature. Fallon, Collier-Meek, Maggin, Sanetti, and Johnson (2015) identified performance feedback as an evidence-based practice for improving educator implementation fidelity. In a comprehensive review of the single-case design literature, they identified 126 single-case design studies. After applying *What Works Clearinghouse* design and evidence standards, they determined that 102 of these studies provide moderate or strong evidence to support the efficacy of performance feedback.

[1]Crane Center for Early Childhood Research and Policy, The Ohio State University, Columbus, USA
[2]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Matthew E. Brock, Department of Educational Studies, Crane Center for Early Childhood Research and Policy, The Ohio State University, 334 PAES Building, 305 West 17th Avenue, Columbus, OH 43210, USA.
Email: brock.184@osu.edu

Action Editor: Daniel Maggin

Noell and colleagues (2014) conducted a meta-analysis of 29 single-case design studies that analyzed intervention implementation in schools. Using multi-level linear modeling, they computed effect sizes in the form of intraclass correlation coefficients (ICC) that represent the proportion of variance between the baseline and treatment conditions. They found that performance feedback alone had the largest effect of any treatment condition; although self-monitoring and a combination of performance feedback, rehearsal, and a meeting cancelation contingency were also associated with statistically significant effects. In contrast, the effect of follow-up meetings without performance feedback was nonsignificant.

Solomon, Klein, and Politylo (2012) conducted a meta-analysis of 36 single-case design studies in which performance feedback was used to improve teacher implementation fidelity. They calculated mean trend (Allison & Gorman, 1993) and the improvement rate difference (IRD; Parker, Vannest, & Brown, 2009), and then used inferential statistics to test the overall efficacy of performance feedback and the degree to which efficacy was moderated by teacher role (i.e., general educator or special educator), the target of the teacher-delivered intervention (i.e., student behavior or academic skills), or the immediacy of feedback. The overall effect of performance feedback was statistically significant, and it tended to be more effective in studies with general educators and studies focused on academic interventions.

In their review of the paraprofessional training literature, Brock and Carter (2013) highlighted modeling and performance feedback as important features of training for paraprofessionals. They identified 13 single-case design studies in which paraprofessional training resulted in accurate implementation of interventions for students with intellectual and developmental disabilities. Effective training most often included a combination of performance feedback and modeling. Modeling included either a video or in-person demonstration of implementation steps. The authors suggest these two strategies are useful in conjunction, because modeling clearly communicates how to implement an intervention and then performance feedback reinforces what practitioners are doing well and helps them to correct their mistakes.

Collectively, these four reviews synthesize strong evidence from the single-case design literature that supports the efficacy of performance feedback. Although these existing reviews make a strong contribution toward better understanding what makes educator training most effective, they have two critical limitations. First, these reviews do not address group-design studies (i.e., randomized controlled trials and quasi-experimental studies). This may be especially problematic given that single-case design studies—due to the nature of the design—almost exclusively utilize a one-to-one coaching model. Although reviewing the single-case design literature may establish a strong understanding of what is effective in the context of coaching, it does little to examine other training formats, or to compare differences between training formats. Furthermore, it is unlikely that coaching alone is a feasible means to affect all educators. Indeed, some school systems choose not to use coaching because of the large amount of resources required to change the behavior of a single educator (Russo, 2004). It would seem that evidence from group-design studies—that more naturally lend themselves to testing educator training in a group format—should be considered alongside the evidence from single-case design studies. Second, researchers have not come to consensus around whether and how to conduct quantitative meta-analyses of single-case design studies (Maggin & Chafouleas, 2013) and researchers have demonstrated that some commonly used methods are problematic (Wolery, Busick, Reichow, & Barton, 2010). In contrast, meta-analysis of group-design studies is more established, and there is much broader consensus among researchers regarding which analytic techniques are appropriate (Borenstein, Hedges, Higgins, & Rothstein, 2011).

In the present systematic review and meta-analysis of group-design studies testing the efficacy of educator training on implementation fidelity, we aim to address these limitations through three research questions:

Research Question 1: What is the summary effect size of training compared with no training or business-as-usual training on implementation fidelity among educators serving students with disabilities?
Research Question 2: What findings about professional development from the single-case design literature are confirmed in the group-design literature?
Research Question 3: Is increased duration of training associated with larger increases in implementation fidelity?

Specifically in question 2, we sought to determine whether performance feedback (e.g., Fallon et al., 2015), a combination of modeling and performance feedback (e.g., Brock & Carter, 2013), and/or a one-to-one coaching format (Kretlow & Bartholomew, 2010) is confirmed as efficacious in the group-design literature.

Method

Study Eligibility Criteria

To be included in this meta-analysis, we required studies to meet the following criteria. First, study participants must have included teachers, pre-service teachers, or paraprofessionals who provided school-based services to students with diagnosed disabilities in the United States.

Together, these three groups represent the individuals who provide—or are being trained to provide—the vast majority of school-based instruction and support to students with disabilities. Disabilities were defined as any disability category listed in the Individuals With Disabilities Education Improvement Act (2004). Second, the independent variable must have been educator training, defined as any training provided to teachers or paraprofessionals designed to change or improve implementation. Third, studies must have included measures of implementation fidelity as an outcome variable, which must have been measured through observation in the context of teaching or providing support to students with diagnosed disabilities. Studies were excluded if they assessed implementation through educator report or record review (e.g., Fuchs, Fuchs, & Hamlett, 1989), given that these methods are less reliable indicators of how educators actually intervened with students. Fourth, studies must have compared the effects of training for an experimental group to the effects of no treatment or business-as-usual-training for a comparison group. For pre-service teachers, business-as-usual was defined as lecture-based college course work and/or general feedback from fieldwork supervisors. For in-service practitioners, business-as-usual was defined as a traditional stand-alone training workshop. Fifth, only randomized controlled trials and quasi-experimental studies (i.e., nonrandom assignment to experimental and control groups) were eligible for inclusion. Results of both groups must have been reported in terms of means and standard deviations. Finally, studies must have been published or written after 1975, when public schools were first required to serve all students with disabilities by the Education for All Handicapped Children Act (1975). We completed search procedures in October 2015.

Search Strategy

We used multiple search strategies to identify all studies meeting the above criteria. These strategies included searching of electronic databases, checking reference lists from studies meeting eligibility criteria, citation searches of studies meeting eligibility criteria, and hand searching all issues of *Exceptional Children, The Journal of Special Education*, and *Teacher Education and Special Education* between 1975 and 2015. We searched five electronic databases, including PsycInfo, Education Resources Information Center (ERIC), ProQuest Dissertation and Theses Database (PQDT), Social Services Abstracts, and Sociological Abstracts. We customized search strings to take advantage of the subject terms provided within each database. Customized search strings and the number of hits from each database are reported in the online supplement. The first author screened all search hits. The second author screened a random 10% sample of all search hits. Point-by-point agreement was 99.1% for screening articles into full-text review, and 100% for identifying articles that met eligibility criteria. We identified 12 studies through systematic searching of electronic databases, checking reference lists from eligible studies, citation searches of eligible studies, and hand searching three prominent journals. Figure 1 displays a flow diagram outlining all search procedures, including the number of studies identified by each strategy.

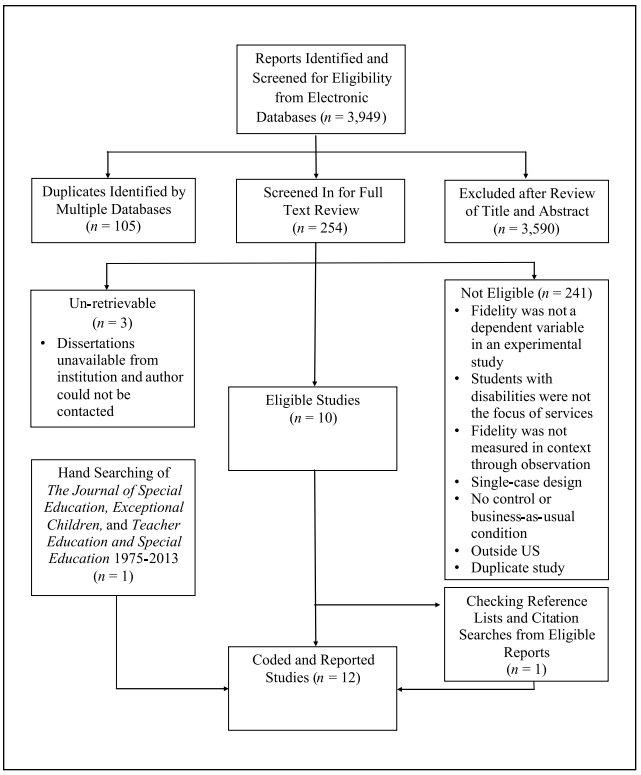

Figure 1. Flowchart of study identification procedures.

Data Collection and Variables

We coded five categories of variables: (a) study design and threats to internal validity, (b) participant and setting characteristics, (c) description of the independent variable, (d) training strategies, and (e) description of the dependent variable. We selected variables that would provide context for understanding the nature of the studies and enable moderator analysis. To ensure reliability, these variables were coded independently by the two authors. Initial agreement was 95.5%, and all disagreements were resolved through consensus. A complete coding manual for this meta-analysis is available as an online supplement.

Study design and threats to internal validity. We classified study designs as randomized controlled trials if the authors reported randomly assigning participants to experimental and comparison groups. Studies in which participants were

nonrandomly assigned to experimental and comparison groups were classified as quasi-experimental.

We used the Cochrane Collaboration's tool for assessing risk of bias (Higgins et al., 2011) to assess threats to internal validity. This tool has six domains, including sequence generation (i.e., use of nonrandom assignment), allocation concealment (i.e., participants or investigators could foresee assignment), blinding (i.e., participants or outcome assessors were not blind to the treatment group, and this knowledge could have affected outcome measures), incomplete outcome data (i.e., missing data and/or participant attrition), selective outcome reporting (i.e., evidence that the authors pre-specified a number of outcomes, but only reported on those with positive results), and other potential threats to validity not captured in the previous five domains. Indicators and guidelines are provided for rating a study on each domain as low risk of bias, high risk of bias, and unclear risk of bias (i.e., inadequate reporting to determine the risk of bias on a given domain). After all individual studies are rated on all domains, an overall judgment can be made across studies. Across-study ratings include low risk of bias (i.e., most studies have low risk of bias), high risk of bias (the proportion of studies at high risk of bias could affect interpretation of results), and unclear risk of bias (most studies have low or unclear risk of bias).

Participant and setting characteristics. We coded practitioner participants by role (i.e., special education teacher, general education teacher, paraprofessional, pre-service special education teacher). We coded student participants by reported disability and grade level (i.e., preschool, elementary, middle, high school). We coded whether fidelity was measured in a general education or special education classroom based on author description. If authors described a setting with only students with disabilities who were served by special educators or special education paraprofessionals, we concluded that the students were served in a special education setting. We coded the urbanicity of schools as rural, suburban, or urban.

Independent variable. We recorded the authors' description of the training (independent variable), the number of hours training occurred, and the number of days between the beginning and end of the training. In five cases, the authors provided a narrative description of the training sessions, but did not indicate specific duration. In these cases, we contacted the first author of each study by email. Two of the authors responded and were able to provide this information. A third author indicated that he did not recall these details and that study records had been destroyed, a fourth author did not respond, and current contact information could not be identified for the fifth author. We coded training format as group training, coaching (i.e., a one-to-one format), or a combination of group and coaching.

Educator training strategies. We coded the presence or absence of training strategies (i.e., independent variable) based on strategies identified in a previous review (Brock & Carter, 2013), adding new categories as necessary based on author description. These strategies included practice description (i.e., verbal or written description of the practice being taught to the practitioner), fidelity checklist (i.e., sharing a printed list of intervention steps), modeling (i.e., in-person or video representation of intervention implementation), performance feedback (i.e., collecting implementation fidelity data of a practitioner implementing the intervention with student[s] with disabilities, and subsequently sharing this data with the practitioner), planning (i.e., directing practitioners to create intervention plans tailored to specific students), question-and-answer session (i.e., providing an opportunity for practitioners to ask questions about implementation of intervention), rationale (i.e., providing an oral or written rationale for why the intervention is important), reading material (i.e., directing practitioners to read written material), role-play (i.e., directing practitioners to practice implementing the intervention with other adults), self-monitoring (i.e., directing practitioners to collect implementation fidelity data about their own performance when implementing the intervention with students with disabilities), and study groups (i.e., directing practitioners to meet in groups to discuss implementation).

Dependent variables. We recorded descriptions of implementation fidelity measures and coded the authors' approach to measuring these variables (i.e., frequency of discrete behavior, percentage of steps on multi-step implementation checklist, duration of behavior, yes/no checklist, rating scale indicating quality of implementation). We recorded the type of student outcome targeted by the intervention (i.e., academic, social, problem behavior, play), whether student-level outcomes were measured, and whether the authors reported a statistically significant difference between groups on a student-level measure.

Effect Size Measures

Calculations. We computed all effect sizes as a standardized mean difference by dividing the difference between the post-treatment experimental and comparison group means by a pooled standard deviation. We then multiplied the standardized mean difference by a correction factor to obtain Hedge's *g*, which accounts for bias due to small sample size (Borenstein et al., 2011). These calculations are expressed in the following formula:

$$g = \frac{\bar{X}_1 + \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \times \left(1 - \frac{3}{4\mathrm{df} - 1}\right).$$

Multiple outcomes. Eight of the 12 studies reported multiple implementation fidelity outcomes. In these cases, we calculated effect sizes for each implementation fidelity outcome targeted for increase (excluding behaviors targeted for decrease) and computed one synthetic effect size for each study from the individual effect sizes. The synthetic effect size was an average of individual effect sizes that accounted for covariance. Computing a synthetic effect size was necessary to obtain a single indication of effect magnitude, because including all outcomes individually would incorrectly assume independence of observations (Borenstein et al., 2011). Because only one study reported an ICC for the variables of interest, we estimated the ICC to be .70 because (a) measures of fidelity related to the same training should be moderately to highly correlated, (b) the only study that included any correlation coefficients among fidelity measures reported correlations between .50 and .70 (Weiner, 2010), and (c) selecting the highest correlation in a possible range of correlations provides the most conservative estimate of standard error with the widest confidence intervals (Borenstein et al., 2011). We computed synthetic effect sizes by adding the sum of the outcome variances and the sum of the products of the outcome variances and correlation coefficients, and multiplying by the inverse number of groups squared. These calculations are expressed in the following formula from Borenstein et al. (2011):

$$\left(\frac{1}{m}\right)^2\left(\sum_{i=1}^{m} V_i + \sum_{i\neq j}\left(r_{ij}\sqrt{V_i}\sqrt{V_j}\right)\right).$$

Special cases. In three cases, we used slightly different procedures to compute effect sizes. Collier (2008) reported means for two separate comparison groups receiving the same treatment. Using a random number generator, we randomly selected one of the comparison groups to compute an effect size. Peterson and McConnell (1996) did not report mean and standard deviations across treatment groups with teachers as the unit of analysis but did report the results from ANOVA that compared these means. We converted the F value to Cohen's d and then Hedge's g (Borenstein, 2009).

Analytic Strategies

We used three primary strategies to analyze effect sizes across studies: (a) calculation of a random-effects mean effect size, (b) meta-regression to test the influence of moderators associated with a priori hypotheses, and (c) sensitivity analysis to examine the potential influences of publication bias and inclusion of studies identified with high risks of bias using the tool by the Cochrane Collaboration (Higgins et al., 2011).

Calculating random-effects mean effect size. To address the first research question, we used a random-effects model to calculate a mean effect size across all studies. Unlike a fixed-effects model, this model is not constrained by the assumption that all unexplained variance is a result of sampling error. Instead, a random-effects model calculates both within-study and between-study variance to estimate a distribution of true effects (Borenstein et al., 2011).

Moderator analysis. To address the second research question, we used meta-regression. We built four separate regression models, each with study-level effect size as the dependent variable. For independent variables, we included performance feedback in the first model, presence of a combination of coaching and performance feedback in the second model, coaching (i.e., presence of one-to-one training) in the third model, and duration of training in the fourth model. The first three models included all 12 studies from this meta-analysis; the fourth model only included the nine studies for which data were available (see "Independent variable" section above).

Sensitivity analysis. We used sensitivity analysis to examine (a) the possibility of publication bias and (b) the influence of studies with notable threats to internal validity on the mean effect size. To examine publication bias, we used three different approaches: (a) a funnel plot to visually analyze the relationship between effects size and precision of measurement (i.e., standard error), (b) an Egger test (Egger, Smith, Schneider, & Minder, 1997) to statistically test whether standard error systematically predicts effect size, and (c) a trim-and-fill analysis (Duval & Tweedie, 2000) to estimate a theoretically unbiased mean effect size. To examine the influence of studies with notable threats to internal validity on the mean effect size, we computed a random-effects mean effect size after excluding studies identified as having high risks of bias using the tool from the Cochrane Collaboration.

Results

Study Design and Threats to Internal Validity

Threats to internal validity were assessed using the Cochrane Collaboration's tool for assessing risk of bias (Higgins et al., 2011; see Table 1). Of the 12 studies, eight randomized participants to experimental conditions and four did not. Of the eight studies that randomized participants, only three specified the method of randomization in enough detail to determine that the allocation sequence was adequately generated (i.e., assignment was truly random) and that allocation was adequately concealed (i.e., participants and investigators could not foresee assignment). Only three studies described steps taken to blind observers to treatment

Table 1. Analysis of Studies Using Cochrane Collaboration's Tool for Assessing Risk of Bias.

Study	Sequence generation[a]	Allocation concealment[b]	Blinding[c]	Incomplete outcome data[d]	Selective outcome reporting[e]	Other sources of bias[f]
Ascione and Borg (1980)	High risk (no random assignment)	High risk (teacher volunteers)	Unclear (not reported)	Unclear (attrition not reported)	Low risk	None identified
Ascione and Borg (1983)	Low risk (determined by lot)	Low risk (randomized after consent)	Unclear (not reported)	High risk (18% attrition)	Low risk	None identified
Brock and Carter (2015)	Low risk (random number generator)	Low risk (randomized after consent)	Unclear (not reported)	Low risk (7% attrition)	Low risk	None identified
Collier (2008)	High risk (no random assignment)	High risk (assigned to condition based on school)	Unclear (not reported)	Unclear (attrition not reported)	Low risk	None identified
Dixon (1983)	High risk (no random assignment)	High risk (assigned to condition based on availability on weekends)	Unclear (not reported)	Low risk (no attrition)	Low risk	None identified
Fink (1980)	Unclear (random assignment method not described)	Unclear (random assignment method not described)	Low risk (observers blinded)	High risk (36% attrition)	Low risk	None identified
Hindman and Polsgrove (1988)	Unclear (random assignment method not described)	Unclear (random assignment method not described)	Unclear (not reported)	Low risk (7% attrition)	Low risk	None identified
Lawton and Kasari (2012)	Low risk (random number generator)	Low risk (randomized after entry assessments)	Low risk (observers blinded)	Low risk (no attrition)	Low risk	Randomization at classroom level
Peterson and McConnell (1996)	Unclear (random assignment method not described)	Unclear (random assignment method not described)	Unclear (not reported)	Low risk (no attrition)	Low risk	None identified
Sutherland and Wehby (2003)	Unclear (random assignment method not described)	Unclear (random assignment method not described)	Low risk (observers blinded)	Low risk (no attrition)	Low risk	None identified
Weiner (2010)	Unclear (random assignment method not described)	Unclear (random assignment method not described)	Unclear (not reported)	Low risk (no attrition)	Low risk	None identified
Whitten (1986)	High risk (no random assignment)	High risk (assigned by school)	None reported	Unclear (attrition not reported)	Low risk	None identified

Note. Studies coded based on indicators listed in Cochrane Collaboration's tool for assessing risk of bias (Higgins et al., 2011).
[a]Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups. [b]Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of or during enrollment. [c]Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. [d]Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. [e]State any important concerns about bias not addressed in the other domains in the tool.

condition, and seven studies reported recruitment in sufficient detail to determine that attrition was not a significant threat to internal validity. There was no evidence of selective outcome reporting. One study (i.e., Lawton & Kasari, 2012) randomized participants at the classroom level, introducing a mismatch between level of assignment and level of analysis. We only identified five studies that did not have a high risk of bias in at least one category (i.e., Brock & Carter, 2015; Hindman & Polsgrove, 1988; Peterson & McConnell, 1996; Sutherland & Wehby, 2001; Weiner, 2010). However, each of these five studies received an unclear rating in at least one category because of insufficient description to judge risk of

Table 2. Participant and Setting Characteristics From Included Studies.

Study	Practitioners	Student disabilities	Setting	Grade level	Urbanity of school
Ascione and Borg (1980)	18 general educators	EBD, LD	General education	Elementary	Urban
Ascione and Borg (1983)	33 general educators	EBD, ID, LD	General education	Elementary	Urban
Brock and Carter (2015)	25 paraprofessionals	Not reported	Combination of settings	Elementary, middle, and high	Suburban
Collier (2008)	18 special educators and paraprofessionals	LD	Special education	High	Rural
Dixon (1983)	20 special educators	LD	Special education	Elementary	Urban
Fink (1980)	64 general educators	LD	General education	Elementary	Suburban
Hindman and Polsgrove (1988)	25 pre-service special educators	Not reported	Not reported	Not reported	Not reported
Lawton and Kasari (2012)	16 special educators and paraprofessionals	ASD	Combination of settings	Preschool	Urban
Peterson and McConnell (1996)	16 special educators	Not reported	Combination of settings	Preschool	Urban
Sutherland and Wehby (2003)	20 special educators	EBD, ID, LD	Special education	Elementary and middle	Urban
Weiner (2010)	31 paraprofessionals	ASD, EBD, LD, OHI, OI, SL, TBI, VI	General education	Elementary	Not reported
Whitten (1986)	12 special educators	HI, ID, OI, VI	Epecial education	Elementary, middle, and high	Not reported

Note. EBD = emotional/behavioral disturbance; LD = learning disability; ID = intellectual disability; ASD = autism spectrum disorder; OHI = other health impairment; OI = orthopedic impairment; SL = speech and language; TBI = traumatic brain injury; VI = visual impairment; HI = hearing impairment.

bias. Based on the Cochrane Collaborations proposed approach for summary assessment, a high risk of bias exists across studies (i.e., the majority of studies have a high risk of bias for one or more domains).

Participant and Setting Characteristics

Number of practitioner participants per study ranged from 12 to 64 and included special educators, general educators, paraprofessionals, and pre-service special educators. Student disabilities ranged from student participants with high-incidence disabilities (i.e., learning disabilities, emotional disturbance) to low-incidence disabilities (i.e., autism spectrum disorder, intellectual disability, visual and/or hearing impairments). Study-level coding of participant and setting characteristics is reported in Table 2.

Independent Variable

Duration of training ranged from 3.75 and 45.00 hr across studies. In most (n = 6) studies, training opportunities were implemented in a combination of group and one-to-one formats; three occurred only in a group format, and two occurred only in a one-to-one format. Study-level descriptions of the independent variable are reported in Table 3.

Educator Training Strategies

Across studies, training included a variety of strategies. Nine studies included performance feedback, and seven studies included a combination of modeling and performance feedback; both were hypothesized moderators of effect size. Study-level coding of training strategies is reported in Table 4.

Dependent Variable

Study-level descriptions of dependent variables—including both the measurement strategy (i.e., how fidelity was measured) and the focus of training (i.e., what was measured)—are reported in Table 3.

Measurement strategy. Researchers used a range of strategies to measure implementation fidelity, including frequency of discrete practitioner behaviors (n = 8 studies), percentage of steps implemented correctly from a

Table 3. Study Designs and Descriptions of Independent and Dependent Variables.

Study	Design	Training description[a]	Hours of training	Training format	Dependent variable(s)	Student outcome(s)
Ascione and Borg (1980)	QED	Training course	21.0	Group	Frequency of nine positive teaching behaviors associated with student self-concept[b]	Social[c]
Ascione and Borg (1983)	RCT	Training course	NR	Group and one-to-one	Frequency of eight positive teaching behaviors associated with student self-concept[b]	Social[d]
Brock and Carter (2015)	RCT	Workshop, video-modeling, and Coaching	3.75	Group and One-to-one	Percentage of constant time delay steps implemented correctly	Academic
Collier (2008)	QED	Coaching	NR	One-to-one	Rating scale indicating quality of implementation of literacy direct instruction[b]	Academic[d]
Dixon (1983)	QED	Workshop series	45.0	Group	Frequency of asking two types of higher order questions[b]	Academic[c]
Fink (1980)	RCT	Workshop series	18.0	Group	Frequency of implementing individualized instruction	Academic
Hindman and Polsgrove (1988)	RCT	Workshop and feedback	7.5	Group and one-to-one	Duration of active instruction (modeling, physically moving around room, visual monitoring)	Engagement[c]
Lawton and Kasari (2012)	RCT	Workshop and coaching	6.0	Group and one-to-one	Frequency of behaviors associated with JASPER in a live observation and a taped observation[b]	Social and play[d]
Peterson and McConnell (1996)	RCT	Workshop and coaching	NR	Group and one-to-one	Rating scale indicating quality of implementation of social interaction skills packages	social[c]
Sutherland and Wehby (2003)	RCT	Feedback and self-monitoring	4.0	One-to-one	Frequency of teacher praise and student opportunities to respond[b]	Academic[d]
Weiner (2010)	RCT	Workshop series and coaching	16.0	Group and one-to-one	Frequency of prompting and descriptive praise[b]	Academic and behavior modification[c]
Whitten (1986)	QED	Workshop series	21.0	Group	Frequency of 15 different prompting and direct support behaviors[b]	Academic and social[c]

Note. QED = quasi-experimental design (i.e., no randomization to experimental conditions); RCT = randomized controlled trial; NR = not reported; JASPER = Joint Attention and Symbolic Play/Engagement and Regulation.
[a]Based on author description and word choice. [b]To calculate an overall effect size, multiple dependent variables were combined into a single synthetic effect size. [c]Reported student outcome measure. [d]Reported statistically significant difference between groups on student outcome measure.

multi-step implementation checklist (n = 1 study), duration of targeted practitioner behavior (n = 1 study), or a rating scale indicating quality of implementation (n = 2 studies). Most studies involved teaching practitioners to implement interventions targeting academic student outcomes (n = 6 studies), while other interventions targeted social outcomes (n = 3 studies), student engagement (n = 1 study), or a combination of outcomes (n = 2 studies). Ten studies measured student outcomes, and four of these studies detected a statistically significant difference between student treatment groups. Study-level descriptions of dependent variables are reported in Table 3.

Focus of training. The focus of training varied widely across studies. Studies focused on (a) a group of practitioner behaviors designed to target a specific student outcome (i.e., Ascione & Borg, 1980, 1983; Lawton & Kasari, 2012; Peterson & McConnell, 1996; Sutherland & Wehby, 2003; Weiner, 2010), (b) a group of practitioner behaviors unified by a theory (Collier, 2008; Dixon, 1983; Fink, 1980; Hindman

Table 4. Professional Development Strategies From Included Studies.

Study	Description	Fidelity checklist	Modeling	Performance feedback	Planning	Q&A session	Rationale	Reading material	Role play	Self-monitoring	Study groups
Ascione and Borg (1980)	X		X	X		X	X	X		X	
Ascione and Borg (1983)	X		X	X		X	X	X		X	
Brock and Carter (2015)	X	X	X	X	X		X		X		
Collier (2008)	X		X	X							
Dixon (1983)	X		X		X		X	X	X		
Fink (1980)	X		X			X	X				
Hindman and Polsgrove (1988)	X			X							
Lawton and Kasari (2012)	X		X	X		X	X				
Peterson and McConnell (1996)	X			X	X	X	X	X			
Sutherland and Wehby (2003)	X		X	X	X		X			X	
Weiner (2010)	X		X	X		X	X		X		
Whitten (1986)	X						X				

Note. Description = verbal or written description of the intervention; fidelity checklist = sharing printed list of intervention steps; modeling = in-person or video representation of intervention implementation; performance feedback = trainer collects implementation fidelity data of practitioner implementing the intervention with student(s) with disabilities, and subsequently sharing these data with the practitioner; planning = trainer directs practitioner to create intervention plan tailored to specific student(s); question-and-answer session = trainer provides opportunity for practitioners to ask questions about implementation of intervention; rationale = trainer provides oral or written rational why they intervention is important; reading material = trainer directs practitioners to read written reading material; role-play = trainer directs practitioners to practice implementing the intervention with other adults; self-monitoring = trainer directs practitioners to collect implementation fidelity data about their own performance when implementing the intervention with student(s) with disabilities; and study groups = practitioners meet groups at schedule times to discuss implementation of the intervention.

& Polsgrove, 1988), (c) loosely related practitioner behaviors thought to be associated with high-quality instruction (Whitten, 1986), or (d) a specific, well-developed behavioral intervention strategy (Brock & Carter, 2015).

Overall Mean Effect

Study-level effect sizes across individual studies ranged from $g = .06$ to 2.57, with a mean effect size of $g = 1.08$, 95% CI = [0.71, 1.46]. Although magnitude of effect size is best interpreted in the context of the intervention being evaluated and the outcome variables (see "Discussion"; Hill, Bloom, Black, & Lipsey, 2007), this mean effect size is large according to commonly used benchmarks (e.g., Cohen, 1988). Figure 2 shows study-level effect sizes and CIs, as well as the overall mean effect size and CI.

Moderator Analysis

Moderator analysis involved using meta-regression to consider whether certain features of training accounted for the variability in the magnitude of their impact. Before running meta-regression models with predictor variables, we ran a null model without predictors. Estimates from the null model suggested a wide distribution of effect sizes across studies ($\tau^2 = .26$), and that the majority of this variance ($I^2 = 55.6\%$) can be attributed to true heterogeneity among studies. The I^2 statistic was computed as $Q\text{-}df/Q$ ($Q = 24.77$; $df = 11$). In other words, these statistics suggest substantial differences in the magnitude of effects across studies and that these differences could potentially be explained by a moderating variable.

All analyses were conducted on 12 studies except for the model including duration of training, for which data were only available for nine studies. Performance feedback alone was not a significant predictor of effect size, $\beta = .68$, $t(10) = 1.62$, $p = .14$, although it did explain random variance in the model ($R^2 = 19.53\%$). Presence of a combination of modeling and performance feedback was the strongest single predictor. This combination was statistically significant, $\beta = .77$, $t(10) = 2.24$, $p = .04$, and it explained a substantial proportion of the random variance in the model ($R^2 = 42.03\%$). Use of a one-to-one training format (i.e., coaching) alone was not a statistically significant predictor of effect size, $\beta = .67$, $t(10) = 1.74$, $p = .11$, although it did explain random variance in the model ($R^2 = 27.12\%$). Duration of training

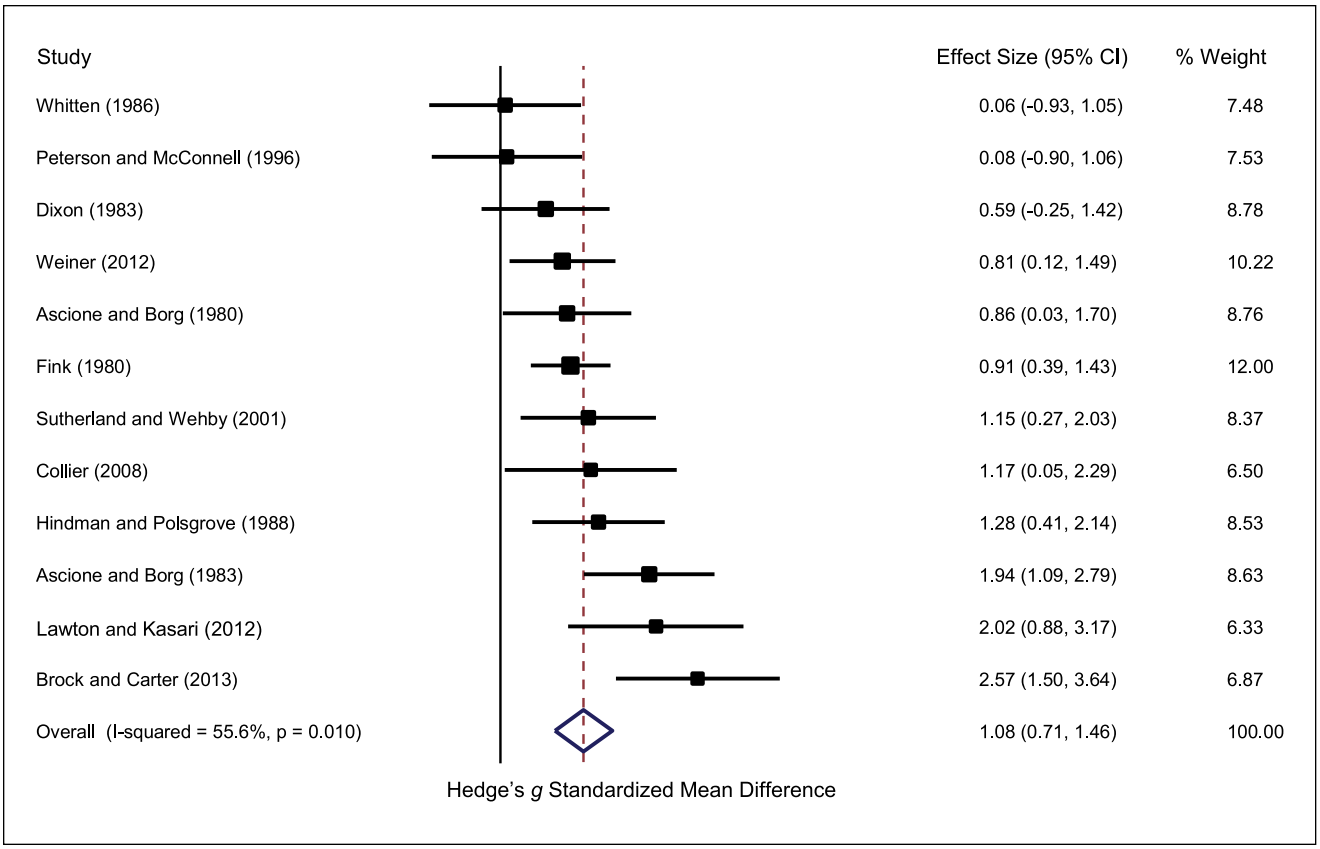

Figure 2. Forest plot displaying study-level effect sizes and the random-effects mean effect size across all eligible studies.
Note. 95% confidence intervals are noted for all effect sizes. Size of square indicates the relative weight of a study-level effect size when computing the overall effect size. Weights are calculated as the inverse of study variance, which includes an estimates of within-study and between-study (τ^2) variance. CI = confidence interval.

(calculated for the subset of nine studies for which this information was available) was not a statistically significant predictor of effect size, $\beta = -.02$, $t(7) = -1.45$, $p = .19$, although it did explain random variance in the model ($R^2 = 26.48\%$). Notably, the coefficient was negative, indicating a (statistically nonsignificant) association between shorter duration of training and increased effect size.

Sensitivity Analysis

Publication bias. The plot of effect size as a function of standard error (see Figure 3) is neither clearly symmetrical nor asymmetrical. This pattern does not provide evidence—but does not rule out the possibility—that smaller studies with smaller or null effects may not have reached publication or dissemination. Results from an Egger's suggest the positive association between effect size and standard error is not greater than what would be expected by chance, $\beta = 1.16$, $t(10) = 0.86$, $p = .41$. No studies were removed or added through trim-and-fill analysis, so this analysis did not provide any evidence of publication bias.

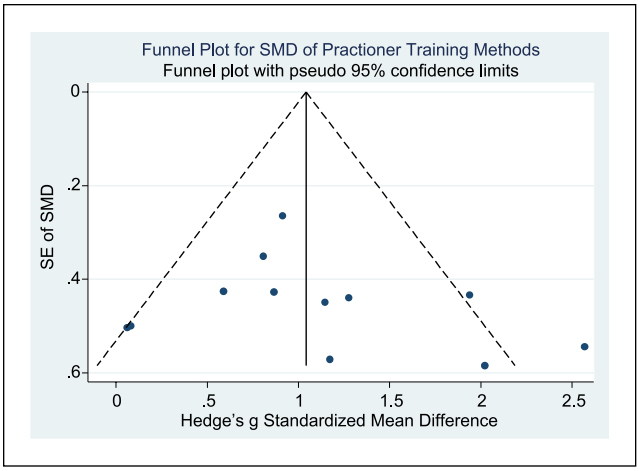

Figure 3. Funnel plot displaying effect size as a function of standard error to assess small-study bias.
Note. Solid line indicates mean effect size, and dotted lines indicate pseudo 95% confidence intervals. Asymmetrical funnel plots suggest small-study bias, which may be explained by publication bias where smaller studies with small or null effects are unlikely to be published or disseminated. SE = standard error; SMD = standardized mean difference.

Influence of studies with risk of bias. The random-effects mean effect size of only the five studies without known risks of bias (as identified through the Cochrane collaboration tool; see Table 1) was actually greater than the mean effect size that included all studies (g = 1.14; 95% CI = [0.45, 1.83]). This suggests that inclusion of studies with risk of bias deflates the mean effect size.

Discussion

Through systematic review of the research literature, we identified 12 randomized controlled trials or quasi-experimental studies testing the efficacy of training to improve fidelity of interventions implemented by educators (i.e., in-service teachers, pre-service teachers, and paraprofessionals) for students with disabilities. Across studies, this body of literature had a high risk of bias according to the Cochrane Collaboration's tool for assessing risk of bias. The mean effect size was both large and statistically significant, and interventions including a combination of modeling and performance feedback tended to have larger effect sizes. Although performance feedback and a coaching format each explained variance individually, these effects were not statistically significant—perhaps due to the very small number of studies that were included in this meta-analysis. For the nine studies with known duration of training, longer duration of training was not related to larger effects. These findings extend the literature in several ways.

First, this meta-analysis highlights the scarcity of high-quality group-design studies testing the impact of practitioner training on implementation fidelity. We only identified 12 studies meeting inclusion criteria, and seven of these studies had a high risk of bias on one more domain of the Cochrane Collaboration's tool for assessing risk of bias. One might expect these methodological limitations to inflate the mean effect size. For example, one might expect that allowing participants to volunteer for the experimental group would increase the likelihood that the experimental group would be composed of highly motivated participants. Similarly, a high rate of attrition in the experimental group might leave the participants most likely to perform well. However, our sensitivity analysis does not support this hypothesis. Collectively, the seven studies with a high risk of bias actually had a lower mean effect size than the other five studies. Given this result, it is unclear whether and how methodological weaknesses affected findings. Nonetheless, we emphasize that the reader should interpret all findings in light of the small number of studies in this meta-analysis, and the large proportion of studies that involve a risk of bias.

Second, this is the first meta-analysis of the special educator training group-design literature to demonstrate that training can have a substantial impact on practitioner behavior. Given the lack of precedent for interpreting effect sizes related to practitioner implementation, one way to interpret the mean effect size in this meta-analysis is to convert the standardized mean difference back into the same metric as one of the original outcome measures (Hill et al., 2008). Based on the pre-treatment distributions reported in included studies, an effect size of g = 1.08 equates to special educators providing 11 additional opportunities for students to respond to academic requests every 10 min (Sutherland & Wehby, 2001), special educators and paraprofessionals implementing two new strategies associated with Joint Attention and Symbolic Play/Engagement and Regulation (JASPER; Lawton & Kasari, 2012), or special educators and paraprofessionals implementing direct instruction with a 20% improvement in quality (Collier, 2008). Although such effects seem meaningful, our analysis does not allow us to define a relationship between change in practitioner implementation and student outcomes. Such a relationship would almost certainly be moderated by the nature of the intervention and the level of baseline implementation.

Third, studies involving a combination of modeling and feedback were most strongly associated with larger effects, consistent with the observation from the single-case design literature that these two strategies may promote improved implementation fidelity. These two strategies may work particularly well in tandem. Modeling of steps prior to implementation likely reduces the errors that need to be addressed with performance feedback. In addition, re-modeling steps during performance feedback likely aides the clarity of the feedback. Notably, seven of the eight studies with this combination of training strategies also included training in a one-to-one coaching format. The overlap between training strategies and format does not allow us to disentangle effects between the two, meaning that we are unable to conclude whether these training strategies would be as effective in a group training format. The overlap between format and strategies is not surprising given that it is easier to provide performance feedback in a coaching format. In a coaching format, a trainer can simply visit each educator's classroom individually to observe and provide performance feedback. In a group format, more creative solutions are required. For example, Ascione and Borg (1980, 1983) demonstrated it is possible to provide feedback in a group training format by directing teachers to audio record a sample of their teaching, bring this sample to a group training, and then take turns listening and critiquing each other in partners and small groups with input from the trainer. Other possible strategies might include video recording or inviting students with disabilities to attend a portion of the group training, although no study in this review included either approach. Given that coaching is not a feasible means to train all educators to implement all of the evidence-based practices needed to promote optimal student outcomes (Russo, 2004), further research is needed to better understand whether and how promising training

strategies can be delivered effectively in the context of a group training format.

Fourth, increased duration of training was not associated with a larger effect size. Surprisingly, the direction of the (statistically nonsignificant) relationship was in the opposite direction of what we expected. This finding suggests that increasing the length of training alone—without considering training strategies or format—is unlikely to affect implementation fidelity. The (sizable, but statistically nonsignificant) relationship between shorter training time and larger effect sizes may have been driven by the tendency for studies that had some of the largest effect sizes to also involve brief coaching with modeling and performance feedback. For example, the two studies with the largest individual effect sizes (i.e., Brock & Carter, 2013; Lawton & Kasari, 2012) included these features and also had the shortest and third shortest duration of training relative to other studies. Therefore, it seems that decisions by a few research teams to design shorter trainings with the most promising features might explain the unexpected direction of this relationship, especially given that this particular analysis only included the nine studies for which this information was available.

Limitations and Directions for Future Research

Reviewed literature. Several limitations of the reviewed literature could be addressed in future research. Measures of implementation fidelity can be based on different conceptualizations and vary widely in quality. For all studies included in this review, researchers developed their own measures of implementation fidelity using different strategies. The validity and reliability of these measures are not clear, nor is it known for certain whether these measures all capture the same underlying construct. Furthermore, it is likely that certain kinds of measures such as frequency counts of teacher behavior are not normally distributed (DeMaris, 2004). Future research should focus on how to best approach measurement of implementation fidelity and collect descriptive data on larger samples of practitioners to ascertain whether these measures conform to normal distributions. In addition, the studies included in this review did not include maintenance outcomes, so we can only make conclusions about immediate changes from educator training. Future research should focus on how educator training affects implementation fidelity after all training has been withdrawn. Moreover, we were unable to explore the role of training providers (i.e., school staff vs. researchers) as training in just one study was delivered by educators. Finally, not all of the practices implemented by educators in these 12 studies could be called evidence-based. Although some practices have been shown to promote positive effects across many studies and research groups, including opportunities to respond (for a review, see MacSuga-Gage & Simonsen, 2015) and time delay (for a review, see Wong et al., 2015), others do not have strong empirical support, nor did the authors measure student outcomes and detect a statistically significant student-level effect.

Meta-analysis method. Several limitations of the meta-analysis methodology could be addressed in future research. Both the small quantity of studies and the proportion of these studies with threats to internal validity limit generalization of findings. The small sample size is a particular concern for analysis of training duration, because this analysis only included nine studies. However, sensitivity analysis did not identify evidence of publication bias, and inclusion of studies with threats to internal validity actually resulted in a more conservative estimate of the overall effect size. Furthermore, a strength of this study was the inclusion of theses and dissertations in an effort to include all relevant studies. Nonetheless, we advise the reader to interpret the findings of this meta-analysis with caution. In addition, this review included studies in which in-service teachers, preservice teachers, and paraprofessionals were trained to implement practices. It is possible that effects might differ across these three types of practitioners, although we did not identify any such pattern across studies in this review.

Implications for Practice

Based on the findings of this review, we make two specific recommendations for administrators, technical assistance providers, teacher and paraprofessional training programs, and policy makers. However, we make these recommendations cautiously given the small number of studies and the identified threats to internal validity. First, the reach of professional development programs should be measured not in terms of the number of training hours, but in terms of observable change in educator behavior. Results of this meta-analysis suggest increased training time alone—without considering quality indicators—does little to change educator behavior in the classroom. Therefore, a measure of training hours alone may be completely unrelated to improved implementation fidelity. Practical tools exist for observing and documenting improved implementation of evidence-based practices in the classroom (e.g., Neitzel & Wolery, 2009). These tools could be used to measure the success of training opportunities and to identify educators who would benefit from further training and support.

Second, training opportunities should be designed to include a combination of modeling and performance feedback. This recommendation is not based solely on the present study, but rather on aligned findings across this meta-analysis of group-design literature and previous reviews of the single-case design literature. One practical and cost-efficient means to achieve this goal is to provide an initial training workshop with follow-up training in a

one-to-one coaching format (cf. Brock & Carter, 2013). This was the approach taken by many studies in this review. Many school systems already offer stand-alone training workshops that could be adapted to emphasize modeling of implementation steps and role-play with performance feedback. A follow-up component could be added to include in-classroom coaching with both modeling and performance feedback.

Closing the research-to-practice gap in special education requires effective methods for training educators to implement evidence-based practices. The results of this meta-analysis suggest that a combination of effective strategies (i.e., modeling and performance feedback) and format (i.e., coaching) can promote improved practitioner implementation. Efforts to promote improved outcomes for students with disabilities hinge on using effective strategies to enable educators to implement evidence-based practices with fidelity.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Partial support for this research came from the Office of Special Education Programs, U.S. Department of Education, through Grant H325D100010 to Vanderbilt University.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior Research and Therapy*, *31*, 621–631.

*Ascione, F. R., & Borg, W. R. (1980). Effects of a training program on teacher behavior and handicapped children's self-concepts. *Journal of Psychology*, *104*, 53–65.

*Ascione, F. R., & Borg, W. R. (1983). A teacher-training program to enhance mainstreamed, handicapped pupils' self-concepts. *Journal of School Psychology*, *21*, 297–309. doi:10.1016/0022-4405(83)90043-2

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York, NY: Russell Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. West Sussex, UK: Wiley.

Brock, M. E., & Carter, E. W. (2013). A systematic review of paraprofessional-delivered instruction to improve outcomes for students with intellectual and developmental disabilities. *Research and Practice for Persons With Severe Disabilities*, *38*, 211–221. doi:10.1177/154079691303800401

*Brock, M. E., & Carter, E. W. (2015). Effects of a professional development package to prepare special education paraprofessionals to implement evidence-based practice. *The Journal of Special Education*, *49*, 39–51. doi:10.1177/0022466913501882 2013

Brock, M. E., Huber, H. B., Carter, E. W., Juarez, A. P., & Warren, Z. E. (2014). Statewide assessment of professional development needs related to educating students with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities*, *29*, 67–79. doi:10.1177/1088357614522290

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

*Collier, P. R. (2008). *The impact of literacy coaching on teacher fidelity and students with learning disabilities' reading achievement* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Full Text. (UMI No. 3347608)

Cook, B. G., & Cook, S. C. (2013). Unraveling evidence-based practices in special education. *The Journal of Special Education*, *47*, 71–82. doi:10.1177/0022466911420877

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, *79*, 135–144. doi:10.1177/001440291307900201

Cook, B. G., Smith, G. J., & Tankersley, M. (2012). Evidence-based practices in education. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook* (Vol. 1, pp. 495–528). Washington, DC: American Psychological Association.

DeMaris, A. (2004). *Regression with social data: Modeling continuous and limited response variables*. Hoboken, NJ: Wiley.

*Dixon, M. E. (1983). *Questioning strategy instruction participation and reading comprehension of learning disabled students* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Full Text. (UMI No. 303114608)

Duval, S., & Tweedie, R. (2000). Trim-and-fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463. doi:10.2307/2669529

Education for All Handicapped Children Act, U.S.C. § 1400 (1975).

Egger, M., Smith, D. G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634. doi:10.1136/bmj.315.7109.629

Fallon, L. M., Collier-Meek, M. A., Maggin, D. M., Sanetti, L. M., & Johnson, A. H. (2015). Is performance feedback for educators an evidence-based practice? A systematic review and evaluation based on single-case research. *Exceptional Children*, *81*, 227–246. doi:10.1177/0014402914551738

*Fink, J. (1980). *An investigation into the effects of an inservice program in learning disabilities* (Unpublished doctoral dissertation). Vanderbilt University, Nashville, TN.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *The Journal of Educational Research*, *83*, 103–110.

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., & Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, *343*, 1–9. doi:10.1136/bmj.d5928

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177. doi:10.1111/j.1750-8606.2008.00061.x

*Hindman, S. E., & Polsgrove, L. (1988). Differential effects of feedback on preservice teacher behavior. *Teacher Education and Special Education*, *11*, 25–29. doi:10.1177/088840648801100104

Individuals With Disabilities Education Improvement Act, 20 U.S.C. § 1400 (2004).

Kretlow, A. G., & Bartholomew, C. C. (2010). Using coaching to improve the fidelity of evidence-based practices: A review of studies. *Teacher Education and Special Education*, *33*, 279–299. doi:10.1177/0888406410371643

*Lawton, K., & Kasari, C. (2012). Teacher-implemented joint attention intervention: Pilot randomized controlled study for preschoolers with autism. *Journal of Consulting and Clinical Psychology*, *80*, 687–693. doi:10.1037/a0028506

MacSuga-Gage, A. S., & Simonsen, B. (2015). Examining the effects of teacher-directed opportunities to respond on student outcomes: A systematic review of the literature. *Education & Treatment of Children*, *38*, 211–239.

Maggin, D. M., & Chafouleas, S. M. (2013). Issues and advances of synthesizing single-case research. *Remedial and Special Education*, *34*, 3–8. doi:10.1177/0741932512466269

Neitzel, J., & Wolery, M. (2009). *Implementation checklist for time delay*. Chapel Hill, NC: The National Professional Development Center on Autism Spectrum Disorders.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Noell, G. H., Gansle, K. A., Mevers, J. L., Knox, R. M., Mintz, J. C., & Dahir, A. (2014). Improving treatment plan implementation in schools: A meta-analysis of single subject design studies. *Journal of Behavioral Education*, *23*, 168–191. doi:10.1007/s10864-013-9177-1

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, *75*, 135–150. doi:10.1177/001440290907500201

*Peterson, C. A., & McConnell, S. R. (1996). Factors related to intervention integrity and child outcome in social skills interventions. *Journal of Early Intervention*, *20*, 146–164. doi:10.1177/105381519602000206

Russo, A. (2004). School-based coaching. *Harvard Education Letter*, *20*, 1–4.

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of the single-case literature. *School Psychology Review*, *41*, 160–175.

*Sutherland, K. S., & Wehby, J. H. (2001). The effect of self-evaluation on teaching behavior in classrooms for students with emotional and behavioral disorders. *The Journal of Special Education*, *35*, 161–171. doi:10.1177/002246690103500306

*Weiner, K. B. (2010). *Improving instructional assistant effectiveness in inclusive settings* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Full Text. (UMI No. 964173012)

Whitten, T. M. (1986). *The effect of inservice training on the instructional behaviors of teachers of profoundly handicapped students* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Full Text. (UMI No. 303466321)

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*, 18–28. doi:10.1177/0022466908328009

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., . . . Shultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disabilities*, *45*, 1951–1966. doi:10.1007/s10803-014-2351-z