Exceptional Children
2017, Vol. 84(1) 7–26

DOI: 10.1177/0014402917698008
journals.sagepub.com/home/ecx
SAGE

Findings Across Practitioner Training Studies in Special Education: A Comprehensive Review and Meta-Analysis

Matthew E. Brock[1], Helen I. Cannella-Malone[2], Rachel L. Seaman[2], Natalie R. Andzik[3], John M. Schaefer[2], E. Justin Page[4], Mary A. Barczak[2], and Scott A. Dueker[2]

Abstract

Existing reviews address important questions about subsets of practitioner training studies in special education but leave important questions about the broader literature unanswered. In this comprehensive review, we identified 118 peer-reviewed single-case-design studies in which researchers tested the efficacy of practitioner training on implementation of educational practices to students with disabilities. We found publication of studies has proliferated in recent years, and most studies involved a multiple-baseline or multiple-probe design, researchers as training agents, in-service special education teachers or paraprofessionals as trainees, and students with learning disabilities or autism spectrum disorder as recipients of intervention. Through visual analysis, we detected 521 effects out of 626 opportunities across studies. The mean *d*-Hedges-Pustejovky-Shadish effect size was $d = 2.48$. Behavioral-skills training was associated with the most consistent improvement of implementation fidelity. We found statistically significant associations between implementation fidelity and modeling, written instructions for implementation, and verbal performance feedback.

Despite a legal mandate for special educators to implement practices that have been shown to improve outcomes for students with disabilities (i.e., the No Child Left Behind Act of 2001[NCLB], 2006), many practitioners struggle to implement these evidence-based practices with fidelity (Cook & Odom, 2013). When practices are not implemented with fidelity, or as they were originally designed, then they are no longer supported by research evidence. From their measurement of baseline performance, researchers have documented that many teachers, paraprofessionals, and related service personnel do not implement these practices with implementation fidelity (e.g., Odom, Cox, Brock, & National Professional Development Center on Autism Spectrum Disorders, 2013). It is imperative that researchers identify the best training approaches for closing this research-to-practice gap. *Training* refers to the provision of any activity or material designed to promote or improve implementation of an educational practice (Brock & Carter, 2016). These training approaches should address two key issues. First, they must be effective. Training should enable practitioners to implement practices with fidelity and promote positive outcomes for students with disabilities. Second, given that resources are limited, training must be efficient

[1]Crane Center for Early Childhood Research and Policy, The Ohio State University
[2]The Ohio State University
[3]Northern Illinois University
[4]Duquesne University

Corresponding Author:
Matthew E. Brock, PhD, 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.
E-mail: brock.184@osu.edu

and feasible. For many individuals tasked with providing training (e.g., administrators who support teachers, teachers who supervise paraprofessionals), training is only one of their many responsibilities. They need training strategies that have been optimized to produce the most effective results in the least amount of time. Researchers need to provide solutions that are both effective and efficient.

In order to identify these solutions, researchers should synthesize the existing research evidence on practitioner training and compare the relative efficacy of different training approaches. Although there are a number of existing reviews on practitioner training that provide initial insight into effective training practices, the limited scope of the reviews does not allow for these important comparisons. Instead, most existing reviews focus on a narrow subset of training strategies, practitioners, or training contexts. For example, several research groups have conducted reviews of studies testing the efficacy of performance feedback as a training tool (e.g., Fallon, Collier-Meek, Maggin, Sanettti, & Johnson, 2015; Scheeler, Ruhl, & McAfee, 2004; Solomon, Klein, & Politylo, 2012). Across reviews, these researchers found that performance feedback meets What Works Clearinghouse criteria as an evidence-based practice, that more immediate feedback tends to be more effective than delayed feedback, and that performance feedback tends to have larger effects when teachers are targeting academic skills compared to addressing challenging behavior. Other literature reviews have focused on training specific practitioners and student populations, such as paraprofessionals implementing interventions with students with all developmental disabilities (Brock & Carter, 2013) or only with children with autism spectrum disorder (Rispoli, Neely, Lang, & Ganz, 2011). Across reviews, these researchers found that modeling, role-play, and feedback were common features of effective training. Other reviews have focused on contextual features, such as training that occurred in a one-to-one coaching context (e.g., Kretlow & Bartholomew, 2010), and have found that effective coaching often included small-group initial training followed by repeated one-to-one coaching visits involving observations, modeling, and feedback.

Training should enable practitioners to implement practices with fidelity and promote positive outcomes for students with disabilities.

Although these existing literature reviews are useful for drawing conclusions about subsets of the practitioner training literature, they leave broader questions unanswered. First, the focus on subsets of the literature, in combination with a focus on small windows of time (e.g., articles published in the past 20 years; Kretlow & Bartholomew, 2010), does not allow one to discern overall patterns of publication in practitioner training. Better understanding trends in how evidence has accumulated over time may provide context for making recommendations for future research. Second, reviews that focus on a single training strategy or types of practitioners in isolation do not provide a clear picture of the diversity of training strategies that have been used across studies, the range of practitioners who have been trained, the heterogeneity of practices that these practitioners were trained to implement, or the variety of different profiles of students with whom they intervened. Third, and perhaps most important, previous reviews do not allow one to gauge the relative efficacy of one kind of training compared to another. For example, Fallon and colleagues (2015) clearly established that performance feedback is a training strategy with a strong evidence base, but the relative efficacy of performance feedback compared to other types of training remains unclear. In addition, findings by Solomon and colleagues (2012) suggest that for the body of research they reviewed, practitioners might acquire implementation of some kinds of practices faster than others. This raises questions about whether it is easier to train practitioners to implement certain practices and whether different types of training are better suited to

targeting particular practices. These limitations of the existing literature can be addressed only with a broad, comprehensive review of practitioner training studies.

One recent review of the literature provides some initial insight into these unanswered questions—but only for group design studies. In a recently published meta-analysis, Brock and Carter (2016) began to address these limitations by conducting a comprehensive review and meta-analysis of the group design literature (i.e., randomized controlled trials and quasiexperimental trials). In their analysis of 12 group design studies, they found that practitioner training had a sizable overall effect on implementation fidelity and that a combination of modeling and performance feedback training strategies tended to have a larger effect size. However, the strength of their conclusions was limited by the very small number of group design studies on this topic and the large proportion of these studies that had significant threats to internal validity. In addition, it is clear from the aforementioned reviews of subsets of the practitioner training literature that single-case-design studies on this topic vastly outnumber those with a group design.

In this article, we describe a comprehensive review of the single-case-design literature and discuss how our findings intersect with the aforementioned comprehensive review of the group-design literature. In this way, we are able to address broad questions that can be answered only by reviewing the entire practitioner training literature in special education. Specifically, we address the following research questions. First, what are the patterns of publication for studies that test the efficacy of practitioner training on implementation fidelity across peer-reviewed journals and time? Second, what types of practitioners were trained in these studies, what were the disabilities of the students who received practitioner-implemented practices, and where did they implement these practices? Third, what practices were practitioners trained to implement, and what student outcomes were targeted by these practices? Fourth, who trained the practitioners, what kinds of training strategies did they use, and how long did training last? Fifth, what were the effects of practitioner training on initial practitioner implementation fidelity, training on maintenance of implementation fidelity, and practitioner implementation on student outcomes? Finally, how did effects differ based on (a) the training strategies used and (b) the practices that practitioners were trained to implement?

Method

This literature review was implemented by a team of two faculty members and six advanced doctoral students in special education. All individuals had advanced training in applied behavior analysis and experience with conducting and evaluating single-case-design studies.

Study Eligibility Criteria

To be included in this review, we required studies to meet the following criteria. First, studies had to be available in English and published in a peer-reviewed journal. Second, studies must have included a single-case design with at least three opportunities to demonstrate and replicate an experimental effect (i.e., the standard for a high-quality single-case design; Kratochwill et al., 2010) for practitioner training. Examples of designs with three opportunities to demonstrate effects include multiple-baseline designs with at least three staggered tiers or ABAB withdrawal designs. We define *practitioner training* as the provision of any training activity or material designed to promote or improve implementation of an educational practice (Brock & Carter, 2016); this definition includes training activities during or after initial implementation (e.g., performance feedback). Single-case-design studies with only two opportunities for demonstration of an experimental effect, qualitative studies, and descriptive studies were excluded. Group-design studies were excluded because they have already been captured in a parallel review (i.e., Brock & Carter, 2016). Third, the dependent variable must have been implementation fidelity of an educational practice

delivered to a student with an identified disability by a school-based practitioner (e.g., in-service or preservice teachers, paraprofessionals, or related service personnel). We define *implementation fidelity* as measurement of "how well an intervention is implemented in comparison with the original program design" (O'Donnell, 2008, p. 33). For examples, see Approach to Measuring Implementation Fidelity in the Results section. We define *educational practices* as any teacher behavior that aims to improve student outcomes; this definition includes both focused interventions as well as comprehensive treatment models (Wong et al., 2015). Studies were excluded in which practitioners implemented an educational practice, but only student outcomes (and not practitioner implementation fidelity) were measured and graphed as the dependent variable. Studies that included only collection of assessment data (e.g., preference assessment, functional behavior assessment) but lacked delivery of an intervention were excluded.

Search Strategy

We used multiple search strategies to ensure all studies meeting the above criteria were identified (see Figure 1). First, in March 2016, we searched four electronic databases: PsycINFO, ERIC, Social Services Abstract, and Education Research Complete Academic. The complete search string is available from the first author. This electronic search yielded 5,223 hits. We used a two-step process to screen articles. Based on review of the title and abstract, we excluded any study that clearly (a) contained no original data, (b) did not include practitioner implementation, or (c) utilized qualitative methodology. We measured reliability of this initial screening by having a second person screen 20% of all initial hits. Overall agreement was 83%. Any article for which two screeners disagreed was retained for subsequent review; however, in no case did a disagreement relate to an article that was ultimately included after full-text review. For any article not yet excluded, we also reviewed the results section(s) and excluded the article if it did not include any graphical representation of data. After screening, the remaining 825 articles were reviewed at the full-text level to determine if they met the three aforementioned study eligibility criteria. Eighty-eight articles met eligibility criteria. Next, we reviewed the reference list of each of these 88 articles and each publication that cited one of these articles, and identified 24 additional articles that met inclusion criteria. Finally, we conducted a hand search of the two journals that accounted for the most identified articles through the aforementioned methods (i.e., *Teacher Education and Special Education* and *Journal of Applied Behavior Analysis*) to ensure that we had not overlooked studies. We identified two additional articles that met inclusion criteria through this hand search.

Data Collection and Variables

To address the first research question about publication patterns, we coded publication dates, journals of publication, experimental designs, and whether researchers assessed maintenance and generalization. To address the second research question, related to participants and key variables, we coded practitioner roles and setting, approaches to measuring implementation fidelity, reliability of dependent variables, student demographics, and approaches to measuring student outcomes. To address the third research question, we categorized educational practices and student outcomes. To address the fourth research question, related to training procedures, we coded trainer roles, specific training strategies, and duration of training. To address the final research question, related to efficacy of training procedures and practitioner implementation, we coded both the consistency and magnitude of experimental effects.

Trends in publication. To track patterns of publication across journals and time, we recorded the journal of publication and the publication year for each study.

Experimental design. We coded the type of single-case design used in each study. Specifically, we categorized studies as multiple baseline across participants (i.e., introduction

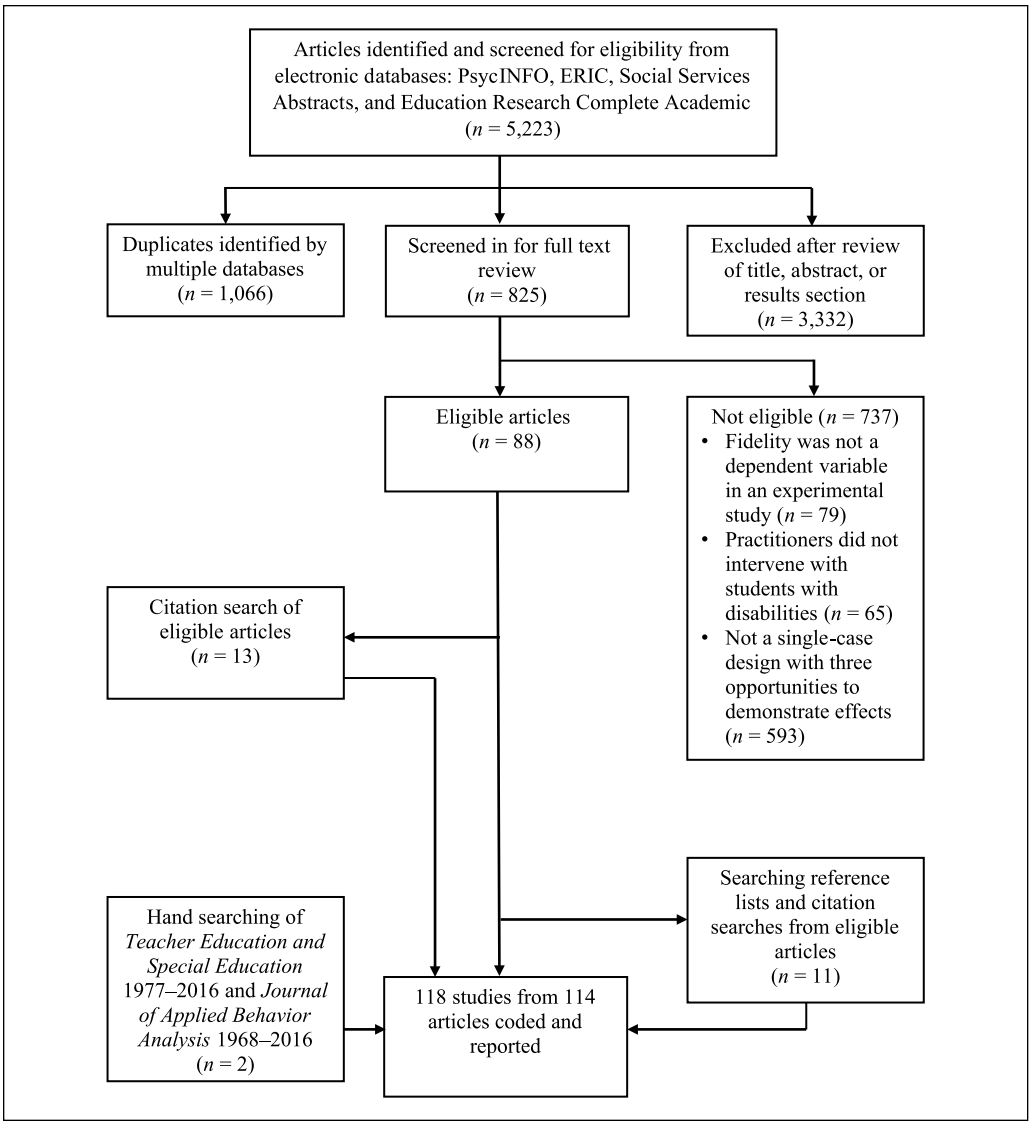

Figure 1. Flow diagram of study search procedures.

of training was staggered across practitioners), multiple baseline across behavior (i.e., introduction of training was staggered across targeted behaviors), multiple baseline across setting (i.e., introduction of training was staggered across settings), withdrawal designs (i.e., training was repeatedly introduced and withdrawn), alternating treatments (i.e., two different training methods were delivered in a rapidly alternating order), changing-criterion designs (i.e., practitioners were trained to meet a series of predetermined criteria), or a combination of the aforementioned designs (e.g., a changing-criterion design embedded within a multiple-baseline-across-participants design).

Reliability of dependent variables. We coded whether the authors of each study (a) reported interobserver agreement (IOA) and (b) if mean IOA exceeded 80%. When authors reported multiple means for IOA for practitioner implementation fidelity but did not report an overall mean, we examined the lowest reported mean. We used 80% as a threshold because this level of agreement has been

established as a benchmark (Kratochwill et al., 2010).

Maintenance and generalization. We coded whether the authors collected and reported maintenance data for practitioner implementation fidelity and how much time elapsed between termination of training and the final data point in the maintenance phase. We categorized the amount of time that elapsed into categories of 0 to 2 weeks, 2 to 4 weeks, 1 to 6 months, 6 to 12 months, and longer than 12 months. Some studies included a maintenance phase but did not provide sufficient description to categorize the time that had elapsed; in these cases, we recorded that the duration was unclear. In addition, we recorded whether the authors reported data for practitioner generalization of implementation to new learners or situations.

Practitioner roles and setting. We coded how many practitioners participated, their respective roles, and the setting in which they intervened with students with disabilities. Based on author description, we categorized roles as preservice special education teacher, in-service special education teacher, preservice general education teacher, in-service general education teacher, paraprofessional, related service provider, trainee working toward certification in behavior analysis, behavior therapist, or administrator. If the authors did not state whether a teacher was a general or special education teacher, but the authors described the teacher working in a self-contained special education classroom, we assigned the category of special education teacher. If authors did not provide sufficient description to discern the role of a practitioner (e.g., describing him or her only as a practitioner or clinician), we categorized the role as unclear. Based on author description, we categorized intervention settings as self-contained special education classrooms, general education classrooms, nonclassrooms (e.g., cafeteria), a combination of settings, or not reported.

Practitioner outcomes. We began coding using the categories and definitions developed by Wong et al. (2015) and added additional categories for practitioner behaviors that were not captured by this framework. Categories we adopted from Wong et al. included antecedent-based intervention, differential reinforcement, discrete trial training, extinction, functional communication training, modeling, naturalistic intervention, peer-mediated instruction and intervention, Picture Exchange Communication System (PECS), pivotal response training, prompting, reinforcement, response interruption or redirection, scripting, structured play group, time delay, and visual supports. When authors described a practice that included components that could also be categorized as focused practices (e.g., functional communication training includes both prompting and reinforcement), we coded only the broader practice and not the components. We also developed additional categories, including opportunities to respond (i.e., inviting a student response), appropriate curricular focus (i.e., implementation of instruction that aligned with the general education curriculum or particular student goals), and prescribed instructional sequence (i.e., implementing sequential steps of a researcher-developed lesson that did not align with a practice described in any other study). In addition, we coded how practitioner implementation fidelity was measured and indexed. Specifically, we categorized each implementation fidelity variable as (a) proportion of sequential steps that were completed, (b) proportion of nonsequential components demonstrated, (c) frequency of behavior, or (d) duration of behavior. The first two categories are aligned with adherence to implementation, and the second two categories are aligned with duration (O'Donnell, 2008).

Student demographics. We coded if students had a diagnosis of autism spectrum disorder, deafblindness, emotional disturbance, hearing impairment, intellectual disability, multiple disabilities, orthopedic impairment, health impairment, learning disabilities, speech or language impairment, traumatic brain injury, visual impairment, developmental delay, or developmental disability. We also coded the number of students with disabilities for whom a label was not reported and the number of students without disabilities who received

intervention alongside students with disabilities. We coded the number of students by grade-level category, including preschool, elementary school (Grades K–5), middle school (Grades 6–8), and high school (Grades 9–12). If authors reported age but not grade level, we categorized children ages 3 to 5 as preschool, ages 6 to 11 as elementary, ages 12 to 14 as middle school, and ages 15 to 18 as high school, and students 19 years of age or older separately.

Student outcomes. Regardless of whether a study directly measured student behavior, we coded the student-level outcome that was targeted by the practitioner-implemented intervention. We coded whether an intervention targeted an academic (i.e., academic skill or content area), social (i.e., social interactions or networks), communication (i.e., verbal, augmentative, or alternative communication), problem behavior (i.e., problem behavior or alternative behaviors), daily living skills (i.e., skills related to daily living, such as dressing or toileting), on-task behavior (i.e., student engagement in appropriate on-task behavior), transition time (i.e., time that students transitioned between activities), play (i.e., play skills, such as pretend or parallel play), imitation (i.e., imitation of actions, movements, or verbalizations; authors explicitly stated that the purpose was imitation and not mastery of the skill being imitated), or vocational outcome (i.e., skill related to a job or vocation).

Trainer roles. We coded if the individual(s) who delivered the training were researchers, consultants, peers (i.e., other practitioners), school administrators, or university supervisors.

Training procedures. On the basis of author description, we coded the strategies used to train practitioners. We began with the coding and framework from Brock and Carter (2016) and added new categories when we identified strategies that did not fit into an existing category. Categories included oral description of instructions (i.e., oral instruction related to defining the practice and its implementation); written description of instructions (i.e., written material that described how to implement the strategy); other written information (i.e., additional written material given to the practitioner); performance feedback (i.e., reinforcement of correct implementation or suggestions for improving future implementation), which we further categorized as verbal, bug-in-ear, written, or video based on delivery; modeling, which we further categorized as live or video; planning (i.e., the trainer collaborated with the practitioner to create intervention plans for applied settings); question-and-answer session (i.e., trainers answered practitioners questions about implementation of the practice); skill rehearsal (i.e., practitioner practiced implementing the practice in the context of the training); script (i.e., practitioner was given a script to repeat verbatim); self-monitoring (i.e., trainers directed practitioners to collect and review implementation fidelity data about their own performance); study groups (i.e., practitioners meet as groups at scheduled times to discuss and/or practice implementation of the practice); behavioral skills training (BST; i.e., authors explicitly stated that BST was used); and goal setting (i.e., practitioners were directed to set performance goals related to their implementation of the practice).

We coded the length of training by categorizing (a) the length of initial training (i.e., training prior to practitioners attempting implementation) and (b) the length of any follow-up training subsequent to initial training. We categorized the length of initial training into categories of 2 hr or less, more than 2 but less than 4 hr, more than 4 but less than 8 hr, and more than 8 hr. When authors reported length in terms of days (i.e., one half day or 3 days) but not hours, we reported the number of days. We also noted when insufficient information was reported to discern duration of training. To capture studies that included further training after practitioners attempted implementation with students, we coded the number of additional training sessions as none, one, multiple, or not reported.

Efficacy of strategies. We used two different approaches to characterize the effectiveness of training on initial practitioner implementation

fidelity and maintenance, and practitioner implementation on student outcomes. First, we calculated success estimates. Originally proposed by Reichow and Volkmar (2010), a success estimate summarizes visual analysis of data as a ratio of the number of times an experimental effect was demonstrated (numerator) to the number of opportunities that an experimental effect could have been demonstrated given the experimental design (denominator). We chose this metric because it is based on established visual analysis methods and is not dependent on the controversial assumptions of other quantitative metrics. We emphasize that a success estimate is not an effect size, because it does not describe the magnitude of effect; it simply summarizes the consistency of effects detected through visual analysis. All success estimates were coded through a consensus process (see reliability section). When visually analyzing data, coders analyzed the level, trend, and variability of data within phases and examined patterns across similar phases (i.e., immediacy of effect, overlap, and consistency of data) in order to determine whether an experimental effect was demonstrated (What Works Clearinghouse, 2014). To examine the relative efficacy of training based on a specific factor, we calculated conditional success estimates for subsets of studies. This involved summing success estimates for subgroups of studies that included a given factor. We interpret success estimates as an indicator of the consistency of effects.

Second, to estimate the magnitude of effects, we calculated a *d*-Hedges-Pustejovsky-Shadish (DHPS) effect size (Hedges, Pustejovsky, & Shadish, 2013). DHPS is calculated using a hierarchical model to produce a between-subjects effect size. This method can be used to calculate an effect size for withdrawal designs that include at least three alternations of baseline and intervention phases (e.g., ABAB) across three participants or multiple-baseline-across-participant experiments with at least three tiers. In addition, there must be adequate variability among cases to produce an estimate of variance greater than zero. Twenty studies (17% of all studies) from our review were excluded from DHPS analysis because the designs did not meet these requirements (e.g., reversal design with only one participant, multiple baseline across behaviors within one participant). The developers of DHPS designed the parameter to correspond to Cohen's *d*, an effect size metric that is commonly used in analysis of group design studies. Marso and Shadish (2014) developed a macro for SPSS software to calculate DHPS. To utilize this macro, we had to first input each individual data point from each study into an SPSS database in the format prescribed by Marso and Shadish. We used a digitizer computer application (i.e., Engauge) to extract precise data points from electronic images of graphs.

After calculating DHPS effect size and variance for each individual study, 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-studies variance to estimate a distribution of true effects (Borenstein, Hedges, Higgins, & Rothstein, 2011). We computed the mean effect size using the "metan" macro for Stata.

Next, we ran meta-regression models using the "metareg" macro for Stata. The purpose of this analysis was to calculate correlations between variables of interest and study-level effect sizes. First, we ran a null model (without any predictor variables) to calculate the distribution of effect sizes across studies and determine if this variance could be attributed to true heterogeneity among studies. Then we ran separate single-predictor models with each training strategy or practice as the independent variable and study-level effect sizes as the dependent variable.

Coder Training and Reliability

The first author trained all coders by (a) providing a detailed coding manual, (b) reviewing the coding manual through oral instruction, (c) assigning practice studies to code, and (d) providing detailed feedback on disagreements. Coders did not begin coding studies

for this review until they achieved 95% agreement with the first author on a practice study.

We computed reliability for (a) study characteristics, (b) visual analysis, and (c) digitization. To compute the reliability for study characteristics, a second independent coder analyzed 22% of studies. Point-by-point agreement (i.e., exact agreements divided by opportunities for agreement) was calculated for each variable. Average agreement across all variables was 95.6% (range: 83%–100%). Reviewers resolved all discrepancies through consensus. A second independent coder visually analyzed the results of every study to compute a success estimate. Initial agreement for visual analysis was 90.3%, and any disagreements between the two coders were resolved through consensus. To ensure accuracy of digitized data for DHPS analysis, a second coder digitized 21.6% of all studies. We calculated point-by-point agreement, with an agreement defined as a value within 2% of the maximum value on the *y*-axis. Agreement was 96.2%, and all disagreements were resolved by having the two coders reexamine the data together and come to consensus.

Results

Overall Patterns in Publication

We identified 118 studies in 114 articles published in 36 different peer-reviewed journals. A complete list of these articles is available from the first author. Nearly one third of all studies were published in two journals: *Journal of Applied Behavior Analysis* ($n = 20$; 17% of studies) and *Teacher Education and Special Education* ($n = 15$; 13%). The number of studies that tested training on practitioner implementation has increased rapidly in recent years. Nearly two thirds (i.e., $n = 76$; 64%) of all studies were published in the past 10 years, with only 42 studies published prior to 2007.

Experimental Design and Measurement

The majority of studies ($n = 98$; 83%) utilized a multiple-baseline- or multiple-probe-across-participants design. Other studies used a multiple-baseline- or multiple-probe-across-behavior design ($n = 12$; 10%), a multiple-baseline- or multiple-probe-across-settings design ($n = 1$; 1%), a withdrawal design ($n = 2$; 2%), an alternating-treatment design ($n = 1$; 1%), a changing-criterion design ($n = 1$; 1%), or a combination of the aforementioned designs ($n = 3$; 3%). Over half of all studies reported maintenance of practitioner implementation ($n = 67$; 57%), less than a third reported generalization of practitioner implementation ($n = 34$; 29%), and less than half reported student outcome data ($n = 55$; 47%). When maintenance data were reported, 13 studies collected the last data point 2 to 4 weeks after training ended, 29 studies 1 to 6 months later, three studies 6 to 12 months later, and one study more than 1 year later. In 21 studies, author description was not sufficient to determine when maintenance data were collected. Nearly all (117 of the 118 studies) reported IOA. Of those 117 studies, 105 (89.7%) reported average agreement of 80% or above on practitioner implementation fidelity variables.

Practitioners, Students, and Settings

Across studies, 475 practitioners were trained to implement interventions with 642 students with disabilities. The numbers of practitioners in each role and students in each disability category are reported in Table 1. In addition, 971 students without identified disabilities received interventions alongside their peers with disabilities (i.e., in some studies an intervention was delivered to an entire general education classroom). In 22 studies, the authors did not report the number of students to whom the practice was delivered. Practitioners implemented practices with 549 preschool students (41% of all students), 176 elementary students (13%), 69 middle school students (5%), 546 high school students (41%), and three students 19 to 22 without a specified grade level (<1%). In 33 studies, authors did not report the number of students in a specific grade level, nor did they report student age.

Table 1. Practitioner Roles and Student Disability Labels Across Eligible Studies.

Variable	Number	% of total
Practitioners who received training	475	100
Special education in-service teachers	207	44
Paraprofessionals	106	22
General education in-service teachers	76	16
Special education preservice teachers	55	12
General education preservice teachers	13	3
Related service personnel	10	2
Students in applied behavior analysis	3	1
University students (no further description)	3	1
Clinician (no further description)	2	<1
Administrator	1	<1
Behavior therapist	1	<1
Students who received intervention	642	100
Learning disability	129	20
Autism spectrum disorder	125	19
Unspecified disability	112	17
Intellectual disability	98	15
Emotional disturbance	38	6
Developmental delay	37	6
Multiple disability categories	35	5
Developmental disability	33	5
Other health impairment	12	2
Speech or language impairment	9	1
Multiple disabilities	6	1
Orthopedic impairment	4	1
Traumatic brain injury	2	<1
Deafblind	1	<1
Hearing impairment	1	<1
Visual impairment	0	0

Note. A student was assigned to the category of "multiple disabilities" when authors simply reported that the student had multiple disabilities; a student was assigned to the category of "multiple disability categories" when authors indicated the student met criteria for two or more categories (e.g., both autism spectrum disorder and intellectual disability).

Practitioners implemented practices with students in self-contained special education classrooms (n = 54; 46% of studies), general education classrooms (n = 43; 36%), nonclassroom school settings (e.g., playground or lunchroom; n = 2; 2%), or a combination of the aforementioned settings (n = 8; 7%). In 11 studies, authors provided insufficient description to determine where practices were implemented.

Practitioner-Implemented Practices and Targeted Student Outcomes

Practitioners were trained to implement 25 different practices with students with disabilities. Twenty-two of these practices could be categorized as focused intervention practices; the other three (i.e., appropriate curricular focus, prescribed instructional sequence, and comprehensive social competence intervention) involved a shift in what teachers were targeting (rather than the teaching procedure), an emphasis on following a researcher-designed lesson that did not include a specific intervention practice, and a comprehensive treatment model for which the component focused interventions were not reported. The most commonly targeted practices were reinforcement (n = 55; 47% of studies) and prompting (n = 43; 36%), and the most common combination of strategies

was reinforcement and prompting (n = 30; 25%). In more than half of studies (n = 60; 51%), practitioners were trained to implement a single practice. Practitioners were trained to implement two practices in 34 studies (29%), three practices in 16 studies (13%), four practices in six studies (5%), six practices in one study (1%), and seven practices in one study (1%). The number of studies that focused on each strategy is reported in Table 2.

In most studies, practitioners targeted a single student outcome (n = 81; 69% of studies). Practitioners targeted two outcomes in 17 studies, three outcomes in seven studies, and five outcomes in two studies. Practitioners targeted communication outcomes in 39 studies, academic outcomes in 34 studies, modification of problem behavior in 30 studies, on-task behavior in 15 studies, social outcomes in 12 studies, daily living skills in seven studies, play skills in five studies, decreased transition time in two studies, vocational skills in two studies, imitation skills in two studies, ambulation skills in one study, following directions in one study, leisure skills in one study, independence with instructional activities in one study, and fine motor skills in one study. In eight studies, the authors did not provide sufficient description to categorize the targeted student outcome (e.g., teachers used discrete trial training to target individualized student outcomes).

Approach to Measuring Implementation Fidelity

Sixty-two studies involved measurement of the adherence dimension of implementation fidelity (O'Donnell, 2008). Specific approaches included measuring the percentage or proportion of nonsequential components (36 studies; 31%) and the percentage or proportion of sequential steps (26 studies; 22%). Fifty-six studies involved measurement of the duration dimension of implementation fidelity. In 55 studies (47%), implementation fidelity was measured as the frequency or rate of one or more discrete practitioner behaviors, and in one study (1%), as the duration of a practitioner behavior.

Trainers, Strategies Used to Train Practitioners, and Length of Training

In most studies, the training agent was a researcher (n = 97; 82%). Practitioners were trained by peers (e.g., another teacher) in six studies, a school administrator in two studies, a consultant in two studies, and a university supervisor in two studies. In eight studies, it was unclear who provided training.

Training agents used 22 different training strategies. The number of studies in which each training strategy was used is reported in Table 3. The most common strategy was performance feedback (n = 102; 86% of studies), and the most common type of performance feedback was verbal feedback (n = 88; 64%). Trainers always used a combination of strategies. A combination of two strategies was used in four studies (3%), three strategies in six studies (5%), four strategies in 17 studies (14%), five strategies in 12 studies (10%), six strategies in 22 studies (19%), seven strategies in 17 studies (14%), eight strategies in 16 studies (14%), nine strategies in 14 studies (12%), 10 strategies in five studies (4%), 11 strategies in three studies (3%), and 12 strategies in two studies (2%).

Duration of training varied greatly across studies. In eight studies, no initial training session was provided prior to implementation (e.g., training involved only individualized performance feedback after implementation was attempted). When reported in hours, we categorized initial training sessions into categories of less than 2 hr (n = 62), 2 to 4 hr (n = 8), 4 to 8 hr (n = 4), and more than 8 hr (n = 3). In some cases, duration of training was not reported in hours but was reported in days. In these cases, training lasted 1 day (n = 2), 3 days (n = 2), or 10 days (n = 1). In 33 studies, authors did not clearly describe the length of the initial training session in hours or days. Most studies (n = 109; 92%) involved follow-up training after practitioners attempted implementation. Only nine studies (8%) involved no follow-up training, four studies (3%) involved one follow-up training session, and 94 studies (80%) involved multiple training sessions. In 11 studies (9%), authors

Table 2. Conditional Success Estimates and Variance Explained by Practitioner-Implemented Practices.

		Initial fidelity			Maintenance of fidelity		Student outcomes		
Intervention or practice	No. of studies	Success estimate	%	DHPS adjusted R^2	Success estimate	%	Success estimate	%	DHPS adjusted R^2
Reinforcement	55	256/329	78	5.84%*	99/135	73	68/115	59	6.32%
Prompting	43	221/262	84	1.32%	89/118	75	52/73	71	–2.95%
Antecedent-based intervention	21	91/101	90	0.80%	41/49	84	17/37	46	7.30%
Naturalistic intervention	21	99/118	84	2.24%	32/44	73	35/57	61	22.35%*
Discrete trial training	14	48/57	84	–1.41%	22/27	81	11/17	65	0.79%
Modeling	8	42/58	72	–0.69%	16/29	55	6/9	67	–2.12%
Time delay	7	44/63	70	–1.66%	26/36	72	8/9	89	–2.63%
Visual supports	6	24/27	89	–1.67%	10/13	77	10/20	50	3.92%
Response interruption/redirection	6	34/46	74	0.88%	7/8	88	3/7	43	–2.55%
Differential reinforcement	5	29/33	88	–1.61%	11/12	92	3/7	43	–2.21%
Functional communication training	4	21/22	95	–1.36%	6/6	100	3/11	27	–2.21%
Opportunities to respond	4	15/18	83	–1.38%	6/6	100	6/9	67	–0.91%
Pivotal response training	4	16/25	64	–0.99%	13/19	68	12/13	92	10.82%*
Picture exchange comm. system	2	12/12	100	4.59%	9/9	100	3/3	100	–2.77%
Structured play groups	2	11/11	100	–1.64%	0/0		6/7	86	
Appropriate curricular focus	2	12/12	100	–1.28%	12/12	100	0/0		
Prescribed instructional sequence	2	12/12	100	6.47%*	0/0		4/6	67	–2.49%
Extinction	2	9/12	75	–1.74%	6/9	67	0/0		
Peer-mediated intervention	1	3/4	75	–0.79%	0/0		7/8	88	–2.80%
Scripting	1	2/3	67	–1.22%	0/0		1/3	33	–1.16%
Punishment[a]	1	6/9	67		0/0		0/0		
Consequence-based intervention[b]	1	3/3	100		0/0		0/0		
Safety skills[c]	1	3/3	100		3/3	100	3/3	100	
Delivery of corrective feedback	1	6/6	100		0/0		0/0		
Comprehensive social competence intervention	1	2/3	67		2/2	100	0/0		

Note. Success estimates, originally proposed by Reichow and Volkmar (2010), summarize visual analysis of data as a ratio of the number of times an experimental effect was demonstrated (numerator) to the number of opportunities that an experimental effect could have been demonstrated given the experimental design (denominator). DHPS (Hedges, Pustejovsky, & Shadish, 2013) is a between-subjects effect size metric based on a simple hierarchical linear model, and adjusted R^2 represents that percentage variance in the overall DHPS effect size predicted by an individual variable.
[a]Time-out procedures. [b]Authors did not specify whether procedure involved reinforcement, punishment, or a combination of the two. [c]Guarding behaviors to ensure safety during ambulation.
*$p < .05$.

Table 3. Conditional Success Estimates and Variance Explained by Training Strategies.

Training strategy	No. of studies	Initial fidelity			Maintenance of fidelity		Student outcomes		
		Success estimate	%	DHPS adjusted R^2	Success estimate	%	Success estimate	%	DHPS adjusted R^2
Performance feedback	102	449/534	84	–1.57%	208/255	82	156/232	67	–2.88%
Verbal feedback	88	393/470	84	6.46%*	186/228	82	131/193	68	0.20%
Written feedback	32	127/157	81	–1.01%	55/75	73	31/60	52	13.47%*
Video feedback	12	53/62	85	–0.94%	30/30	100	19/31	61	–1.75%
Bug-in-ear feedback	6	28/39	72	0.00%	14/19	74	1/5	20	–1.21%
Oral description/lecture	87	384/452	85	3.36%	185/243	76	136/198	69	0.30%
Modeling	79	372/440	85	5.93%*	168/217	77	120/187	64	–2.67%
Live modeling	63	305/358	85	8.26%*	141/186	76	107/170	63	–3.02%
Video modeling	22	93/119	78	–1.50%	26/36	72	24/28	86	–2.57%
Written instructions	56	249/284	88	9.71%*	114/142	80	81/122	66	0.07%
Skill rehearsal	54	248/297	84	1.28%	115/147	78	85/121	70	–2.77%
Question/answer session	44	206/235	88	–1.23%	98/123	80	63/105	60	–0.50%
Rationale	37	176/205	86	–1.62%	68/101	67	42/60	70	–1.78%
Planning	31	123/159	77	2.11%	48/66	73	45/69	65	–2.88%
Self-monitoring	18	68/77	88	–1.35%	26/42	62	24/37	65	–2.32%
Other written material	17	69/92	75	–1.12%	37/61	61	15/24	63	11.18%*
Behavioral skills training	7	32/32	100	2.15%	20/20	100	29/46	63	–3.37
Intervention script	4	20/28	71	0.28%	6/7	86	5/7	71	–2.70%
Goal setting	4	11/16	69	1.43%	4/5	80	6/16	38	1.40%
Study groups	1	3/3	100		3/3	100	0/0		

Note. Success estimates, originally proposed by Reichow and Volkmar (2010), summarize visual analysis of data as a ratio of the number of times an experimental effect was demonstrated (numerator) to the number of opportunities that an experimental effect could have been demonstrated given the experimental design (denominator). DHPS (Hedges, Pustejovsky, & Shadish, 2013) is a between-subjects effect size metric based on a simple hierarchical linear model, and adjusted R^2 represents that percentage variance in the overall DHPS effect size predicted by an individual variable.

*$p < .05$.

reported providing follow-up training, but it was unclear if they provided one or multiple training sessions.

Efficacy of Training Strategies

Success estimates. Across all studies, the success estimate for training on initial practitioner implementation was 521/626 (83%), training on maintenance of practitioner implementation was 232/294 (79%), and implementation on student outcomes was 166/249 (67%). Conditional success estimates were calculated based on how practitioners were trained and what they were trained to do and are reported in Tables 2 and 3.

Conditional success estimates were also calculated based on whether studies involved no follow-up training, one follow-up training session, or multiple follow-up training sessions after the initial training. For studies with no follow-up training, the success estimate for training on initial practitioner implementation was 44/46 (96%), training on maintenance of practitioner implementation was 26/29 (90%), and implementation on student outcomes was 12/24 (50%). For the four studies with only one follow-up training session, the success estimate for training on initial practitioner implementation was 25/29 (86%), and implementation on student outcomes was 3/3 (100%); there were no opportunities for effects of training on maintenance of practitioner implementation. For the 93 studies with multiple follow-up training sessions, the success estimate for training on initial practitioner implementation was 412/501 (82%), training on maintenance of practitioner implementation was 181/234 (77%), and implementation on student outcomes was 128/193 (67%).

DHPS effect size for practitioner implementation fidelity. Study-level effect sizes across individual studies ranged from $d = 0.32$ to 43.6, with a mean effect size of $d = 2.48$, 95% confidence interval (CI) [2.22, 2.74]. This mean effect size is very large according to commonly used benchmarks (e.g., Cohen, 1988), although there is limited precedent for interpreting DHPS effect sizes for single-case-design studies. Moderator analysis involved using meta-regression to determine 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 ($\tau^2 = 2.23$) and that the majority of this variance ($I^2 = 96.6\%$) can be attributed to true heterogeneity among studies. Three training strategies had significantly stronger effects on practitioner implementation fidelity, including verbal feedback, $\beta = .84$, $t(97) = 2.10$, $p = .04$; modeling, $\beta = .82$, $t(97) = 2.27$, $p = .03$; and written directions, $\beta = .91$, $t(97) = 2.62$, $p = .01$. Two intervention practices had significantly stronger effects on practitioner implementation fidelity, including reinforcement, $\beta = -.93$, $t(97) = -2.61$, $p = .01$, and prescribed instructional sequence, $\beta = 3.07$, $t(97) = 2.44$, $p = .02$. The proportion of variance (i.e., adjusted R^2) explained by each training strategy and intervention practice is reported in Tables 2 and 3.

DHPS effect size for student outcomes. Study-level effect sizes across individual studies ranged from $d = -0.33$ to 6.45, with a mean effect size of $d = 1.65$, 95% CI [1.41, 1.90]. Estimates from the null meta-regression model suggested a wide distribution of effect sizes across studies ($\tau^2 = 2.02$) and that the majority of this variance ($I^2 = 99.0\%$) can be attributed to true heterogeneity among studies. Two training strategies had significantly stronger effects on student outcomes, including other written material, $\beta = 1.84$, $t(40) = 2.23$, $p = .03$, and written feedback, $\beta = -1.20$, $t(40) = -2.32$, $p = .03$. Two intervention practices had significantly stronger effects on student outcomes, including naturalistic intervention, $\beta = 2.02$, $t(40) = 3.32$, $p < .01$, and pivotal response training, $\beta = 1.83$, $t(40) = 2.19$, $p = .04$. The proportion of variance (i.e., adjusted R^2) explained by each training strategy and intervention practice is reported in Tables 2 and 3.

Discussion

A number of published literature reviews have identified findings about subsets of the practitioner training literature but leave broader questions unanswered. This comprehensive review of the single-case-design literature, a parallel review of a published meta-analysis of the group-design literature, addresses these broader questions. Specifically, we reviewed all single-case-design studies that tested the efficacy of training on practitioner implementation fidelity. We found the number of these studies has proliferated in recent years, and most studies involved a multiple-baseline or multiple-probe design, researchers as training agents, in-service special education teachers or paraprofessionals as trainees, and students with learning disabilities or autism spectrum disorders as recipients of intervention. Through visual analysis, we detected relatively consistent effects of practitioner training on implementation fidelity and less consistent effects of practitioner implementation on student outcomes. BST was associated with the most consistent improvement of implementation fidelity. Through DHPS analysis, we determined that practitioner training has a very large effect size on implementation fidelity. We found that the use of modeling, written instructions for implementation, and oral description of implementation steps were statistically significant predictors of effects. In addition, naturalistic intervention and pivotal response training tended to have the largest effects on student outcomes. Key findings from this review provide new perspective on the state of research focusing on training practitioners to implement practices with students with disabilities.

First, there has been an increased focus on research on practitioner training in recent years. Nearly two thirds of single-case-design studies in this review were published in the past 20 years, and half of the group-design studies were published in the same time frame (Brock & Carter, 2016). This trend corresponds with the timing of a legislative mandate for scientifically based instructional approaches (NCLB, 2006) and a subsequent focus by researchers on implementation science both across disciplines, such as education, mental health, and substance abuse (e.g., Fixsen, Naoom, Blase, Friedman, & Wallace, 2005), and specifically within the field of special education (e.g., Odom, 2009). Indeed, in January 2013, this journal published an entire issue focused on implementation of evidence-based practices (i.e., Cook & Odom, 2013). Although there has been a longer tradition of developing theory and frameworks around what makes practitioner training most effective, findings from this literature review suggest an increasing emphasis on testing the efficacy of specific training methods through rigorous experimental research. Much of this research uses similar methodology. Single-case-design studies outnumber group-design studies nearly 10:1, and the bulk of these studies utilized a multiple-baseline-across-participants design. The large number of multiple-baseline designs stems from researchers most often targeting learned (i.e., nonreversible) behaviors; withdrawal designs were appropriate only in the small number of cases in which researchers used contingencies to manipulate the rate of implementation behaviors (e.g., rate of praise or opportunities to respond). Given the scarcity of comparative designs, it is unsurprising that most studies involved a comparison of experimental practitioner training to business as usual.

It is unclear if the training described in many of these studies would be feasible under typical circumstances.

Despite the commonalities in research methods, most studies did not coalesce around a specific type of practitioner training. Although a few strategies were used more often than others (e.g., modeling and performance feedback), most studies used an idiosyncratic combination of four to 12 training strategies. Only a handful of studies used a similar training package, including studies published in the same article or those involving BST ($n = 8$).

Second, it is unclear if the training described in many of these studies would be feasible under typical circumstances. When

authors reported who provided training, the training agent was almost always a member of an external research team. Further, the majority of studies involved multiple follow-up sessions with practitioners. In addition, because most studies involved a multiple-baseline-across-participants design, practitioners nearly always received training in a one-to-one format. This raises questions about how feasible the training processes described in these studies would be for typical training agents, particularly for in-service practitioners. Results from survey research suggest that typical training for in-service teachers consists of a stand-alone training workshop in a group format without any follow-up training (Brock, Huber, Carter, Juarez, & Warren, 2014) and that most paraprofessionals do not receive any formal training on instructional practices (Carter, O'Rourke, Sisco, & Pelsue, 2009). This difference likely stems—at least in part—from the cost associated with providing repeated one-to-one training sessions in order to train a single practitioner in a single practice. A handful of studies in this review propose some possible solutions that might be more feasible in everyday practice, including peer coaching (e.g., Tschantz & Vail, 2000), teachers training paraprofessionals (Brock, Biggs, Carter, Cattey, & Raley, 2016), or training practitioners to self-monitor their own implementation and correct their own errors instead of depending on a trainer to repeatedly observe and provide feedback (e.g., Bingham, Spooner, & Browder, 2007).

Although repeated performance feedback might be more tenable in the context of university fieldwork supervision for preservice teachers, dedicating multiple feedback sessions to each intervention practice that a teacher needs to master would likely require more frequent and focused supervision than is feasible for university supervisors. Mentor teachers might be able to provide more frequent and focused feedback, but this model is untested and presumes mentor teachers have mastered the given instructional practice and are skilled in providing feedback.

Third, a number of training strategies were associated with practitioner implementation fidelity or student outcomes. Four training strategies were associated with increased implementation fidelity. BST was associated with the most consistent effects across practitioners (based on success estimates derived from visual analysis). This is unsurprising, given that BST is a well-established combination of promising strategies that has shown promise not only for training school-based practitioners but also for training residential staff (e.g., Lambert, Bloom, Clay, Kunnavatana, & Collins, 2014) and parents (e.g., Seiverling, Williams, Sturmey, & Hart, 2012). Modeling, written instructions (e.g., an implementation checklist), and verbal feedback were statistically significantly associated with larger magnitude of effects (based on meta-regression on the DHPS effect size). A component of BST, modeling is one of the most efficient ways to clearly communicate correct implementation of steps and has been identified as a critical feature of training in previous reviews (e.g., Brock & Carter, 2013). In this review, trainers used modeling both during initial training to promote initial implementation fidelity and in follow-up training sessions to correct errors. Although use of an implementation checklist has not been heavily emphasized in the teacher training literature, in other fields, such as medicine, implementation checklists have been designed to be the active ingredient in training practitioners to high fidelity (e.g., Mayer et al., 2016). Further, a written implementation checklist may enhance the efficacy of other training strategies; after practitioners achieve initial mastery of implementation steps, a written checklist likely helps practitioners to review and recall these steps. It is not surprising that verbal feedback promoted increased magnitude of effects, given that a number of previous reviews have concluded that performance feedback is an effective training strategy (e.g., Fallon et al., 2015; Solomon et al., 2012). However, it is somewhat surprising that only the subset of studies that used verbal feedback (and not the larger category of

performance feedback) was associated with larger effects. We suspect that might be related to inclusion of studies that used written performance feedback, which was often used as a low-intensity alternative to verbal feedback. It was often delivered less frequently (e.g., weekly by email) and did not include the type of modeling and role-play that is often paired with in-person feedback.

Two strategies associated with student effects were statistically significant. Written performance feedback was associated with decreased magnitude of student effects, and use of other written materials (i.e., written materials other than intervention instructions) was associated with increased magnitude of student effects. We attribute the negative association between written feedback and student effects to the low-intensity nature of this approach (see discussion in previous paragraph). Other written materials typically included general information about a disability (e.g., characteristics of students with autism spectrum disorder) or about the broader context for the intervention practice (e.g., principles of applied behavior analysis and systematic instruction). If practitioners lacked a firm grasp on this kind of background information, one can envision how a better understanding might facilitate more effective implementation with students and perhaps even increased motivation to implement interventions with high fidelity.

Fourth, in a number of cases, the effects were related to the intervention practice that practitioners were being trained to implement. Practitioner implementation fidelity increased most consistently for PECS, structured play groups, appropriate curricular focus, and prescribed instructional sequence (based on success estimates derived from visual analysis). Reinforcement was statistically significantly associated with smaller fidelity gains, and a prescribed instructional sequence was associated with larger fidelity gains. We offer two explanations for these associations. One explanation is that it simply is easier to train a practitioner to implement some practices than others. Indeed, it would likely be easier to train someone to focus on more appropriate curricular goals than to implement discrete trial training with high fidelity. Alternatively, due to the nature of some practices, it might be appropriate to expect a different magnitude of improvement in fidelity. For example, when researchers measured baseline practitioner performance for practices with regimented steps that were likely to differ from business as usual (e.g., PECS, prescribed instructional sequence), they typically documented extremely low baseline performance with little or no variability (i.e., zero or near-zero levels). Because these practices were often indexed as a percentage of steps implemented accurately, practitioners had the opportunity to progress from zero to 100. Even moderate implementation fidelity (e.g., 50%) represented a clear and substantial shift from baseline performance. In contrast, researchers often gauged implementation of reinforcement by measuring frequency of praise during instruction. Practitioners provided at least some praise during the baseline condition with moderate variability, and it is neither possible nor optimal to produce the same magnitude of change as observed for interventions indexed as a percentage of steps (i.e., 100 praise statements per minute). Therefore, differences in magnitude of improvement may be appropriate and expected.

Naturalistic intervention and pivotal response training were both statistically significantly associated with larger improvements of student outcomes. A commonality of these two practices is that they both leverage children's natural interests and motivation, which may enhance their effectiveness. Given the lack of association between these practices and practitioner implementation fidelity, these practices might be well designed to produce positive student outcomes even when practitioner implementation fidelity is less than optimal.

Implications for Practice

Findings from this literature review have important implications for providers of both preservice and in-service practitioner training. First, training agents should use promising strategies for improving practitioner

implementation fidelity and promoting student outcomes. Our findings suggest such strategies include BST (i.e., the combination of modeling, rehearsal, and performance feedback), modeling, verbal feedback, and written directions for implementation. Further, supplemental written material may be helpful when practitioners are unfamiliar with a disability category or the larger framework in which the practice is situated. Second, if naturalistic intervention or pivotal response training is a good match for high-priority student goals and student profiles, trainers might consider prioritizing training on these practices. Our analysis suggests that these practices in particular might be well designed to produce strong student effects even when practitioner implementation fidelity is variable. Third, providers should gauge the impact of training not only by measuring practitioner implementation fidelity but also by tracking the degree to which practitioner implementation improves student outcomes. Findings across studies in this review show that implementation fidelity does not guarantee improved student outcomes. Trainers can increase the likelihood of success by focusing on evidence-based practices that are a good match for student profiles and instructional targets, measuring student growth, and following up with practitioners to help them adjust their implementation when students do not make optimal progress. When adjustments do not promote sufficient student progress, providers should consider whether an alternative evidence-based practice might be a better choice.

Limitations and Future Directions

We identified several limitations of the literature we reviewed. First, many articles did not include precise descriptions of variables, which limited our ability to both develop more precise categories and code all variables across studies. We recommend that authors of future studies provide more explicit descriptions of how, when, where, with whom, and how long the training occurred in accordance with established standards (e.g., Kratochwill et al., 2010). A related limitation exists with measurement. For example, 10% the studies did not report IOA at an acceptable level (i.e., <80%). When measurement does not meet minimum standards, readers cannot be confident that reported findings can be attributed to the intervention. A third—potentially more significant—limitation is that relatively few studies reported student outcomes, and those that did had mixed results. It is essential that researchers measure both practitioner and student outcomes to build a strong case that practitioner training was effective. As noted earlier, accurate implementation of a practice does not guarantee student progress.

Providers should gauge the impact of training not only by measuring practitioner implementation fidelity but also by tracking the degree to which practitioner implementation improves student outcomes.

There are also limitations of our review process. First, we intentionally cast a wide net to survey the broader practitioner literature, and this wide scope leaves many more targeted research questions unanswered. In the future, researchers may wish to address more targeted questions related to individual training practices or particular types of practitioners. Second, although we carefully selected two approaches to gauging the efficacy of studies, each of these approaches has limitations. Success estimates summarize visual analysis as a dichotomous variable that reflects the consistency, but not magnitude, of effects. DHPS is a recommended approach for estimating the magnitude of effects in single-case-design studies, but this method has known limitations (e.g., it does not analyze trend; it cannot analyze certain experimental designs). In addition, given its relatively limited use, there may be other limitations that have not yet been identified. Further research is needed to mitigate existing limitations and better understand whether other limitations exist. Third, we included only published studies in this review, which may have biased our findings. In the future, researchers might

examine data from dissertations or other unpublished datasets.

Conclusion

Bridging the research-to-practice gap in special education is crucial, and improved practitioner training is one avenue for achieving this goal. In this review and meta-analysis, we identified a number of training strategies that are associated with greater consistency or magnitude of effects on practitioner implementation fidelity and student outcomes. We also found that different intervention practices may be more or less difficult to train to fidelity relative to other practices. Although these findings make a strong contribution to designing improved training opportunities, there is still much to learn about what training strategies are most effective, how to make training more feasible, and which combinations of training and practices best promote student outcomes. Further addressing these questions will not be easy and will require accumulation of evidence across research groups who are focused on optimizing practitioner training through high-quality research. Given the increased focus on implementation and publication of research on practitioner training over the past 20 years, we are optimistic the next 20 years will bring clearer answers to these and other important questions.

References

Bingham, M. A., Spooner, F., & Browder, D. (2007). Training paraeducators to promote the use of augmentative and alternative communication by students with significant disabilities. *Education and Training in Developmental Disabilities*, *42*, 339–352.

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. (2016). A meta-analysis of practitioner training to improve implementation of interventions for students with disabilities. *Remedial and Special Education*. Advance online publication. doi:10.1177/0741932516653477

Brock, M. E., Biggs, E., Carter, E. W., Cattey, G., & Raley, K. (2016). Implementation and generalization of peer support arrangements for students with significant disabilities in inclusive classrooms. *The Journal of Special Education*, *49*, 221–232. doi:10.1177/0022466915594368

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

Carter, E. W., O'Rourke, L., Sisco, L. G., & Pelsue, D. (2009). Knowledge, responsibilities, and training needs of paraprofessionals in elementary and secondary schools. *Remedial and Special Education*, *30*, 344–359. doi:10.1177/0741932508324399

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

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

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

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Hedges, L.G., Pustejovsky, J., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*, 324–341. doi:10.1002/jrsm.1086

Kratochwill, T. R., Hitchcock, J., Homer, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W.R. (2010, June). *Single-case designs technical documentation*. Retrieved

from http://files.eric.ed.gov/fulltext/ED510743.pdf

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

Lambert, J. M., Bloom, S. E., Clay, C. J., Kunnavatana, S. S., & Collins, S. D. (2014). Training residential staff and supervisors to conduct traditional functional analyses. *Research in Developmental Disabilities*, *35*, 1757–1765. doi:10.1016/j.ridd.2014.02.014

Marso, D., & Shadish, W. (2014). *User guide for DHPS, D_Power, and GPHDPwr SPSS macros (Version 1.0)*. Merced, CA: Author.

Mayer, E. K., Sevdalis, N., Rout, S., Caris, J., Russ, S., Mansell, J., . . . Moorthy, K. (2016). Surgical checklist implementation project: The impact of variable WHO checklist compliance on risk-adjusted clinical outcomes after national implementation: A longitudinal study. *Annals of Surgery*, *263*, 58–63. doi:10.1097/SLA.0000000000001185

No Child Left Behind Act of 2001, 20 U.S.C. §§ 6301 *et seq*. (2006 & Supp. V. 2011).

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*, 33–84. doi:10.3102/0034654307313793

Odom, S. L. (2009). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education*, *29*, 53–61. doi:10.1177/0271121408329171

Odom, S. L., Cox, A. W., Brock, M. E., & National Professional Development Center on Autism Spectrum Disorders. (2013). Implementation science, professional development, and autism spectrum disorders. *Exceptional Children*, *79*, 233–251. doi:10.1177/001440291307900207

Reichow, B., & Volkmar, F. R. (2010). Social skills interventions for individuals with autism: Evaluation for evidence-based practices within a best evidence synthesis framework. *Journal of Autism and Developmental Disorders*, *40*, 149–166. doi:10.1007/s10803-009

Rispoli, M., Neely, L., Lang, R., & Ganz, J. (2011). Training paraprofessionals to implement interventions for people autism spectrum disorders: A systematic review. *Developmental Neurorehabilitation*, *14*, 378–388. doi:10.3109/17518423.2011.620577

Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education*, *27*, 396–407. doi:10.1177/088840640402700407

Seiverling, L., Williams, K., Sturmey, P., & Hart, S. (2012). Effects of behavioral skills training on parental treatment of children's food selectivity. *Journal of Applied Behavior Analysis*, *45*, 197–203. doi:10.1901/jaba.2012.45-197

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.

Tschantz, J. M., & Vail, C. O. (2000). Effects of peer coaching on the rate of responsive teacher statements during a child-directed period in an inclusive preschool setting. *Teacher Education and Special Education*, *23*, 189–201. doi:10.1177/088840640002300302

What Works Clearinghouse. (2014). *Procedures and standards handbook (Version 3.0)*. Washington, DC: Author.

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., Brock, M. E., . . . 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

Authors' Note

Support for this research came from the Institute of Educational Sciences, U.S. Department of Education, through Grant R324B160009 to The Ohio State University. We would like to thank Dr. Paula Chan and Dr. Megan Miller for their help with initial analysis of literature. In addition, we would like to thank David Marso for providing technical assistance with the DHPS macro for SPSS. We also would like to thank Dr. Samuel Odom for his feedback and mentorship.

Manuscript received May 2016; accepted February 2017.