



Preschool Interventions for Children with Autism Spectrum Disorder: a Review of Effectiveness Studies

Jessica Tupou¹ · Larah van der Meer¹ · Hannah Waddington¹ · Jeff Sigafoos¹

Received: 8 January 2019 / Accepted: 9 April 2019 / Published online: 6 May 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Early intervention is widely recommended for young children with autism spectrum disorder, but the effectiveness of interventions when implemented in inclusive preschool settings needs to be considered. To this end, we identified 16 studies that provided early intervention to children with autism spectrum disorder who were attending inclusive preschools. Effectiveness studies, that is, studies evaluating interventions that were implemented under real-world conditions, were targeted, provided they met the pre-determined inclusion criteria. Included studies were summarized in terms of (a) participant characteristics, (b) intervention procedures, (c) procedures for training teachers, (d) study design, (e) outcomes, (f) study quality, and (g) social validity. A range of intervention procedures were implemented, and positive outcomes for children and/or teachers were reported in most studies. However, methodological limitations were identified for most of the included studies, indicating the need for additional higher-quality research. Still, the present review suggests that teachers may be able to effectively deliver early intervention to young children with autism spectrum disorder in inclusive preschool settings.

Keywords Autism spectrum disorder · Early intervention · Teacher-implemented intervention · Preschool · Effectiveness research · Systematic review

Autism spectrum disorder (ASD) is a type of developmental disability characterized by deficits in social interaction and communication skills and by the presence of repetitive/ritualistic behaviors and a restricted range of interests (American Psychiatric Association 2013). Although ASD was once considered to be a relatively rare condition (Simpson 2004), its prevalence is now estimated at up to 1 in every 59 individuals (Baio et al. 2018). With increasing numbers of children being diagnosed with ASD, there is a corresponding need for provision of effective intervention (Woods and Wetherby 2003).

Various intervention approaches have been developed and evaluated to assess their efficacy and/or effectiveness for promoting improved adaptive behavior functioning and reducing ASD symptoms. Efficacy research refers to the evaluation of intervention effects when the intervention is conducted under

ideal or controlled conditions, such as when delivered by researchers in a clinical setting. In contrast, effectiveness research refers to the evaluation of intervention effects when the intervention is conducted under real-world conditions, such as when delivered in preschool settings by the usual teaching personnel (Singal et al. 2014).

The range of interventions that has been evaluated in efficacy or effectiveness research includes pharmacological agents, dietary interventions, occupational and speech-language therapies, interventions based on the principles of applied behavior analysis, and developmentally orientated interventions (Ospina et al. 2008). In addition, hybrid intervention approaches have been evaluated, such as early intervention programs that make use of behavior analytic instructional tactics (e.g., reinforcement, response prompting) within naturalistic and developmentally appropriate social interactions and play routines (Debodinance et al. 2017; Odom et al. 2010). Interventions that fall into this latter, behavioral-developmental, category appear to be the most widely used in contemporary practice, perhaps due to the growing evidence base supporting their efficacy (Ospina et al. 2008).

Positive outcomes across a range of domains of functioning (e.g., social, play, cognitive, and communication

✉ Jessica Tupou
Jessica.Tupou@vuw.ac.nz

¹ School of Education, Victoria University of Wellington, Wellington 6147, New Zealand

skills) from such behavioral/developmentally oriented early intervention programs have been documented in a number of studies (e.g., Dawson 2008; Debodinance et al. 2017; Drahota et al. 2012; Estes et al. 2015; Keenan and Dillenburger 2011; Reichow 2012). However, in most existing studies, the treatment procedures were implemented by researchers or therapists in a one-on-one format or within specialist small-group arrangements, rather than by usual personnel (e.g., teachers) in more naturalistic/inclusive group settings (Young et al. 2016). That is, most research to date has examined efficacy rather than effectiveness. To advance evidence-based practice with respect to the implementation of early intervention programs, it would seem critical to evaluate outcomes from effectiveness studies, that is, studies in which the treatment procedures were implemented under more natural, real-world conditions, such as in inclusive preschool settings with teachers serving as the intervention agents.

An inclusive preschool-based approach to the delivery of interventions would seem to offer several potential advantages over delivery of intervention by experts in more structured/clinical settings. First, international guidelines recommend delivery of interventions in settings that (a) provide ongoing opportunities for interaction with typically developing peers and (b) are the least restrictive environment for meeting the individual's needs (Broderick 2017; United Nations 2006). Inclusive preschools are perhaps more likely to meet these requirements than clinical settings as they offer the opportunity for children with ASD to be part of the same learning environment as their typically developing peers and to practice the social behaviors needed to interact with these peers (Vivanti et al. 2017). Opportunities for interaction with peers would seem especially important for young children with ASD because of the associated social communication impairments, which are likely to interfere with forming positive peer relationships. The presence of typically developing peers in an inclusive preschool setting might be useful for promoting more positive peer relations and offering models of age-appropriate play, communication, and social behavior (Koegel et al. 2001).

Providing early intervention in an inclusive preschool setting would also seem to offer some potential cost saving, as intervention could be delivered to more than one child at a time using existing resources (e.g., existing teaching staff, physical spaces, and equipment). Additionally, there is a documented disparity in children's access to early intervention services (Thomas et al. 2007) and healthcare (Kramer et al. 2017). Specifically, ethnicity, parental education, and geographic location may all impact on families' ability to access ASD-related services (Thomas et al. 2007). Addressing this disparity has been cited as an important and understudied area of research in

the area of ASD intervention (Interagency Autism Coordinating Committee (IACC) 2013). Delivering intervention in an inclusive preschool environment might help to address this disparity to some extent, particularly if the young child with ASD is already attending an inclusive preschool where early intervention could be provided.

Existing teaching staff would seem to be the most logical intervention agents for interventions conducted in inclusive preschool settings because they are likely to have knowledge of early development and learning as well as familiarity with the unique needs and interests of the children they teach (Lawton and Kasari 2012). It is important to note, however, that some research suggests that typical pre-service training may not equip teachers with the knowledge, skills, and/or levels of confidence required to meet the needs of children with ASD (Mitchell and Hegde 2007). Thus, it seems important to examine how teaching staff in inclusive preschools might be trained to meet the needs of the young children with ASD that they teach.

Given the increasing need for, and potential positive outcomes from, early intervention programs for young children with ASD—and the potential advantages of inclusive preschool-based delivery of interventions by teaching staff—it seems important to explore whether such interventions can be delivered in this type of setting and whether doing so is likely to produce positive outcomes for the child. In this review, we sought to identify studies involving the provision of early intervention to children with ASD who were attending inclusive preschool settings. We also sought to appraise the quality of the identified studies and evaluate their effects on child outcomes. The strategies used in training teaching staff to implement these interventions with fidelity were a particular focus of the review as well. The specific questions addressed in this review are:

1. What types of intervention procedures/programs have been used in the included studies?
2. What are the design characteristics and rigor of the included studies?
3. What were the characteristics of the participants involved in the studies?
4. How were teaching staff trained to implement the intervention procedures and to what extent did this training enable the staff to implement the interventions with fidelity?
5. What were the range of outcomes for the children with ASD who received intervention?
6. To what extent were the interventions/staff training procedures perceived by stakeholders to be effective and acceptable (i.e., socially valid)?

Method

Search and Screening Procedures

Searches were carried out by the first author using the PsycINFO, ERIC, Scopus, PubMed, and ProQuest databases. For all databases, the following search terms were entered into the “Anywhere” field: *Autis** OR *ASD* AND “Teacher led” OR “teacher implemented” AND *Intervention* OR *program** OR *treatment* AND “Early intervention” OR *preschool* OR “early childhood.” Results were limited to journal articles published in English between 2000 and 2017. These initial electronic searches returned 351 articles.

The titles and abstracts for these 351 articles were then reviewed to screen them for their potential eligibility for inclusion. At this stage, 16 articles were deemed eligible for inclusion, and consequently, the full text of each of these articles was reviewed to ensure that each study met all of the inclusion criteria. Ten articles met all of the inclusion criteria. An ancestral search of the reference lists of the included articles from the database search produced a further four articles for inclusion. Finally, author searches were conducted on the authors of the included articles from the database and ancestral searches. These author searches produced an additional five articles for inclusion. At this stage, three articles were excluded because they involved both inclusive and non-inclusive preschool settings and the data from non-inclusive preschools could not be separated from the data from the inclusive preschools. In total, 16 articles met the inclusion criteria.

Inclusion Criteria

To be included in the review, studies needed to meet several criteria. First, the study had to have evaluated outcomes from interventions that were conducted in inclusive preschool settings. An inclusive preschool setting was defined as an educational and/or care setting for preschool-aged children (typically aged 1 to 6 years) that included both children with and without disabilities. In addition, the interventions had to have been implemented by the staff who regularly worked in that setting. Staff could include teachers, paraeducators, teaching assistants, or similar personnel. The study also had to include at least one child participant who was (a) aged between 12 and 72 months or attending the preschool and (b) had a diagnosis of ASD, Autism, Asperger’s Syndrome, or Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). If a study also included one or more participants who did not meet these inclusion criteria, only data from the participants who were eligible for inclusion was extracted and analyzed. If data from eligible participants could not be separated out, the study was excluded. Studies set in special education classrooms and studies focused on interventions implemented by parents, researchers, or specialists that were not

regular preschool staff were excluded as were case studies, studies with qualitative designs, and theses or dissertations.

Data Extraction

The following data were extracted from each included study: (a) type of study design, (b) child characteristics (number, age, and diagnoses), (c) intervention characteristics (intervention type, frequency, and duration), (d) teacher training (method, frequency, and duration), (e) research quality/rigor, (f) child outcomes (type of outcome, method of measurement, and results), (g) teacher outcomes/behavior (type of outcome/behavior, method of measurement, and results), and (h) social validity (method of measurement and results).

Results were classified as positive, mixed, or no effect/negative. For studies with a single-case design, a positive result was coded when positive changes for all primary dependent variables (DV), all participants, and all intervention phases were reported. Mixed results referred to situations where authors reported minimal or no improvement or highly variable results for one or more participants, primary DVs, and/or intervention phases. Finally, a coding of no effect or a negative result meant that the intervention was not associated with any positive results for any of the participants or for any of the primary DVs. For studies using group designs, results for each DV were reported separately. A positive result was coded when authors reported significant improvements for the experimental group (EG) for a given DV. Where control group (CG) data were reported, improvements observed in the EG needed to be significantly better than those observed in the CG for the result to be coded positive. Conversely, a code of no effect or negative result was used when no significant improvements were reported for the EG and/or reported improvements for the EG were not significantly better than those reported for the CG.

A modified version of Goldstein et al.’s (2014) framework was used to evaluate each included study for its quality/rigor. The quality criteria used in this framework are consistent with those proposed by Cook et al. (2009) and the School Psychology Division of APA (Kratochwill and Stoiber 2002). This framework was chosen because it is suitable for use with both single-case and group-design studies and allows for a quantification of the quality/rigor of a study across a comprehensive set of specific quality indicators (Snyder et al. 2015). This framework was also selected because it allows readers to graphically view the strengths and weaknesses of studies and thus provides the ability for readers to quickly assess studies (individually and collectively) across the variables that they are most interested in.

Using this framework, included studies were evaluated across four broad areas: (a) design characteristics and internal reliability, (b) measurement features, (c) general characteristics, and (d) results and external validity. Thirteen quality

criteria, across the four aforementioned areas, were used to evaluate all studies (including group designs and single-case research designs). These 13 criteria are (a) design characteristics, (b) measurement, (c) reliability, (d) intervention fidelity, (e) training fidelity, (f) rationale, (g) robust treatment effects, (h) statistics, (i) maintenance and generalization, (j) implementation site, (k) participant selection, (l) consumer satisfaction, and (m) social validity. For single-case research designs, two additional criteria (quality of baseline and visual analysis) were used. Definitions and guidelines for assigning ratings for each criterion are included in Goldstein et al.'s framework.

The procedures that we used for evaluating the quality of the studies involved having the first author examine each article and assign a rating of between 1 and 4 for each applicable category. A score of 4 represented exemplary performance, 3 represented acceptable performance, 2 represented minimal performance, and a score of 1 represented unacceptable performance. A total mean score was then calculated for each study by dividing the sum of the study's scores for each category by the number of categories scored.

Interrater Agreement

A second reviewer independently reviewed the full text of all articles identified during the database, ancestral, and author searches to check on their eligibility against the inclusion criteria. Before undertaking any agreement checks, the primary rater (first author) and the independent rater (second author) discussed the inclusion criteria and the types of data to be extracted from each article. Interrater agreement on whether or not each identified article did or did not meet the inclusion criterion was calculated by dividing the number of agreements by the number of agreements plus disagreements and multiplying by 100. Overall agreement for all searches was 88.3% (range = 80 to 100%). Agreement on the accuracy of data extraction was also assessed for all included studies and for all variables. Agreement ranged from 96 to 99% with a mean of 97%. Disagreements were discussed to obtain consensus.

Results

The 16 included studies are summarized in Tables 1, 2, 3, and 4. Table 1 summarizes design, child characteristics, and intervention characteristics. Table 2 summarizes child outcomes, Table 3 summarizes teacher outcomes and social validity, and Table 4 provides an evaluation of each study's design characteristics and the presence of specific quality indicators.

Child Characteristics

A total of 809 children participated across the 16 included studies. We classified children in terms of the diagnoses they

had been assigned in the original research reports. Of these participating children, 734 (91%) had a diagnosis of autism/ASD, 25 (3%) had a diagnosis of PDD-NOS, 1 child (< 1%) had a diagnosis of Asperger's Syndrome, and 1 (< 1%) child had no formal diagnosis, but was reported to have displayed autism-like symptoms. A further 48 children (6%) from one study (Schwartz et al. 2004) had a diagnosis of either ASD or PDD-NOS; however, the authors did not specify which diagnosis each child had. The mean age across studies was 45.9 months. This mean does not include the participants from Schwartz et al. (2004) because these authors only provided the range of participants' ages, not the mean. Early intervention was provided to 517 (64%) of the participating children with the remaining 292 children (36%) assigned to CGs that received treatment as usual.

Setting

The early interventions being evaluated in these studies were implemented in the regular classroom environments within inclusive preschool settings for all 16 included studies. However, in one study, part of the intervention was delivered on a one-on-one basis in a designated treatment room at the preschool (Eikeseth et al. 2012). Twelve studies (75%) took place in US-based preschools (Boulware et al. 2006; Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Harjusola-Webb and Robbins 2012; Kern et al. 2007; McBride and Schwartz 2003; Olive et al. 2007; Schwartz et al. 2004; Strain and Bovey 2011; Van DerHeyden et al. 2002; Young et al. 2016). The remaining four studies (25%) were conducted in Italy (D'Elia et al. 2014), Norway (Eldevik et al. 2012), Germany (Kern and Aldridge 2006), and Sweden (Eikeseth et al. 2012).

Intervention Approaches

Various intervention approaches were used across the 16 included studies. Seven studies (44%) delivered some type of comprehensive intervention that targeted a range of developmental areas (Boulware et al. 2006; D'Elia et al. 2014; Eikeseth et al. 2012; Eldevik et al. 2012; Schwartz et al. 2004; Strain and Bovey 2011; Young et al. 2016). These seven studies involved the use of one of six different comprehensive intervention programs, specifically: (a) Developmentally Appropriate Treatment for Autism (DATA; Boulware et al. 2006; Schwartz et al. 2004), (b) Treatment and Education of Autistic and Related Communication Handicapped Children (TEACCH; D'Elia et al. 2014), (c) Early Intensive Behavioral Intervention (EIBI; Eldevik et al. 2012), (d) Learning Experiences and Alternative Program for Preschoolers (LEAP; Strain and Bovey 2011), (e) Comprehensive Autism Program (CAP; Young et al. 2016), and (f) an EIBI intervention described as being based on Lovaas' UCLA model

Table 1 Summary of child characteristics, intervention characteristics, and teacher training for included studies

Study	Design	Child characteristics			Intervention characteristics		Teacher training	
		<i>n</i>	Age (months)	Diagnosis	Type	Frequency (sessions per week), intensity (hours per week), and duration	Method	Frequency (sessions per week), intensity (hours per week), and duration
1. Boulware et al. (2006)	Single-group pre- and post-tests	4	<i>M</i> = 24.8	ASD	Project DATA for toddlers.	Integrated playgroup: 2 (frequency) × 3 (intensity); individualized instruction: 3 × 6; family/home support: NS × 7 for 9–21 months (<i>M</i> = 13.5)	NS	NS
2. D’Elia et al. (2014)	Nonequivalent comparison-group	30 (CG = 15, EG = 15)	CG: <i>M</i> = 49 EG: <i>M</i> = 49	AD (<i>n</i> = 15) PDD-NOS (<i>n</i> = 15)	EG: low intensity TEACCH program. CG: speech and psychomotor therapy	Frequency: NS Intensity: EG: 2 at home and 2 at school. CG: 2 of each type at school. Duration: 24 months	NS	NS
3. Eikeseth et al. (2012)	Nonequivalent comparison-group	59 (CG = 24, EG = 35)	CG: <i>M</i> = 53 EG: <i>M</i> = 47	Autism	EG: treatment based on Lovaas’ UCLA model CG: TAU	Frequency: NS Intensity: EG: 15–37 (<i>M</i> = 23). Additional hours each week (number NS) in the child’s home. Duration: 24 months	NS	NS
4. Eldevik et al. (2012)	Nonequivalent comparison-group	43 (CG = 12, EG = 31)	<i>M</i> = 44	ASD (<i>n</i> = 34) PDD-NOS (<i>n</i> = 8) Asperger (<i>n</i> = 1)	EG: EIBI program. Delivered via 1:1 discrete trial format, shifting to small-group settings with an eventual goal of naturalistic teaching. CG: TAU	EG: 5 × 13.6; CG: NS (but at least 5 h/week) for <i>M</i> = 25 months	Apprenticeship model using modeling, mentoring, and coaching	Consultations: 1 × 1–4; team meetings: 1 × 2 for the duration of intervention (<i>M</i> = 25 months). Plus one-off 3-day workshop
5. Fleury and Schwartz (2017)	Multiple-baseline across groups	9	<i>M</i> = 51	ASD	Dialogic reading techniques	4 × 0.55 for 1.2 months	Initial group training session involving video and live modeling of procedures, practice with other trainees, and feedback. Individual feedback (oral and written) and coaching sessions	Exact frequency, intensity, and duration NS. Initial training: 2 h; feedback and coaching sessions: ≥ 3 sessions
6. Garfinkle and Schwartz (2002)	Multiple-baseline across participants	4	<i>M</i> = 53	ASD (<i>n</i> = 3) no formal diagnosis but display of ASD-like symptoms (<i>n</i> = 1).	Small-group peer imitation intervention	5 × 0.83 Duration: NS	Details of initial training NS but fidelity was monitored throughout the implementation phase and didactic coaching provided as necessary	NS

Table 1 (continued)

Study	Design	Child characteristics		Intervention characteristics		Teacher training		
		<i>n</i>	Age (months)	Diagnosis	Type	Frequency (sessions per week), intensity (hours per week), and duration	Method	Frequency (sessions per week), intensity (hours per week), and duration
7. Gibson et al. (2010)	Single-case ABAB	1	48	Autism	Functional Communication Treatment (FCT)	Frequency: NS Intensity: weekly NS but 1 × 15-min session per day for 12 sessions (monthly NS)	Training was delivered by a consultant via videoconferencing software and included written instruction, modeling, role play, descriptive praise, and corrective feedback.	One-off session continued until teacher and teaching assistant reached 100% fidelity and were confident using procedures (45 min).
8. Harjusola-Webb and Robbins (2012)	Multiple-baseline across participants	3	<i>M</i> = 36	ASD	Naturalistic communication--promoting strategies package based on PRT and EMT	Frequency: NS Intensity: teachers were instructed to embed the intervention throughout the entire daily routine (participants were enrolled at the preschool for 18 h/week). Duration: NS but data collected for 6-months	Intervention strategy manual and regular standardized meetings with researchers	Frequency: 1 × 0.17–0.67 for the entire intervention (<i>M</i> = NS)
9. Kern and Aldridge (2006)	Multiple-baseline across participants	4	<i>M</i> = 47.5	Autism	Embedded music therapy	NS but participants were observed for 10 min/day for 30 (Eric), 56 (Ben), 63 (Phillip), or 71 (Lucas) sessions. Duration: 8 months	Didactic training with the study's author	NS; however, staff training ended when teaching staff indicated that they were comfortable with both the song and the intervention procedure.
10. Kern et al. (2007)	Single-subject withdrawal design	2	<i>M</i> = 39.5	Autism	Music therapy.	Frequency: NS Intensity: weekly NS but 2–10 min/day. Duration: <i>M</i> = 2.5 months	Initial consultation to establish goals and procedures. Direct modeling of songs and interaction with participants	Frequency: NS intensity: NS duration: 0.5 months
11. McBride and Schwartz (2003)	Multiple-probe across participants	2	<i>M</i> = 56	Autism (<i>n</i> = 1) PDD-NOS (<i>n</i> = 1)	Activity-Based Intervention (ABI) and ABI with discrete trials	Frequency: NS Intensity: weekly NS but 30–35 min per session Duration: <i>M</i> = 14.5 sessions (monthly NS)	Initial planning meeting, didactic training session, hands-on training session, and verbal feedback from researcher after each observation	Exact frequency, intensity, and duration NS. Length of initial planning meeting NS; didactic training session: 45 min; Hands-on training session: 30 min. Feedback after observations continued until teachers reached 80% fidelity for 3 consecutive sessions

Table 1 (continued)

Study	Design	Child characteristics		Intervention characteristics		Teacher training		
		<i>n</i>	Age (months)	Diagnosis	Type	Frequency (sessions per week), intensity (hours per week), and duration	Method	Frequency (sessions per week), intensity (hours per week), and duration
12. Olive et al. (2007)	Multiple-probe across participants.	3	53	Autism (<i>n</i> = 2), PDD-NOS (<i>n</i> = 1)	Enhanced Milieu Teaching combined with a voice output communication aid	4 × 0.33	Graduate course on communication intervention	NS.
13. Schwartz et al. (2004)	Single-group pre- and post-tests	48	All aged 36–72 months.	Autism PDD-NOS (<i>n</i> = NS)	Project DATA	Frequency: NS Intensity: 20 Duration: <i>M</i> = 16 months	NS	NS
14. Strain and Bovey (2011)	Clustered randomized design	294 (C = 117, EG = 177)	50.4	ASD	EG: LEAP full training. CG: LEAP manuals, and training materials (e.g., Power Point slides) only	Frequency: integrated into the daily preschool program. Number of sessions per week varied across participants. Intensity: <i>M</i> = 17; Duration: 2 years	Seven-phase educational model including written instruction, discussion, demonstration, in-vivo practice, feedback, evaluation, and follow-up	Frequency: NS intensity: exact hours per week NS but 23 days total. Duration: 24 months
15. Van DerHeyden et al. (2002)	Reversal design (A-B-A-C-A-C)	1	29	ASD	Within-and extra-stimulus prompts to increase targeted toy play behaviors	2 × 1 day (number of hours NS). Duration: NS; however, data were collected over 36 sessions.	NS	NS
16. Young et al. (2016)	Randomized controlled trial.	302 (CG = 124, EG = 178)	<i>M</i> = 52.8	ASD	EG: CAP CG: TAU	Exact frequency and intensity NS but integrated throughout the day as well as 3 × 15 min 1:1 sessions per child per day, children attended school approx. 16 h/week. Duration: one school year	Workshops and collaborative on-site coaching.	Initial workshop: 10 h across 2 days; onsite training: 10 h across 2 days; subsequent onsite coaching (<i>M</i> = 27 h per school). Duration: NS

NS, not specified; DATA, = Developmentally Appropriate Treatment for Autism; TEACCH, Treatment and Education of Autistic and Related Communication Handicapped Children; EIBI, Early Intensive Behavioral Intervention; TAU, treatment as usual; LEAP, Learning Experiences and Alternative Program for Preschoolers and their Parents; PRT, Pivotal Response Training; EMT, Enhanced Milieu Teaching; CAP, Comprehensive Autism Program

Table 2 Summary of child outcomes

Study	Outcome:	Measurement:	Findings:
1. Boulware et al. (2006)	Developmental functioning Behavior Communication, social, and symbolic abilities. Functional outcomes (e.g., following instructions, toilet training, and motor imitation) Autism severity Adaptive behavior	BSID TABS CSBS Specifically developed functional outcome scale ADOS VABS	BSID-II scores increased for all participants ($M = 28$ points). 3 participants improved their overall category rating. 3 participants demonstrated improvement on the TABS, improving from “atypical” ($n = 2$) or “at risk” ($n = 1$). All participants demonstrated gains in at least 1 of the 7 communication clusters, 3 participants made significant gains ($> 2SD$) on at least one cluster. 2 participants demonstrated gains across all 7 clusters. All participants demonstrated gains in at least one functional behavior. Decrease in “autism” diagnoses across both groups, larger decrease in EG Significant improvements in adaptive behavior across both groups, slightly higher increase in EG Significant gains in language skills across both groups with higher gains observed in EG Significant difference between EG and CG across a number of subscales, with both groups showing decrease in score across all subscales Significant improvement over time in all categories except receptive language and personal self-care After 1 year of treatment, EG showed significantly greater improvements on all VABS sub-domains (M effect size = 0.84) and within-group decreases on maladaptive behavior for the EG were significant ($ES = 1.0$). Changes in VABS scores during the 2nd year of treatment were not significant.
2. D’Elia et al. (2014)	Autism symptom severity	CARS	For EG, CARS scores decreased significantly for both the 1st ($ES = 0.92$) and 2nd years of treatment. No CARS data was presented for the CG EG made significantly larger gains on composite scores for intelligence.
3. Eikeseth et al. (2012)	Intellectual functioning Adaptive behavior	BSID for participants aged <42 months, Stanford-Binet Intelligence Scale for older participants VABS	EG made significantly larger gains on adaptive behaviour composite score and communication and socialization sub-domains. The rate of response to questions improved for all participants immediately upon introduction of the intervention ($M ES = 1.02$). Changes in rate of initiations between baseline and intervention were minimal for 8 participants. For the 9th participant (Sam), rate of initiations decreased upon introduction of the intervention.
4. Eldevik et al. (2012)	Autism symptom severity	CARS	For EG, CARS scores decreased significantly for both the 1st ($ES = 0.92$) and 2nd years of treatment. No CARS data was presented for the CG EG made significantly larger gains on composite scores for intelligence.
5. Fleury and Schwartz (2017)	Verbal participation (response to adults and initiating comments/questions)	Video recordings of reading sessions (2 week)—response rate	EG made significantly larger gains on adaptive behaviour composite score and communication and socialization sub-domains. The rate of response to questions improved for all participants immediately upon introduction of the intervention ($M ES = 1.02$). Changes in rate of initiations between baseline and intervention were minimal for 8 participants. For the 9th participant (Sam), rate of initiations decreased upon introduction of the intervention.

Table 2 (continued)

Study	Outcome:	Measurement:	Findings:
6. Garfinkle and Schwartz (2002)	Book-specific vocabulary Primary: social initiations, frequency, and types of verbal responses (positive/negative/no response), independent and prompted peer initiations, nonsocial engagement, physical proximity to peers during play Collateral: initiations of target child by typically developing peers in the generalization setting Frequency of elopement	Researcher-delivered assessment task using illustrations from known books to probe for target vocabulary. Direct observation using live interval recording with 5-min intervals	Book-specific vocabulary increased for all participants during intervention phase (M increase = 1.2 words per week). Minimal increase in social interactions for all participants. Small increase in number of imitations of peer behavior although the number observed were highly variable.
7. Gibson et al. (2010)	Frequency of elopement	Direct observation	Number of imitations increased for all participants.
8. Harjusola-Webb and Robbins (2012)	Expressive communication (including gestures, vocalizations, word approximations, words, and multi-word utterances) Positive peer interactions, play, and engagement with material and equipment	Daily observation via video conferencing software, using 20-s partial interval recording Direct observation for 30 min/day, 2 days/week	The mean percentage of intervals with elopement decreased from baseline phase 1 ($M = 96%$) to intervention phase 1 ($M = 11%$) and from baseline phase 2 ($M = 93%$) to intervention phase 2 ($M = 5%$). Mean frequency of expressive language output increased for all participants.
9. Kern and Aldridge (2006)	Number of steps of the morning arrival routine completed independently and number of unprompted greetings from peers	Video recordings 10 min using 15-s momentary time-sampling recording procedure	No/minimal increases in peer interactions observed during playgroup adaptation intervention phase; however, increases were observed for all participants during teacher-mediated and peer-mediated intervention phases.
10. Kern et al. (2007)	Verbal responses (correct/no response/incorrect), engagement, and proximity to peers	Direct observation (event recording)	Both participants increased the number of steps completed independently during the intervention, improvements declined during withdrawal phase and improved again during the second intervention phase.
11. McBride and Schwartz (2003)	Prompted/independent use of voice output communication aids (VOCA), use of independent communication acts	Video observation using 10-s momentary time sampling	Both children showed increases in mean response to target objectives from baseline to ABI-only and from ABI-only to ABI with discrete trial phase. Results for child engagement and proximity to peers were mixed.
12. Olive et al. (2007)	Prompted/independent use of voice output communication aids (VOCA), use of independent communication acts	Direct observation (event recording) of 5-min child-teacher interactions at the child's preschool	Prompted and independent VOCA use increased for all participants from 0 in baseline to an average of 10.2 times per 5-min session during intervention (range = 7.3–12.8). Total independent spontaneous communicative acts increased from baseline ($M = 0.6$ times per 5-min session) to intervention ($M = 20$ times per 5-min session).

Table 2 (continued)

Study	Outcome:	Measurement:	Findings:
13. Schwartz et al. (2004)	Child developmental progress in functional skills	AEPS	Children made gains across all AEPS developmental domains ($M = 20.5$ percentage points) over the course of 1 school year.
14. Strain and Bovey (2011)	Autism symptom severity	Researcher-designed Functional Outcomes Index CARS	Children also made gains across all 6 functional outcomes on the FOI ($M = 24.8$ percentage points). Participants from EG showed a greater decrease in CARS scores ($M = 6.1$ -point decrease) compared to the CG ($M = 2.8$ -point decrease).
	Cognitive development	Mullen Scales of Early Learning.	EG Mullen scores were significantly higher than CG scores after intervention.
	Language	PLS-4	EG PLS-4 scores were significantly higher than CG scores after intervention; differences were significant at the 5% level.
	Social skills and problem behavior	SSRS	EG showed greater increases on the SSRS (EG: $M = 28.6$ percentile points; CG: $M = 11.9$ percentile points). EG also scored lower on problem behavior ratings than the CG.
15. VanDerHeyden et al. (2002)	Contact with “low contact” activity centers, planted stimulus, toy play, speech, and disruptive behavior	Direct observation (partial-interval occurrence and responses per minute)	Contact with “low-contact” activity centers and the planted stimulus increased for all 3 intervention phases. Disruptive behavior was variable until the final intervention phase where it decreased to 0. No speech data was reported.
16. Young et al. (2016)	Autism severity. Adaptive behavior Cognitive development Expressive language Receptive language Social skills	CARS VABS BDI-2—cognitive domain EOWPVT ROWPVT ASIEP-3 and SSRS	Intervention had a positive impact on receptive language (effect size = 0.13) and social skills (effect size = 0.19). No impact was identified on the remaining measures.

ADOS, Autism Diagnostic Observation Schedule; AEPS, Assessment, Evaluation, and Programming System for Infants and Children; ASIEP-3, Autism Screening Instrument for Educational Planning-3rd ed.; BDI-2, Battelle Developmental Inventory (2nd ed.); BSI-D, Bayley Scales of Infant Development; CARS, Childhood Autism Rating Scale; CBC, Child Behavior Checklist; CDI, MacArthur Communication Development Inventories; CSBS, Communication and Symbolic Behavior Scales; EOWPVT, Expressive One Word Picture Vocabulary Test; PEP-3, Psychoeducational Profile (3rd ed.); PLS-4, Preschool Language Scale (4th ed.); ROWPVT, Receptive One Word Picture Vocabulary Test; SSRS, Social Skills Rating System; TABS, Temperament and Atypical Behavior Scale; VABS, Vineland Adaptive Behavior Scales

Table 3 Summary of teacher and social validity outcomes for included studies

Teacher behavior/outcomes				Social validity	
Study	Outcome:	Measurement:	Findings:	Measurement:	Findings:
1. Boulware et al. (2006)	None	n/a	n/a	None	n/a
2. D'Elia et al. (2014)	Implementation fidelity	Review of each student's individual education plan (IEP)	NS	None	n/a
3. Eikeseth et al. (2012)	None	n/a	n/a	None	n/a
4. Eldevik et al. (2012)	None	n/a	n/a	None	n/a
5. Fleury and Schwartz (2017)	Implementation fidelity	Direct observation at least 3 times per educator across the intervention phase. Observers used a checklist developed specifically for the study by the researchers. Method for measuring rate of dialogic prompt use was NS.	Mean fidelity = 95% (range = 76 – 100%).	Researcher-developed social validity questionnaire consisting of 11 items rated on a 7-point Likert-type scale and several open-ended questions	Respondents reported that children's vocabularies and levels of verbal participation improved as a result of the intervention (mode = 6), and that the intervention was enjoyable (mode = 6) and easy to use (mode = 7). However, the limited range of books available for the intervention and some children's challenging behavior were commonly cited challenges.
6. Garfinkle and Schwartz (2002)	Implementation fidelity	Direct observation	For 3 participants, fidelity was high throughout the study; for the remaining participants, initial extra coaching and guidance were required to achieve fidelity.	Social validity questionnaire for teachers consisting of 5 questions with 5-point Likert-type scale answer ratings and 4 open-ended questions.	Intervention was rated favorably across all five questions ($M = 4.7, 4.7, 4.7, 5.0, 4.5$)
7. Gibson et al. (2010)	Implementation fidelity	Video observation (33% of sessions in each phase)	During baseline conditions fidelity was 100%. During intervention conditions mean fidelity was 90% (range = 87–100%)	Teacher satisfaction was measured using the Behavior Intervention Rating System-Revised (BIRS-R). A specifically developed open-ended questionnaire was used to gather qualitative data on teacher' perceptions of the videoconferencing technology.	Scores on the BIRS-R ranged from 5.0–6.0 out of 6 across all 24 items ($M = 5.7$). Results of the open-ended questionnaire suggested that the videoconferencing software was viewed favorably by the teacher and the teaching assistant.
8. Harjusola-Webb and Robbins (2012)	Use of communication-promoting strategies	Direct observation for 30 min/day, 2 days/week	Increased frequency of communication-promoting strategy use for all teachers during intervention phase	None	n/a
9. Kern and Aldridge (2006)	Implementation fidelity	Direct observation during teacher-mediated intervention phase	Fidelity was very high for 3 of the 4 teachers. For teacher 4, fidelity was < 50% for the second half of condition C.	None	n/a
10. Kern et al. (2007)	Use of prompting	Direct observation event recording	No data reported	None	n/a

Table 3 (continued)

Teacher behavior/outcomes		Social validity	
Study	Outcome:	Measurement:	Findings:
11. McBride and Schwartz (2003)	Primary teacher Behavior: rate of instruction	Video observation: frequency count (rate) of instructional trials presented	Teachers 2 and 3, both, showed an increase in rate of instruction from baseline to ABI-only and a further increase from ABI-only to ABI with discrete trial. However, for teacher 3, the increase from baseline to ABI-only was only minimal.
	Collateral behaviors: obligatory and nonobligatory interactions physical prompts	Interval recording (10-s intervals)	For both teachers, obligatory and nonobligatory interactions increased in ABI-only and/or ABI with discrete trial phases. Teacher 3's use of physical prompts decreased by 38 percentage points from baseline to ABI-only and a further 5 percentage points from ABI-only to ABI with discrete trials. Teacher 2's use of physical prompts remained fairly stable across phases $M = 96.4\%$
12. Olive et al. (2007)	Implementation fidelity	Direct observation by an independent observer using an intervention competency checklist	None
13. Schwartz et al. (2004)	None	n/a	n/a
14. Strain and Bovey (2011)	Implementation fidelity	Direct observation using QPI procedural rating observational protocol	Anecdotal evidence and quotes collected from annual family interviews and letters sent by families Teacher questionnaire covering 14 dimensions social validity using 5-point Likert-type rating scales
15. VanDerHeyden et al. (2002)	Implementation fidelity	Direct observation for correct implementation of each step of the intervention	Overall, parents were satisfied with the format and content of the program, as well as their child's progress. Overall, teachers rated the intervention highly, meaning each item of the questionnaire ranged from 4 to 4.88. There was also a high correlation ($r = .89$) between teachers' social validity ratings and fidelity scores.
	Adult use of prompts and attention	Direct observation	n/a
16. Young et al. (2016)	Implementation fidelity	Attendance at training workshops, post-intervention teacher interviews, and direct observation using fidelity checklists.	Teacher exit survey. Positive results: 100% of respondents learned new skills to help their students and 91% planned to continue using the intervention.

Table 4 Experimental design and quality indicators for included studies

Citations	Average Rating	Design Characteristics	Measurement and Reliability				General Characteristics and Results						Dimensions of External Validity				
			Measurement	Reliability	Intervention fidelity*	Training fidelity**	Rationale	Robust Treatment Effects	Quality of Baseline	Visual Analysis	Statistics	Maintenance and Generalization	Implementation Site	Participant Selection	Consumer Satisfaction	Social Validity	
1. Boulware et al. (2006)	1.8	●	○	●	●	●	○	○	n/a	n/a	●	○	○	○	○	●	●
2. D'Elia et al. (2014)	2.2	●	○	●	●	●	○	○	n/a	n/a	○	○	○	○	○	●	●
3. Eikeseth et al. (2012)	2.1	●	○	●	●	●	○	○	n/a	n/a	○	○	○	○	○	●	●
4. Eldevik et al. (2012)	2.5	●	○	○	●	○	○	○	n/a	n/a	○	○	○	○	○	●	●
5. Fleury & Schwartz (2017)	2.6	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○
6. Garfinkle & Schwartz (2002)	2.6	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○
7. Gibson et al. (2010)	2.9	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
8. Harjusola-Webb & Robbins (2012)	2.5	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
9. Kern & Aldridge (2006)	2.7	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
10. Kern et al. (2007)	2.3	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
11. McBride & Schwartz (2003)	2.6	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
12. Olive et al. (2007)	2.3	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○
13. Schwartz et al. (2004)	2.2	●	○	●	●	○	○	○	n/a	n/a	○	○	○	○	○	○	○
14. Strain & Bovey (2011)	3.4	○	○	○	○	○	○	○	n/a	n/a	○	○	○	○	○	○	○
15. VanDerHeyden et al. (2002)	2.3	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○
16. Young et al. (2016)	2.9	○	○	●	○	○	○	○	n/a	n/a	○	○	○	○	○	○	○
Average rating by criterion	2.5	2.4	2.6	2.1	2.4	1.9	3.3	3.1	3.4	3.3	2.0	1.8	4.0	2.6	1.9	1.5	

4 = ● Exemplary, 3 = ○ Acceptable, 2 = ● Minimal, 1 = ● Unacceptable

Table adapted from Goldstein et al. 2014 framework designed for evaluating the quality of research studies for systematic reviews

(Eldevik et al. 2012), which in fact appears to have been based on similar principles to the EIBI intervention reported in Eldevik et al. 2012

The remaining nine studies (56%; Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Harjusola-Webb and Robbins 2012; Kern and Aldridge 2006; Kern et al. 2007; McBride and Schwartz 2003; Olive et al. 2007; Van DerHeyden et al. 2002) focused on interventions that we classified as more targeted because the intervention focused on changing a less comprehensive (more specific) set of skill(s) or focused on a smaller number of specific developmental areas than did the comprehensive programs listed before. The specific skills and developmental areas targeted in these interventions included (a) communication (Gibson et al. 2010; Harjusola-Webb and Robbins 2012; Kern et al. 2007; McBride and Schwartz 2003; Olive et al. 2007), (b) play skills (Kern and Aldridge 2006; Van DerHeyden et al. 2002), (c) peer interaction (Garfinkle and Schwartz 2002; Kern and Aldridge 2006; McBride and Schwartz 2003), and (d) reading skills (Fleury and Schwartz 2017).

Frequency of Intervention Sessions

Frequency of intervention refers to the number of sessions per week. Seven studies (44%) included details of the frequency of intervention sessions (Boulware et al. 2006; Eldevik et al.

2012a, 2012b; Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Olive et al. 2007; VanDerHeyden et al. 2002); the mean number of sessions per week across these studies was 4.2 (range: 2 to 5). A further four studies (25%; Gibson et al. 2010; Kern and Aldridge 2006; Strain and Bovey 2011; Young et al. 2016) provided some information on intervention frequency, but not enough for the mean number or sessions per week to be calculated.

Intensity of Intervention Sessions

Intervention intensity, that is, the mean number of hours of intervention per week, was not specified for one study (Kern and Aldridge 2006) and was unclear for a further study (VanDerHeyden et al. 2002). Another three studies (Gibson et al. 2010; Kern et al. 2007; McBride and Schwartz 2003) provided the number of minutes per intervention session, but not enough information for the mean number of hours per week to be calculated. Two studies (Harjusola-Webb and Robbins 2012; Young et al. 2016) reported the number of hours per week that children were enrolled to attend preschool but not the intensity of the intervention received. For the remaining nine studies (56%), mean intensity of intervention that was delivered in the preschool setting was 9.6 h/week (range 0.3 to 23 h/week).

For three studies (19%), the intervention also included a family/home component that involved an additional number

of hours delivered by the child's family and/or in the child's home (Boulware et al. 2006; D'Elia et al. 2014; Eikeseth et al. 2012). The number of hours per week was not specified by Eikeseth et al. (2012), but for the Boulware et al. (2006) study, the family/home component involved an additional 7 h/week of intervention, and for the D'Elia et al. (2014) study, participants received an additional 2 h of intervention per week delivered in their home.

Duration of Intervention Sessions

Duration refers to the amount of time (in months) over which the intervention was conducted. The mean duration of intervention was reported in 10 studies (63%) and ranged from 1.2 to 25 months with a mean of 15 months. The duration of intervention was not specified or was not clearly specified in four studies (25%; Garfinkle and Schwartz 2002; Harjusola-Webb and Robbins 2012; Olive et al. 2007; VanDerHeyden et al. 2002). Two studies (13%; Gibson et al. 2010; McBride and Schwartz 2003) specified the total number of sessions of intervention received, but not the period of time over which the sessions were delivered.

Staff Training

Five studies (31%) did not report any details on the approach used for training teachers to implement the intervention procedures (Boulware et al. 2006; D'Elia et al. 2014; Eikeseth et al. 2012; Schwartz et al. 2004; Van DerHeyden et al. 2002). For the remaining studies (69%), a range of methods was used to train teaching staff, including (a) providing a formal graduate course in communication intervention (Olive et al. 2007); (b) didactic coaching, mentoring, and/or training (Eldevik et al. 2012; Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Kern and Aldridge 2006; McBride and Schwartz 2003; Young et al. 2016); (c) modeling (Eldevik et al. 2012; Fleury and Schwartz 2017; Gibson et al. 2010; Kern et al. 2007; Strain and Bovey 2011); (d) use of role play (Fleury and Schwartz 2017; Gibson et al. 2010); (e) individual feedback (Fleury and Schwartz 2017; Gibson et al. 2010; Harjusola-Webb and Robbins 2012; McBride and Schwartz 2003; Strain and Bovey 2011); (f) written instruction and/or feedback (Fleury and Schwartz 2017; Gibson et al. 2010; McBride and Schwartz 2003; Strain and Bovey 2011); (g) group instruction and/or coaching (Fleury and Schwartz 2017; Young et al. 2016); (h) workshops (Young et al. 2016); and (i) in vivo practice and/or coaching (McBride and Schwartz 2003; Strain and Bovey 2011). In one study, training was delivered via videoconferencing (Gibson et al. 2010), another study used video modeling (Fleury and Schwartz 2017), and a written training manual was provided to staff in the Harjusola-Webb and Robbins (2012) study. Finally, in the study by Kern et al. (2007), staff training began

with an initial consultation meeting to establish intervention goals and procedures for each participating child.

The frequency, intensity, and duration of staff/teacher training varied across the included studies. Seven studies (44%) did not specify the frequency, intensity, or duration of training (Boulware et al. 2006; D'Elia et al. 2014; Eikeseth et al. 2012; Garfinkle and Schwartz 2002; Olive et al. 2007; Schwartz et al. 2004; VanDerHeyden et al. 2002), and the authors of two studies (13%) did not provide enough information for the exact frequency, intensity, and duration of training to be calculated (Fleury and Schwartz 2017; McBride and Schwartz 2003). For one study (6%; Gibson et al. 2010) training consisted of one 45-min session, this study is not included in the calculations of training frequency, intensity, or duration in this section. Three studies (19%) provided details of the frequency of teacher training (Eldevik et al. 2012; Harjusola-Webb and Robbins 2012; McBride and Schwartz 2003). Training was delivered weekly in two of these studies (Eldevik et al. 2012; Harjusola-Webb and Robbins 2012), and training sessions decreased to bi-weekly for one study as the intervention progressed (Eldevik et al. 2012). The intensity of teacher training was specified in two studies (Eldevik et al. 2012; Harjusola-Webb and Robbins 2012) and ranged from 0.17 to 6 h/week. A further two studies (McBride and Schwartz 2003; Strain and Bovey 2011) provided some information on the intensity of teacher training, but not enough to calculate the mean number of hours per week.

The duration of teacher training was reported in five studies (31%; Eldevik et al. 2012; Kern and Aldridge 2006; Kern et al. 2007; McBride and Schwartz 2003; Strain and Bovey 2011). In two of these studies (Kern and Aldridge 2006; McBride and Schwartz 2003), training ended once staff reached a pre-determined level of fidelity or indicated that they were confident with using the intervention. For the remaining three studies (19%), the mean durations of teacher training were 25 months (Eldevik et al. 2012); 8 months (Kern et al. 2007); and 0.5 months (Strain and Bovey 2011).

Child Outcomes

Table 2 indicates that all 16 studies reported on child outcomes. Outcomes have been grouped according to study design. Specifically, the nine studies (56%) with single-case designs primarily used direct observation to measure child outcomes so these studies have been grouped together. The remaining seven studies (44%), which all had group designs, used a range of instruments other than direct observation to assess child outcomes. The findings from these studies have been grouped according to domains of functioning.

Child Outcomes for Single-Case Design Studies The nine studies (56%) with a single-case design included at least one child

outcome measured by in vivo or video observation of child behavior (Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Harjusola-Webb and Robbins 2012; Kern and Aldridge 2006; Kern et al. 2007; McBride and Schwartz 2003; Olive et al. 2007; Van DerHeyden et al. 2002). The specific outcomes measured varied across studies and included frequency of elopement (Gibson et al. 2010); verbal participation in the target reading activity (Fleury and Schwartz 2017); social initiations (Garfinkle and Schwartz 2002); verbal responses (Garfinkle and Schwartz 2002; McBride and Schwartz 2003); imitation of peers (Garfinkle and Schwartz 2002); engagement and physical proximity to peers (Garfinkle and Schwartz 2002; McBride and Schwartz 2003); peer imitation of participating children (Garfinkle and Schwartz 2002); frequency of expressive communicative acts (Harjusola-Webb and Robbins 2012); positive peer interactions, play, and engagement with materials and equipment (Kern and Aldridge 2006); number of correctly completed steps of the morning arrival routine (Kern et al. 2007); use of a speech-generating device for communication (Olive et al. 2007); and contact with target activity centers, toy play, and disruptive behavior (Van DerHeyden et al. 2002). One study (Fleury and Schwartz 2017) also included a researcher-developed assessment of book-specific vocabulary.

Due to the wide variety of outcomes measured across the nine single-case design studies, it is not possible to make direct comparisons. Results were coded as positive for five studies (Gibson et al. 2010; Harjusola-Webb and Robbins 2012; Kern et al. 2007; McBride and Schwartz 2003; Olive et al. 2007). Mixed results or minimal improvements were reported in four studies (Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Kern and Aldridge 2006; Van DerHeyden et al. 2002). For example, Kern and Aldridge (2006) reported positive results across all participants, but for only two of the three intervention phases and Fleury and Schwartz (2017) reported minimal improvement in child verbal initiations, but positive results for all other measured child outcomes.

Child Outcomes for Group Design Studies For the seven group design studies, a range of instruments were used to measure child outcomes. Six (86%) of these group studies (Boulware et al. 2006; D’Elia et al. 2014; Eikeseth et al. 2012; Eldevik et al. 2012; Strain and Bovey 2011; Young et al. 2016) measured children’s adaptive/maladaptive behavior using a range of instruments including (a) Temperament and Atypical Behavior Scale (TABS; Bagnato et al. 1999), (b) Vineland Adaptive Behavior Scales (VABS; Sparrow et al. 1984), (c) Child Behavior Checklist (CBC; Achenbach 1991), and (d) the Social Skills Rating System (SSRS; Gresham and Elliott 1990). Five (71%) of these group design studies that measured adaptive/maladaptive behavior reported positive results (Boulware et al. 2006; D’Elia et al. 2014; Eikeseth et al.

2012; Eldevik et al. 2012; Strain and Bovey 2011), while the remaining study was coded as having no effect because there were no significant changes in participant scores for the EG (Young et al. 2016).

Functional skills were measured as outcomes in two (29%) of the group studies (Boulware et al. 2006; Schwartz et al. 2004) and were assessed using (a) Bayley Scales of Infant Development (Bayley 2006); (b) Assessment, Evaluation, and Programming System for Infants and Children (AEPS; Bricker 1994); (c) a researcher-developed functional outcomes index (Schwartz et al. 2004); and (d) a researcher-developed functional outcomes scale (Boulware et al. 2006). Participating children from both studies demonstrated gains across at least one functional outcome, and participants from the Schwartz et al. (2004) study made gains across all six of the functional outcomes measured.

Four (57%) of the seven group studies reported on measures of autism severity and/or symptoms (D’Elia et al. 2014; Eikeseth et al. 2012; Strain and Bovey 2011; Young et al. 2016) using the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 2008), or the Childhood Autism Rating Scale (CARS; Schopler et al. 2002). D’Elia et al. (2014) reported decreases in autism diagnoses across both EGs and CGs, as measured by the ADOS, with a larger decrease observed in the EG. Similarly, in the study by Strain and Bovey (2011), the EG demonstrated a greater decrease in CARS scores than the CG. The Eikeseth study (2012) reported a significant decrease in CARS scores for the EG, but did not report comparison data for the CG. The authors of the final study (Young et al. 2016) did not report any significant change in CARS scores.

Child communication and/or language was measured in five (71%) of the group studies (Boulware et al. 2006; D’Elia et al. 2014; Fleury and Schwartz 2017; Strain and Bovey 2011; Young et al. 2016) via a range of different instruments including (a) Communication, Social, and Symbolic Behavior Scales (CSBS; Wetherby and Prizant 2002); (b) MacArthur Communication Developmental Inventories (CDI; Fenson et al. 1993; Fenson et al. 1994); (c) Preschool Language Scale (PLS; Zimmerman et al. 1991); (d) Expressive One Word Picture Vocabulary Test (EOWPVT; Brownell 2000a); (e) Receptive One Word Picture Vocabulary Test (ROWPVT; Brownell 2000b); and (f) a researcher-delivered book vocabulary assessment (Fleury and Schwartz 2017). Participants demonstrated improvement on at least one communication/language outcome across all five of these studies.

Two (29%) of the group studies (Strain and Bovey 2011; Young et al. 2016) reported on social skills, which were measured via the Social Skills Rating System (SSRS; Gresham and Elliott 1990) and the Autism Screening Instrument for Educational Planning (ASIEP; Krug et al. 2008). Both studies

reported positive results, with the EG making greater improvements than the CG in both cases.

Aspects of child cognition or educational strengths and weaknesses were reported as outcomes in four (57%) of the group studies (D'Elia et al. 2014; Eldevik et al. 2012; Strain and Bovey 2011; Young et al. 2016). Intellectual functioning was measured in one study (Eldevik et al. 2012) using the Bayley Scales of Infant Development (BSID; Bayley 2006) for participants younger than 42 months of age, and the Stanford-Binet Intelligence Scale (Thorndike et al. 1986) for participants older than 42 months. Overall, the EG made significantly greater gains than the CG on composite scores for both instruments. Another study (D'Elia et al. 2014) measured psycho-educational skills using the Psychoeducational Profile: Third Edition (PEP-3; Schopler et al. 2005) and found that EG participants made significant improvements over time across most categories. Finally, child cognitive development was measured in two (29%) of the seven group studies (Strain and Bovey 2011; Young et al. 2016) using the Mullen Scales of Early Learning (Mullen 1995) and the cognitive domain of the Bayley Scales of Infant Development (BDI; Bayley 2006). In the Strain and Bovey (2011) study, EG scores were significantly higher than CG scores after intervention; however, no significant change in scores was reported in the Young et al. study.

Teaching Staff Outcomes

Table 3 shows that four (25%) of the included studies did not report any measures of teacher outcomes (Boulware et al. 2006; Eikeseth et al. 2012; Eldevik et al. 2012; Schwartz et al. 2004). For the purposes of this review, the term “implementation fidelity” has been used as an umbrella term to describe the extent to which interventions were delivered as intended and in line with the program model or prescribed procedures. The included studies used a range of terms to refer to this concept including fidelity of intervention implementation (D'Elia et al. 2014; Olive et al. 2007; Strain and Bovey 2011; Young et al. 2016); procedural fidelity (Fleury and Schwartz 2017; Van DerHeyden et al. 2002); treatment fidelity (Garfinkle and Schwartz 2002); and fidelity of treatment (Kern and Aldridge 2006).

Implementation fidelity was measured as a teacher outcome in 9 (56%) of the 16 included studies (D'Elia et al. 2014; Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Kern and Aldridge 2006; Olive et al. 2007; Strain and Bovey 2011; Van DerHeyden et al. 2002; Young et al. 2016). However, in one of these nine studies (D'Elia et al. 2014), the authors did not actually provide the results of their fidelity checks. Implementation fidelity was measured via direct observation in eight of these studies (Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Kern and Aldridge 2006; Olive et al.

2007; Strain and Bovey 2011; Van DerHeyden et al. 2002; Young et al. 2016), and through review of participating children's individual education plans (IEP) for the ninth study (D'Elia et al. 2014). Young et al. also included data on teachers' rate of attendance at training workshops and responses from teacher exit interviews to support data from the in-class observations. Six studies (38%) reported high levels of implementation fidelity ($M = 90%$, range of means = 73 to 100%; Fleury and Schwartz 2017; Gibson et al. 2010; Olive et al. 2007; Strain and Bovey 2011; VanDerHeyden et al. 2002; Young et al. 2016). Two studies (13%) reported mixed results with some participating teachers failing to reach high fidelity levels (Kern and Aldridge 2006) or requiring additional coaching to reach required fidelity levels (Garfinkle and Schwartz 2002). Young et al. (2016) also reported high rates of attendance at teacher-training workshops.

Other teacher outcomes related to the use of specific teaching techniques include (a) rate of dialogic prompt use (Fleury and Schwartz 2017); (b) use of communication-promoting strategies (Harjusola-Webb and Robbins 2012); (c) use of prompting (Kern et al. 2007); (d) rate of instruction, use of physical prompts and interactions with the target child (McBride and Schwartz 2003); and (e) use of prompts and attention (VanDerHeyden et al. 2002). For most of these studies, outcomes were assessed via direct observation (Harjusola-Webb and Robbins 2012; Kern et al. 2007; Van DerHeyden et al. 2002) or video observation (McBride and Schwartz 2003). However, Fleury and Schwartz (2017) did not specify the method of measurement used. Two studies reported positive results (Harjusola-Webb and Robbins 2012; Van DerHeyden et al. 2002), one study reported mixed results (McBride and Schwartz 2003) and two studies did not report the results (Fleury and Schwartz 2017; Kern et al. 2007).

Social Validity

As displayed in Table 3, nine studies (56%) did not appear to have assessed social validity (Boulware et al. 2006; D'Elia et al. 2014; Eikeseth et al. 2012; Eldevik et al. 2012; Harjusola-Webb and Robbins 2012; Kern and Aldridge 2006; Kern et al. 2007; Olive et al. 2007; Schwartz et al. 2004; Van DerHeyden et al. 2002). Among the seven studies (44%) that did include a measure of social validity, five of these studies did so by administration of a questionnaire that was completed by the teaching staff (Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Strain and Bovey 2011; Young et al. 2016), whereas one study used teacher interviews (McBride and Schwartz 2003), and one study reported on what would appear to have been more anecdotal evidence on social validity that was provided by participants' families (Schwartz et al. 2004). Due to the range of different measures used to evaluate social validity across these

seven studies, results cannot be summarized or compared. However, the reported results were generally positive, with the interventions rated highly by teachers (Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; McBride and Schwartz 2003; Strain and Bovey 2011; Young et al. 2016) and parents (Schwartz et al. 2004). Strain and Bovey (2011) reported a strong correlation ($r = .89$) between teachers' implementation fidelity and their ratings of the social validity of the intervention.

Design Characteristics and Quality Ratings

A detailed analysis of each study's design characteristics and presence of specific quality indicators are presented in Table 4. Nine (56%) of the included articles used single-case research designs (Fleury and Schwartz 2017; Garfinkle and Schwartz 2002; Gibson et al. 2010; Harjusola-Webb and Robbins 2012; Kern and Aldridge 2006; Kern et al. 2007; McBride and Schwartz 2003; Olive et al. 2007; Van DerHeyden et al. 2002). The remaining seven articles (44%) used group designs including a single group pre-test and post-test design (Boulware et al. 2006; Schwartz et al. 2004), nonequivalent comparison-group design (D'Elia et al. 2014; Eikeseth et al. 2012; Eldevik et al. 2012), randomized controlled trial (Strain and Bovey 2011), and cluster-randomized trial (Young et al. 2016).

The quality of each study was assessed using the quality-rating framework developed by Goldstein et al. (2014), and the results are displayed in Table 4. The mean quality rating across all of the included studies was 2.5 out of 4 (range = 1.8 to 3.4). This mean score can be loosely translated as indicating a minimal acceptable level of quality (Goldstein et al. 2014). In order to receive a rating of "minimal quality," studies needed to have an overall mean score of at least 2. Boulware et al.'s (2006) study received a mean rating of 1.8 and was the only included study that did not meet the standards for minimal quality. Only one of the included studies (Strain and Bovey 2011) demonstrated an "acceptable" level of quality, with a mean score of 3.4. The mean scores for the remaining 14 studies ranged from 2.1 to 2.9. With respect to scores for each category, studies tended to score highly for study rationale ($M = 3.3$), robust treatment effects ($M = 3.1$), and the external validity of the implementation site ($M = 4$), and studies with single-case designs tended to score highly for quality of baseline ($M = 3.4$) and visual analysis ($M = 3.3$). Lower mean scores were obtained for social validity ($M = 1.5$), consumer satisfaction ($M = 1.9$) maintenance and generalization ($M = 1.8$), and training fidelity ($M = 1.9$). The mean scores for the remaining categories ranged from 2 to 2.6.

Discussion

The purpose of this review was to gather, summarize, and evaluate the empirical literature regarding teacher-implemented early interventions for young children with ASD. The review was limited to studies conducted in inclusive preschool settings to assess intervention effectiveness, that is, the effects of interventions when implemented under "real-world" conditions. A systematic search of the literature produced 16 articles that met the inclusion criteria. Nine (56%) of the included studies used a single-case research design, three studies (19%) used a nonequivalent comparison-group design, two (13%) used a single-group pre-test post-test design, and two studies (13%) were randomized controlled trials. Various intervention approaches were evaluated across the studies, including six different comprehensive interventions and a range of targeted interventions focused on specific skills or developmental areas. Intervention dosage also varied widely across studies, as did the method, frequency, and duration of teacher training.

Child Outcomes

Overall, the results from the present review suggest that interventions delivered by teaching staff in an inclusive preschool setting can be effective in improving outcomes for young children with ASD. For 14 (88%) of the 16 included studies, the participating children were reported to have made gains in at least one primary outcome variable. For the remaining two studies (13%), the participating children showed minimal improvement or variable gains across the primary outcome measures (D'Elia et al. 2014; Garfinkle and Schwartz 2002). Interestingly, both of these studies reported positive results on secondary or collateral outcomes. The Garfinkle and Schwartz (2002) study also reported some generalization of child behaviors to nontarget activities and/or peers and highly favorable teacher ratings regarding the benefit of the intervention to target children. For the D'Elia et al. study, mean improvements for the EG were only slightly higher than those observed in the CG across most primary outcomes; however, no implementation fidelity data were reported and so it is difficult to know whether teaching staff implemented the intervention with integrity.

More than 20 different child outcomes were measured across the 16 included studies via a range of different methods and instruments. Direct observation (either in vivo or via video) was the most commonly used method. The specific outcomes measured via direct observation varied widely across studies, possibly because of the wide variety of outcomes targeted by different interventions (particularly interventions focused on specific skills/developmental areas). However, there was also a lack of consistency in the way that broader outcomes (e.g., adaptive behavior, autism severity, and

communication skills) were measured. Twenty different instruments were used to measure broad child outcomes across six (38%) of the included studies, but only four of these instruments were used in more than one study. This variation in the approach to the measurement of outcomes across studies would seem to hinder cross-study comparisons of interventions, which is important for informing decisions regarding which treatment/treatments should be considered best practice (Ospina et al. 2008).

It is also important to determine whether a given intervention has a more positive effect than treatment as usual (TAU). Thus, it is useful for intervention research to compare the focus intervention with the treatment(s) that participants would typically receive. Furthermore, studies need to include a clear, detailed description of TAU conditions (Dingfelder and Mandell 2011). Four (25%) of the studies included in this review compared specific interventions with TAU and found that the intervention group performed significantly better than the CG on at least one primary outcome (Eikeseth et al. 2012; Eldevik et al. 2012; Strain and Bovey 2011; Young et al. 2016). This suggests that the interventions being evaluated in these studies appeared to be more effective than the treatment that the target children would have ordinarily received, had they not participated in the study. It is important to note that for the Eldevik et al. (2012) study, the participants were not randomly allocated to a treatment condition and this may have compromised the validity of the comparison between the EG and the CG. However, the authors reported group equivalence with respect to age, gender, duration of intervention, diagnosis, and level of intellectual disability prior to the beginning of the study. A further study (D'Elia et al. 2014) also compared intervention with TAU and reported that EG improvements on primary outcomes were not significantly better than those recorded for the CG.

Research on intervention effectiveness should also include measures of the generalization and maintenance of observed results (Koegel and Rincover 1977). However, only two studies (13%) reported on generalization of child behavior (Garfinkle and Schwartz 2002; Van DerHeyden et al. 2002). VanDerHeyden et al. reported on generalization of child behavior to nontarget activity centers and the study by Garfinkle and Schwartz included data on generalization of child behavior to a nontarget activity and nontarget peers. The study by Garkinkle and Schwartz also included a 2- to 4-week follow-up phase and was the only included study to report on maintenance of child behavior. A further five studies (31%) did not include a follow-up phase, but measured child outcomes across a relatively long-term (> 12 months) intervention phase (D'Elia et al. 2014; Eldevik et al. 2012; Strain and Bovey 2011; Young et al. 2016). However, these data do not show whether target behavior(s) occurred absent the intervention, an important consideration given that it is possible for

improvements made through early intervention to decline after the intervention ends (Estes et al. 2015).

Teacher Outcomes

It is generally accepted that intervention research should include data on implementation fidelity. Such data can help to establish the extent to which the intervention was implemented as intended and thus whether its correct implementation was likely to have been responsible for any positive intervention effect (Ospina et al. 2008). The measurement of fidelity in community-based settings is especially important as it is common for fidelity to be compromised when an intervention is transferred from a more controlled clinical setting to a more natural, applied setting (Breitenstein et al. 2010; Chang et al. 2016). Research has demonstrated that child outcomes are impacted by implementation fidelity (Stahmer et al. 2017). Indeed, two of the included studies highlighted decreases in child performance corresponding with decreased levels of teacher fidelity (Kern and Aldridge 2006; Van DerHeyden et al. 2002).

With respect to reporting on implementation fidelity, eight studies (50%) reported data related to this quality indicator. This number suggests a reporting improvement over time in that an earlier review by Wheeler et al. (2006) reported that only 18% of studies reported on implementation fidelity. Of the eight studies from the current review that did report on implementation fidelity, six studies reported positive results with teachers reaching and maintaining high levels of fidelity during the intervention. For the remaining two studies (Garfinkle and Schwartz 2002; Kern and Aldridge 2006), results were mixed with some teachers failing to reach high levels of fidelity and others requiring extra coaching to meet acceptable fidelity levels. Interestingly, although studies referred to acceptable and high levels of fidelity, there did not appear to be any clear consensus across studies regarding the level of performance required to reach each level. For example, in the study by Young et al. (2016), a mean fidelity rate of 73% was deemed to be high, while in the study by Van DerHeyden et al. (2002), a mean fidelity rate of 72% was described as poor. Further, none of these eight studies provided a minimum standard of fidelity, that is, the level of fidelity required to establish experimental control and/or ensure the effectiveness of the intervention.

Also of note was the absence of generalization and maintenance data for teacher fidelity across the included studies. Only one included study (6%) reported on the generalization of teacher behavior to nontarget activities and children (McBride and Schwartz 2003). Results suggested that teacher fidelity improved from baseline but to a lesser extent than the improvement seen with target activities and children. None of the included studies reported follow-up data for teacher fidelity. This type of data may be important in establishing teacher-

delivered intervention as a feasible long-term option for early intervention for children with ASD. This is because a potentially valuable goal would be for community-based intervention to become largely self-sustaining, with teachers able to implement intervention with only minimal input from outside experts such as researchers. Indeed, the long-term feasibility and sustainability of interventions is an important consideration for any mental health–related intervention research study (Proctor et al. 2009) as current research suggests that many community-implemented interventions do not sustain over time (Dingfelder and Mandell 2011).

Social Validity

Based on the quality rating scale used in this review (Goldstein et al. 2014), social validity refers to the educational or clinical significance of a study while consumer satisfaction refers to stakeholder’s perceptions of the acceptability of the treatment or intervention. We use the term social validity as an umbrella term to encompass both of these aspects. Social validity is an important consideration in the implementation of interventions in community settings because of the well-documented research-to-practice gap (Drahota et al. 2012) and the recognized role of stakeholder perceptions in bridging this gap (Stahmer et al. 2017). Findings from the included study by Strain and Bovey (2011) suggested a high correlation ($r = .89$) between teachers’ ratings of the social validity of the intervention and their implementation fidelity, further highlighting the potential importance of social validity. Indeed, in the Strain and Bovey study, teachers that viewed the intervention as socially valid were more likely to implement it with high levels of fidelity.

It is concerning then that only six studies (38%) included in this review met the quality requirements for acceptable measurement of social validity. All six of these studies reported overall positive results, suggesting that the interventions were viewed favorably by teaching staff. Of the remaining 10 studies, one study included a measure of social validity that was rated as “minimal” quality and the remaining 9 studies did not include any measure of social validity. Clearly, future research could be strengthened by greater attention to the assessment of social validity.

Limitations

Several limitations need to be considered with this review. Firstly, we limited included studies to peer-reviewed articles published in English, and as a result, we may have missed studies that might have otherwise met the criteria for inclusion. Also, the exclusion of unpublished work, such as theses and dissertations, may have increased the likelihood of no effects or negative results due to publication bias. Finally,

included studies may have received different quality ratings if we had used an alternative framework to assess quality.

Implications

The results of this review would seem to suggest that teaching staff might be able to learn how to effectively deliver early intervention, with a reasonable degree of fidelity, to young children with ASD in inclusive preschool environments. Further, the delivery of early intervention in these settings may improve outcomes for the participating children. However, many of the reviewed studies had minimally acceptable levels of quality based on the Goldstein et al. (2014) rating framework. Given the need for high-quality studies to guide evidence-based practice, the results of this review point to the need for additional and higher-quality studies. At the present time, any statements regarding the effectiveness of the early interventions included in this review must be viewed as tentative. Although these results suggest that a range of early intervention programs can be effective when implemented in inclusive preschool settings, further research is needed to establish the generality of the findings of this review. Specifically, there is a need for more high-quality studies that evaluate the long-term effectiveness of interventions and the long-term maintenance of both child and teacher outcomes. These studies should also include assessment of the social validity as well as the long-term feasibility of use of the intervention for different communities (Dingfelder and Mandell 2011). Alongside this broad long-term research agenda, smaller studies that examine the active components of intervention could be informative and may enable providers to make use of only those intervention components that are most likely to be necessary. Eliminating the use of inactive treatment components may reduce the costs and increase the efficiency of intervention efforts. Single-case research designs would seem well-g geared towards identifying the active ingredients of a given intervention, as well as for evaluating the generalization and maintenance of treatment effects (Lord et al. 2005; Ward-Horner and Sturmey 2010).

There may also be value in future comparative research to determine which intervention approach or package, if any, is the most effective when delivered in an inclusive preschool setting. This type of research should also include a thorough assessment of the initial and ongoing costs of different intervention approaches as well as their acceptability to stakeholders (Dingfelder and Mandell 2011). It may also be valuable to compare inclusive preschool-based delivery of interventions with one-on-one or specialist preschool delivery to determine which mode of delivery is most effective. Future research should also include clearly defined child outcomes that are explicitly linked to the expected outcomes of the intervention being studied (Lord et al. 2005). It would be valuable for these outcomes to be measured with a consistent set of

instruments across studies, to allow for future cross-study comparisons. Finally, future studies should include a clearly defined measure of implementation fidelity. Researchers might consider specifying a minimum level of acceptable fidelity. In line with this, future research that more closely examines possible links between implementation fidelity and child outcomes would seem of some value.

Funding Information The preparation of this manuscript was supported in part by a PhD scholarship awarded to Jessica Tupou from Victoria University of Wellington.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

References

(*Study included in the review

- Achenbach, T. M. (1991). *Child behavior checklist*. San Antonio, TX: Psychological Cooperation.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing. <https://doi.org/10.1176/appi.books.9780890425596>.
- Bagnato, S. J., Neisworth, J. T., Salvia, J. J., & Hunt, F. M. (1999). *Temperament and atypical behavior scale (TABS). Early childhood indicators of developmental dysfunction*. Baltimore: Paul H. Brookes Publishing Co.
- Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., ... & Durkin, M. S. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveillance Summaries*, 67(6), 1.
- Bayley, N. (2006). *Bayley scales of infant and toddler development*. Upper Saddle River, NJ: Pearson.
- *Boulware, G. L., Schwartz, I. S., Sandall, S. R., & McBride, B. J. (2006). Project DATA for toddlers: An inclusive approach to very young children with autism spectrum disorder. *Topics in Early Childhood Special Education*, 26, 94–105. doi:<https://doi.org/10.1177/02711214060260020401>
- Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L., & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing & Health*, 33(2), 164–173. <https://doi.org/10.1002/nur.20373>.
- Bricker, D. (1994). *Assessment, evaluation, and programming system (AEPS) for infants and children* (Vol. 1): *AEPS measurement for birth to three years*. Baltimore: Paul H. Brookes Publishing Co.
- Broderick, A. (2017). Article 7 [children with disabilities]. In V. Della Fina, R. Cera, & G. Palmisano (Eds.), *The United Nations Convention on the Rights of Persons with Disabilities* (pp. 195–212). Cham: Springer.
- Brownell, R. (2000a). *Expressive one-word picture vocabulary test*. Novato, CA: Academic Therapy Publications.
- Brownell, R. (2000b). *Receptive one-word picture vocabulary test*. Novato, CA: Academic Therapy Publications.
- Chang, Y. C., Shire, S. Y., Shih, W., Gelfand, C., & Kasari, C. (2016). Preschool deployment of evidence-based social communication intervention: JASPER in the classroom. *Journal of Autism and Developmental Disorders*, 46, 2211–2223. <https://doi.org/10.1007/s10803-016-2752-2>.
- Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children*, 75, 365–383. <https://doi.org/10.1177/001440290907500306>.
- Dawson, G. (2008). Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Development and Psychopathology*, 20, 775–803. <https://doi.org/10.1017/S0954579408000370>.
- Debodinance, E., Maljaars, J., Noens, I., & Van den Noortgate, W. (2017). Interventions for toddlers with autism spectrum disorder: A meta-analysis of single-subject experimental studies. *Research in Autism Spectrum Disorders*, 36, 79–92. <https://doi.org/10.1016/j.rasd.2017.01.010>.
- *D’Elia, L., Valeri, G., Sonnino, F., Fontana, I., Mammone, A., & Vicari, S. (2014). A longitudinal study of the TEACCH program in different settings: The potential benefits of low intensity intervention in preschool children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 44(3), 615–626. doi:<https://doi.org/10.1007/s10803-013-1911-y>
- Dingfelder, H. E., & Mandell, D. S. (2011). Bridging the research-to-practice gap in autism intervention: An application of diffusion of innovation theory. *Journal of Autism and Developmental Disorders*, 41, 597–609. <https://doi.org/10.1007/s10803-010-1081-0>.
- Drahota, A., Aarons, G. A., & Stahmer, A. C. (2012). Developing the autism model of implementation for autism spectrum disorder community providers: Study protocol. *Implementation Science*, 7, 85. <https://doi.org/10.1186/1748-5908-7-85>.
- *Eikeseth, S., Klintwall, L., Jahr, E., & Karlsson, P. (2012). Outcome for children with autism receiving early and intensive behavioral intervention in mainstream preschool and kindergarten settings. *Research in Autism Spectrum Disorders*, 6, 829–835. doi:<https://doi.org/10.1016/j.rasd.2011.09.002>
- *Eldevik, S., Hastings, R. P., Jahr, E., & Hughes, J. C. (2012). Outcomes of behavioral intervention for children with autism in mainstream pre-school settings. *Journal of Autism and Developmental Disorders*, 42, 210–220. doi:<https://doi.org/10.1007/s10803-011-1234-9>.
- Estes, A., Munson, J., Rogers, S. J., Greenson, J., Winter, J., & Dawson, G. (2015). Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54, 580–587. <https://doi.org/10.1016/j.jaac.2015.04.005>.
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59, 1–173. <https://doi.org/10.2307/1166093>.
- Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur communicative development inventories: User’s guide and technical manual*. San Diego, CA: Singular Publishing Group.
- *Fleury, V. P., & Schwartz, I. S. (2017). A modified dialogic reading intervention for preschool children with autism spectrum disorder. *Topics in Early Childhood Special Education*, 37, 16–28. doi:<https://doi.org/10.1177/0271121416637597>
- *Garfinkle, A. N., & Schwartz, I. S. (2002). Peer imitation: Increasing social interactions in children with autism and other developmental disabilities in inclusive preschool classrooms. *Topics in Early Childhood Special Education*, 22, 26–38. doi:<https://doi.org/10.1177/027112140202200103>
- *Gibson, J. L., Pennington, R. C., Stenhoff, D. M., & Hopper, J. S. (2010). Using desktop videoconferencing to deliver interventions to a preschool student with autism. *Topics in Early Childhood*

- Special Education, 29, 214–225. doi:<https://doi.org/10.1177/0271121409352873>
- Goldstein, H., Lackey, K. C., & Schneider, N. J. (2014). A new framework for systematic reviews: Application to social skills interventions for preschoolers with autism. *Exceptional Children, 80*, 262–286. <https://doi.org/10.1177/0014402914522423>.
- Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system (SSRS)*. Circle Pines, MN: American Guidance Service.
- *Harjusola-Webb, S. M., & Robbins, S. H. (2012). The effects of teacher-implemented naturalistic intervention on the communication of preschoolers with autism. *Topics in Early Childhood Special Education, 32*, 99–110. doi:<https://doi.org/10.1177/0271121410397060>
- Interagency Autism Coordinating Committee (IACC). (2013). IACC Strategic Plan for Autism Spectrum Disorder (ASD) Research—2013 Update. U.S. Department of Health and Human Services Interagency Autism Coordinating Committee. Retrieved from <http://iacc.hhs.gov/strategic-plan/2013/index.shtml>
- Keenan, M., & Dillenburger, K. (2011). When all you have is a hammer: RCTs and hegemony in science. *Research in Autism Spectrum Disorders, 5*, 1–13. <https://doi.org/10.1016/j.rasd.2010.02.003>.
- *Kern, P., & Aldridge, D. (2006). Using embedded music therapy interventions to support outdoor play of young children with autism in an inclusive community-based child care program. *Journal of Music Therapy, 43*, 270–294. doi:<https://doi.org/10.1093/jmt/43.4.270>
- *Kern, P., Wolery, M., & Aldridge, D. (2007). Use of songs to promote independence in morning greeting routines for young children with autism. *Journal of Autism and Developmental Disorders, 37*, 1264–1271. doi:<https://doi.org/10.1007/s10803-006-0272-1>
- Koegel, L. K., Koegel, R. L., Frea, W. D., & Fredeen, R. M. (2001). Identifying early intervention targets for children with autism in inclusive school settings. *Behavior Modification, 25*, 745–761. <https://doi.org/10.1177/0145445501255005>.
- Koegel, R., & Rincover, A. (1977). Research on the difference between generalization and maintenance in extra-therapy responding. *Journal of Applied Behavior Analysis, 10*, 1–12. <https://doi.org/10.1901/jaba.1977.10-1>.
- Kramer, M. R., Schneider, E. B., Kane, J. B., Margerison-Zilko, C., Jones-Smith, J., King, K., et al. (2017). Getting under the skin: Children's health disparities as embodiment of social class. *Population Research and Policy Review, 36*, 671–697. <https://doi.org/10.1007/s11113-017-9431-7>.
- Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence based interventions in school psychology: Conceptual foundations of the procedural and coding manual of division 16 and the society for the study of school psychology task force. *School Psychology Quarterly, 17*, 341–389. <https://doi.org/10.1521/scpq.17.4.341.20872>.
- Krug, D. A., Arick, J. R., & Almond, P. (2008). *Autism screening instrument for educational planning*. Austin, TX: Pro-ed.
- 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. <https://doi.org/10.1037/a0028506>.
- Lord, C., Rutter, M., Pamela, C., Dilavore, & Risi, S. (2008). *ADOS: Autism diagnostic observation schedule – ADOS manual*. Los Angeles: Western Psychological Services.
- Lord, C., Wagner, A., Rogers, S., Szatmari, P., Aman, M., Charman, T., ... & Harris, S. (2005). Challenges in evaluating psychosocial interventions for autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 35*, 695–708. doi:<https://doi.org/10.1007/s10803-005-0017-6>
- *McBride, B. J., & Schwartz, I. S. (2003). Effects of teaching early interventionists to use discrete trials during ongoing classroom activities. *Topics in Early Childhood Special Education, 23*, 5–17. doi:<https://doi.org/10.1177/027112140302300102>
- Mitchell, L. C., & Hegde, A. V. (2007). Beliefs and practices of in-service preschool teachers in inclusive settings: Implications for personnel preparation. *Journal of Early Childhood Teacher Education, 28*, 353–366. <https://doi.org/10.1080/10901020701686617>.
- Mullen, E. (1995). *Mullen scales of early learning*. Circle pines, MN: American Guidance Service.
- Odom, S. L., Boyd, B. A., Hall, L. J., & Hume, K. (2010). Evaluation of comprehensive treatment models for individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*, 425–436. <https://doi.org/10.1007/s10803-009-0825-1>.
- *Olive, M. L., de la Cruz, B., Davis, T. N., Chan, J. M., Lang, R. B., O'Reilly, M. F., & Dickson, S. M. (2007). The effects of enhanced milieu teaching and a voice output communication aid on the requesting of three children with autism. *Journal of Autism and Developmental Disorders, 37*, 1505–1513. doi:<https://doi.org/10.1007/s10803-006-0243-6>
- Ospina, M. B., Seida, J. K., Clark, B., Karkhaneh, M., Hartling, L., Tjosvold, L., ... & Smith, V. (2008). Behavioural and developmental interventions for autism spectrum disorder: A clinical systematic review. *PLoS One, 3*(11), e3755. doi:<https://doi.org/10.1371/journal.pone.0003755>
- Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health and Mental Health Services Research, 36*, 24–34. <https://doi.org/10.1007/s10488-008-0197-4>.
- Reichow, B. (2012). Overview of meta-analyses on early intensive behavioral intervention for young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 42*, 512–520. <https://doi.org/10.1007/s10803-011-1218-9>.
- Schopler, E., Lansing, M. D., Reichler, R. J., & Marcus, L. M. (2005). *Examiner's manual of psychoeducational profile* (3rd ed.). Austin, TX: Pro-ed Incorporation.
- Schopler, E., Reichler, R. J., & Renner, B. R. (2002). *The childhood autism rating scale (CARS)*. Los Angeles, CA: Western Psychological Services.
- *Schwartz, I. S., Sandall, S. R., McBride, B. J., & Boulware, G. L. (2004). Project DATA (developmentally appropriate treatment for autism) an inclusive school-based approach to educating young children with autism. *Topics in Early Childhood Special Education, 24*, 156–168. doi:<https://doi.org/10.1177/02711214040240030301>
- Simpson, R. L. (2004). Finding effective intervention and personnel preparation practices for students with autism spectrum disorders. *Exceptional Children, 70*, 135–144. <https://doi.org/10.1177/001440290407000201>.
- Singal, A. G., Higgins, P. D. R., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology, 5*, e45. <https://doi.org/10.1038/ctg.2013.13>.
- Snyder, P. A., Rakap, S., Hemmeter, M. L., McLaughlin, T. W., Sandall, S., & McLean, M. E. (2015). Naturalistic instructional approaches in early learning: A systematic review. *Journal of Early Intervention, 37*(1), 69–97. <https://doi.org/10.1177/1053815115595461>.
- Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales*. Circle Pines, MN: American Guidance Service.
- Stahmer, A. C., Aranbarri, A., Drahota, A., & Rieth, S. (2017). Toward a more collaborative research culture: Extending translational science from research to community and back again. *Autism, 21*, 259–261. <https://doi.org/10.1177/1362361317692950>.
- *Strain, P. S., & Bovey, E. H. (2011). Randomized, controlled trial of the LEAP model of early intervention for young children with autism spectrum disorders. *Topics in Early Childhood Special Education, 31*, 133–154. doi:<https://doi.org/10.1177/0271121411408740>
- Thomas, K. C., Ellis, A. R., McLaurin, C., Daniels, J., & Morrissey, J. P. (2007). Access to care for autism-related services. *Journal of Autism*

- and *Developmental Disorders*, 37, 1902–1912. <https://doi.org/10.1007/s10803-006-0323-7>.
- Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale*. Rolling Meadows, IL: Riverside Publishing.
- United Nations (2006). Convention on the rights of persons with disabilities. Retrieved from: <http://www.un.org/disabilities/documents/convention/convoptprot-e.pdf>
- *Van DerHeyden, A. M., Snyder, P., DiCarlo, C. F., Stricklin, S. B., & Vagianos, L. A. (2002). Comparison of within-stimulus and extra-stimulus prompts to increase targeted play behaviors in an inclusive early intervention program. *The Behavior Analyst Today*, 3, 188. doi:<https://doi.org/10.1037/h0099967>
- Vivanti, G., Duncan, E., Dawson, G., & Rogers, S. J. (2017). *Implementing the group-based Early Start Denver Model for pre-schoolers with autism*. Cham, Switzerland: Springer International Publishing.
- Ward-Homer, J., & Sturmey, P. (2010). Component analysis using single-subject experimental designs: A review. *Journal of Applied Behavior Analysis*, 43, 685–704. <https://doi.org/10.1901/jaba.2010.43-685>.
- Wetherby, A. M., & Prizant, B. M. (2002). *Communication and symbolic behavior scales: Developmental profile* (1st normed ed.). Baltimore: Paul H. Brookes Publishing Co.
- Wheeler, J. J., Baggett, B. A., Fox, J., & Blevins, L. (2006). Treatment integrity: A review of intervention studies conducted with children with autism. *Focus on Autism and Other Developmental Disabilities*, 21, 45–54. <https://doi.org/10.1177/10883576060210010601>.
- Woods, J. J., & Wetherby, A. M. (2003). Early identification of and intervention for infants and toddlers who are at risk for autism spectrum disorder. *Language, Speech, and Hearing Services in Schools*, 34, 180–193. [https://doi.org/10.1044/0161-1461\(2003\)015](https://doi.org/10.1044/0161-1461(2003)015).
- *Young, H. E., Falco, R. A., & Hanita, M. (2016). Randomized, controlled trial of a comprehensive program for young students with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 46, 544–560. doi:<https://doi.org/10.1007/s10803-015-2597-0>
- Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1991). *Pre-school language scale – 3*. San Antonio, TX: The Psychological Corporation.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.