J Autism Dev Disord
DOI 10.1007/s10803-008-0596-0

ORIGINAL PAPER

Comprehensive Synthesis of Early Intensive Behavioral Interventions for Young Children with Autism Based on the UCLA Young Autism Project Model

Brian Reichow · Mark Wolery

Abstract A 3-part comprehensive synthesis of the early intensive behavioral intervention (EIBI) for young children with autism based on the University of California at Los Angeles Young Autism Project method (Lovaas in Journal of Consulting and Clinical Psychology, 55, 3–9, 1987) is presented. The three components of the synthesis were: (a) descriptive analyses, (b) effect size analyses, and (c) a meta-analysis. The findings suggest EIBI is an effective treatment, on average, for children with autism. The conditions under which this finding applies and the limitations and cautions that must be taken when interpreting the results are discussed within the contextual findings of the moderator analyses conducted in the meta-analysis.

Keywords Autism · Early intensive behavioral intervention · Applied behavior analysis · Lovaas

Introduction

Recently, the Committee on Educational Interventions for Children with Autism of the National Research Council (NRC) reviewed ten comprehensive intervention programs for young children with autism (Lord et al. 2001). Some of these programs were based on applied behavior analysis, which is a method that has been used to treat children with autism for many years. Recent survey data suggest interventions based on applied behavior analysis are some of the most frequently used interventions in autism (Green et al. 2006; Stahmer et al. 2005).

Many of the programs had supporting empirical evidence, but the NRC did not recommend a single program and cited a need for more research on them (Lord et al. 2001). Instead, consensus guidelines were listed stating children with autism should receive a comprehensive intervention program beginning as soon as they are diagnosed. The program should (a) address the individual's unique deficit areas, (b) use low teacher to student ratios, (c) include a family component, (d) be provided for at least 20–25 h per week, and (e) conduct ongoing assessment and revision of intervention goals and objectives (Lord et al.). Similar guidelines have been recommended by others (Dawson and Osterling 1997; Iovannone et al. 2003; Volkmar et al. 1999) and are generally consistent with recommended practices in early intervention (Sandall et al. 2005).

One comprehensive intervention program reviewed by the NRC (Lord et al. 2001) was early intensive behavioral intervention (EIBI) based on the University of California at Los Angeles Young Autism Project model (UCLA YAP; Lovaas 1981, 1987, 2003). This program was an intensive home-based program using the manual published by Lovaas (1981). The program typically lasted at least 2 years and involved upwards to 40 h of therapy each week. The first results from the program were noteworthy; Lovaas (1987) reported an average difference of 31 points on IQ tests between the treatment and control group, and classified 9 of 19 (47%) participants as having achieved recovery (defined as post-intervention IQ in the normal range—i.e., >85—and successful completion of first grade in a regular education classroom or unassisted placement in a regular education setting). This study, and the claims made by Lovaas (i.e., recovery) caused much debate among

B. Reichow (✉) · M. Wolery
Department of Special Education, Vanderbilt University, Peabody Box 228, Nashville 37203, TN, USA
e-mail: brian.reichow@vanderbilt.edu

researchers. Criticisms focused on methodological limitations including assignment to groups, non-uniform assessment protocol, and selection bias (e.g., Gresham and MacMillan 1998; Mundy 1993; Schopler et al. 1989). Critics often cited the need for additional replications.

Since the 1987 study, replications have occurred including those conducted as part of the National Institute of Mental Health Multi-Site Young Autism Project [MYAP, and independent replications (e.g., Birnbrauer and Leach 1993; Anderson et al. 1987)]. The replications have attempted to address methodological criticisms levied against the original study and have incorporated stronger methods including random assignment to groups (Sallows and Graupner 2005; Smith et al. 2000). Variations of the original intervention protocol also have been examined, including examination of home-based EIBI (Sheinkopf and Siegel 1998), community-based EIBI (Magiati et al. 2007), school-based EIBI (Eikeseth et al. 2007; Eldevik et al. 2006) and parent-managed EIBI (Bibby et al. 2001; Sallows and Graupner 2005).

The purpose of this paper is to provide a comprehensive synthesis of the studies on EIBI. This synthesis includes an examination of the characteristics of the experimental methods, participants, and intervention program (i.e., EIBI), as well an analysis of the effects of EIBI on participants (e.g., outcome data). To accommodate both descriptive and statistical analyses, this synthesis was conducted on multiple levels: (a) descriptive analysis, (b) effect size analyses, and (c) a meta-analysis.

Method

Study Selection

The selection of studies for this review involved seven inclusion criteria: (a) study specified the EIBI was based on the UCLA YAP model by describing the study as a replication of Lovaas (1987), citing intervention techniques and/or curriculum based on one of the Lovaas manuals (Lovaas 1981, 2003), reference to funding from the MYAP, and/or through personal communication with experts who worked with Lovaas on the UCLA YAP or directed MYAP replication sites (J. Wynn, October 9, 2007; M. Amerine-Dickens, March 5, 2007; T. Smith, March 5, 2007, personal communication); (b) participants had diagnoses of autistic disorder, autism spectrum disorder (ASD), pervasive developmental disorder (PDD), or pervasive developmental disorder not otherwise specified (PDD-NOS); (c) participant samples receiving EIBI treatment had a mean chronological age less than 84 months at the beginning of treatment; (d) mean duration of EIBI was greater than or equal to 12 months; (e) at least one child outcome measure was reported; (f) experimental research designs (e.g., pre-test/post-test multiple-group design) or quasi-experimental research designs (i.e., nonequivalent control group design, one-group pre-test/post-test design) were used (Campbell and Stanley 1963); and (g) publication in English in a peer-reviewed journal. A four-step literature search was conducted in the following order: (a) electronic database search, (b) review of references from review articles on comprehensive early intervention programs for children with autism and eligible reports, (c) hand search of selected journals, and (d) expert contact.

Fourteen research reports were located meeting all inclusion criteria and are shown in Table 1. Two reports, Lovaas (1987) and McEachin et al. (1993) used the same participants. It was therefore decided to limit the reports such that each individual (participant of a study) only contributed one result to the synthesis. The Lovaas (1987) report was used because the data were more consistent with other studies. In the Sallows and Graupner (2005) study, two arrangements of EIBI were compared (clinic-coordinated EIBI and parent-coordinated EIBI). In summary, data of 14 samples from 13 research reports were analyzed.[1]

Coding of Study Reports

The study characteristics and outcome data were coded using a manual and forms created for this synthesis. Three study level characteristics (research methods, participant characteristics, and intervention characteristics) were defined and coded to provide information about each study. Outcome data were coded for both samples receiving EIBI and for comparisons between groups receiving EIBI and non-EIBI groups. All coded data (including effect sizes) were obtained directly from the study reports or via contact with a study researcher.

Interobserver agreement (IOA) was assessed on 4 of 14 samples (29%) for the coding of study reports by two independent recorders. IOA was calculated as the product of the quotient of agreements by disagreements and 100. The range of IOA by sample was 85.5–93%. The mean IOA for the four samples reviewed was 91.6%.

Descriptive Analysis

Methodological Characteristics

To assess the influence experimental methods had on study outcomes, five methodological areas were analyzed. First,

[1] Unless otherwise noted, analyses were done using the data from the 14 distinct samples [i.e., the data from the two groups of the Sallows and Graupner (2005) study were treated separately]. When analyses were done using the data on the 13 studies, it is indicated as such.

Table 1 Methodological characteristics of studies

Study	Rigor	Design	Group assignment	Procedural fidelity			Measurement constructs by timing of measurement					
				Adherence	Differentiation	Competence	IQ	AB	Lang	AP	Psy	DR
Lovaas (1987)	Adequate	Quasi-experimental prospective multiple-group comparison	Therapist availability	Indirect measures, treatment manual	Not reported	Indirect measures	Pre/post		Pre	Post	Pre	Post
Anderson et al. (1987)	Weak	Quasi-experimental prospective one-group pre/post design	Parent selection	Direct measures, treatment manual	Not applicable (one-group study)	Direct measures	Pre/post	Pre/post	Pre	Post	Pre	Post
Birnbrauer and Leach (1993)	Weak	Quasi-experimental prospective multiple-group comparison	Parent selection	Indirect measures, treatment manual	Indirect measures	Not reported	Pre/post	Pre/post	Pre/post		Pre/post	
Smith et al. (1997)	Adequate	Quasi-experimental retrospective multiple-group comparison	Therapist availability	Indirect measures, treatment manual	Not reported	Indirect measures	Pre/post	Pre/post	Pre		Pre/post	Post
Sheinkopf and Siegel (1998)	Weak	Quasi-experimental retrospective multiple-group comparison	Parent selection	Indirect measures, treatment manual	Not reported	Not reported	Pre/post			Post	Pre/post	
Smith et al. (2000)	Strong	Experimental multiple-group comparison	Random assignment	Indirect measures, treatment manual	Not reported	Indirect measures	Pre/post	Pre/post	Pre/post	Post	Pre/post	Post
Bibby et al. (2001)	Weak	Quasi-experimental retrospective one-group pre/post design	Parent selection	Indirect measures, treatment manual	Not applicable (one-group study)	Indirect measures	Pre/post	Pre/post	Pre		Pre/post	Post
Boyd and Corley (2001)	Weak	Quasi-experimental retrospective one-group pre/post design	Parent selection	Indirect measures, treatment manual	Not applicable (one-group study)	Direct measures			Pre	Post	Pre/post	Post
Sallows and Graupner (2005)	Strong	Experimental multiple-group comparison	Random assignment	Indirect measures, treatment manual	Indirect measures	Direct measures	Pre/post	Pre/post	Pre/post	Post	Pre/post	Post
Cohen et al. (2006)	Strong	Quasi-experimental prospective multiple-group comparison	Parent selection	Indirect measures, treatment manual	Indirect measures	Indirect measures	Pre/post	Pre/post	Pre/post	Post	Pre	Post
Eldevik et al. (2006)	Adequate	Quasi-experimental retrospective multiple-group comparison	Parent selection	Indirect measures, treatment manual	Indirect measures	Indirect measures	Pre/post	Pre/post	Pre/post		Pre/post	
Eikeseth et al. (2007)	Adequate	Quasi-experimental prospective multiple-group comparison	Therapist availability	Indirect measures, treatment manual	Indirect measures	Indirect measures	Pre/post	Pre/post	Pre	Post	Pre/post	
Magiati et al. (2007)	Adequate	Quasi-experimental prospective multiple-group comparison	Parent selection	Indirect measures, treatment manual	Indirect measures	Not reported	Pre/post	Pre/post	Pre/post	Post	Pre/post	

AB adaptive behavior, *Lang* expressive and receptive language, *AP* academic placement, *Psy* psychopathology, *DR* diagnostic recovery

an overall rating of experimental rigor was obtained using the Evaluative Method for Determining Evidence-Based Practices in Autism (Reichow et al. in press). Second, the study design was categorized (i.e., experimental multiple-group comparison, quasi-experimental prospective multiple-group comparison, quasi-experimental retrospective multiple-group comparison, quasi-experimental prospective one-group pre/post design, or quasi-experimental retrospective one-group pre/post design). Third, the method used for group assignment was categorized into three groups (random assignment, therapist availability, and parent selection).

Fourth, procedural fidelity (Billingsly et al. 1980) was analyzed using the conceptual systems of treatment integrity proposed by Perepletchikova and Kazdin (2005) and Gresham (2005). Perepletchikova and Kazdin (2005) defined three components of treatment integrity. Fidelity of treatment adherence was defined as evidence the characteristics of treatment were delivered consistently as planned across and within participants of a sample. Treatment differentiation was defined as evidence the groups of a comparative study received different levels of the treatment package. Therapist competence was defined as evidence of therapist training and/or evaluation of therapist performance. In a response to these components, Gresham (2005) outlined three methods of measuring treatment integrity: (a) direct measures, (b) indirect measures, and (c) manualized treatments. The final methodological characteristic was the measures used, which was categorized into six constructs (IQ, adaptive behavior, language, academic placement, psychopathology, and diagnostic reclassification).

Participant Characteristics

Participant characteristics were assessed by examining the pre-treatment assessments on six variables: (a) diagnosis, (b) chronological age, (c) IQ, (d) adaptive behavior, (e) language, and (f) other treatments received. These data were used to illustrate differences between samples and as moderator variables for the meta-analysis.

Intervention Characteristics

Nine intervention characteristics were identified for this review. Three intervention characteristics pertained to the intensity of the intervention. Intervention density was defined as the total number of hours per week participants received EIBI. Intervention duration was defined as the total number of months each participant received EIBI. The total hours of therapy was calculated by multiplying the product of intervention density and duration by 4.3 (converter for months to weeks). Not all studies reported the mean intervention density and/or duration. When the mean data for density and/or duration were not provided, an estimated value was determined from information in the study report and used for all subsequent analyses.

Three intervention characteristics described the organization of intervention services. The model of supervisor training was a dichotomous variable; studies were either categorized as being consistent with the UCLA/MYAP training protocol, including an internship at an affiliated clinic site (i.e., UCLA or MYAP), or studies were categorized as using other training models (e.g., inservice, on-the-job, workshop-based). The second organizational intervention characteristic categorized the type of service coordination model as being clinic-coordinated, community-coordinated, or parent-coordinated. Parental role was defined by the type of involvement expected for each participant's parents (usually mother). These included conducting therapy, service-coordination, and assisting therapists.

The remaining three intervention characteristics describe aspects of the EIBI therapy. The educational and/or training qualifications of therapist were categorized as parent, undergraduate college student, lay person, or paraprofessional. The location of therapy was coded as the location intervention occurred across the entire intervention period and included three categories (home, school, community). Finally, the use of physical aversives was recorded as occurring, not occurring, or not reported for each sample.

Outcome Data

Descriptive analyses were conducted on constructs with no pre-intervention assessment (academic placement, diagnostic reclassification) and for constructs using many different measures (psychopathology). Because the calculation of an effect size was not appropriate for these constructs, they were analyzed using descriptive statistics. These analyses were conducted on the sample data,[2] thus the results reflect the changes within a sample without reference to a control group. The data for placement were analyzed by reporting the range of the percentage of participants from each sample in regular education classrooms and other educational settings (e.g., special education settings, aphasic classrooms). The data for psychopathology were analyzed by comparing the mean scores of the pre- and post-intervention assessments for each sample, which were then categorized by the type of change. The data for diagnostic classification were analyzed by reporting the

[2] Because the data on these measures for each sample of the Sallows and Graupner (2005) study were not reported separately for these constructs, they were analyzed as an aggregated sample.

range of the percentage of participants meeting Lovaas' (1987) criteria of recovery (i.e., post-intervention IQ in the normal range—i.e., greater than 85—and successful completion of first grade in a regular education classroom or unassisted placement in a regular education setting) for each sample.

Effect Size Analyses

Effect sizes were calculated for the outcome data from the constructs of IQ, adaptive behavior, expressive language, and receptive language. Two types of effect sizes were used: The standardized mean change effect size and the standardized mean difference effect size. The formulae for these are shown in Table 2. Three steps were taken to help ensure the most conservative effect sizes were calculated. First, effect sizes were calculated only when the data necessary for its calculation were available. If a sample or study was missing the necessary data for the calculation of an effect size, no effect size was calculated for that study. Hence, no data were extrapolated or interpolated for the calculation of effect sizes. Second, Hedge's g (Hedges and Olkin 1985) was used as the effect size metric, which calculates a more conservative (i.e., smaller) estimate of the effect size than Glass' Δ or Cohen's d (Grissom and Kim 2005). Finally, because effect sizes based on small samples are known to be biased (Lipsey and Wilson 2001), effect sizes were multiplied by the small sample correction factor (Hedges and Olkin 1985).

The first effect size analyses were calculated using the standardized mean change effect size and examined the difference between the average gains made by distinct samples. This comparison showed the absolute difference within a sample without regard to a comparison or control group. For these analyses, the effect sizes were analyzed with reference to the research report rigor rating (i.e., strong, adequate, and weak) of the study containing the sample.

For the ten studies using between-group designs, the standardized mean difference effect size (g_d) was used (shown in Table 2). This effect size showed the magnitude of difference between the group receiving EIBI and the

Table 2 Formulas used in the statistical analyses

Formula	Equation	Where
Standardized mean change effect size with small sample adjustment	$g_c = d\ \{1 - [3/(4 \times df - 1)]\}$	$d = (Y_2 - Y_1)/s_p^2$ df = Degrees of freedom Y_1 = pre-treatment mean Y_2 = post-treatment mean $s_p^2 = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/n_1 + n_2 - 2]}$ n_1 = Number of participants at pre-treatment n_2 = Number of participants at post-treatment s_1^2 = Pre-treatment variance s_2^2 = Post-treatment variance
Standardized mean difference effect size with small sample adjustment	$g_d = d\ \{1 - [3/(4 \times df - 1)]\}$	$d = (Y_2 - Y_1)/s_p^2$ df = Degrees of freedom Y_1 = mean for comparison group Y_2 = mean for EIBI group $s_p^2 = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/n_1 + n_2 - 2]}$ n_1 = Number of participants in comparison group n_2 = Number of participants in EIBI group s_1^2 = Variance of comparison group s_2^2 = Variance of EIBI group
Small sample correction	$1 - [3/(4 \times df - 1)]$	df = Degrees of freedom
Mean effect size	$\overline{ES} = \Sigma ES_i \times w_i/\Sigma w_i$	$ES_i = g_c$ for studies $i = 1$ to k w_i = Inverse variance weight = $1/SE^2$
Q-statistic	$Q = \Sigma w_i\ (ES_i - \overline{ES})^2$	w_i = Inverse variance weight = $1/SE^2$ $ES_i = g_c$ for studies $i = 1$to k $\overline{ES}$ = standardized mean effect size
Proportion of variance accounted for by between-study variance	$I^2 = Q/(df - 1)/(Q/df)$	$Q = \Sigma w_i\ (ES_i - \overline{ES})^2$ df = Degrees of freedom = $(n - 1)$

comparison group. For the analyses of the comparative studies, effect sizes for each construct were analyzed with reference to the characteristics of the comparison group. Three types of comparison groups were used across studies. Two studies (Lovaas 1987; Smith et al. 1997) compared intensity of behavioral intervention (i.e., high intensity vs. low intensity). Six studies (Birnbrauer and Leach 1993; Cohen et al. 2006; Eldevik et al. 2006; Eikeseth et al. 2007; Magiati et al. 2007; Sheinkopf and Siegel 1998) compared EIBI with other treatments (e.g., treatment as usual, eclectic treatment, specialist nursery school). The remaining comparative studies (Sallows and Graupner 2005; Smith et al. 2000) examined two service coordination models (clinic- vs. parent-coordination) of EIBI.

Meta-analysis

Meta-analytic techniques (Lipsey and Wilson 2001) were used to conduct a quantitative analysis of the samples for which data from the within-group analysis of changes in IQ were available [i.e., standardized mean change effect size with small sample adjustment (g_c)]. Four statistical analyses were conducted as a part of the meta-analysis: (a) mean effect size, (b) publication bias, (c) homogeneity of the data, and (d) analyses of moderator variables. The calculations and manipulation of the data for these analyses are explained below and the formulae used in the analyses are shown in Table 2.

Mean Effect Size

The calculation of the mean effect size involved three steps. First, the effect size and variance were calculated for each sample. Second, a weight (indicated as w_i in Table 2) for each sample was determined by taking the inverse of its variance. These weights were used in all statistical analyses of the meta-analysis, and provide an estimate of their precision (Lipsey and Wilson 2001). Finally, the mean effect size (indicated as $\overline{ES}$ in Table 2) was calculated by summing the product of each effect size by its weight and dividing by the sum of the weights.

Publication Bias

The trim and fill method of Duval and Tweedie (2000) was used to detect and estimate publication bias. The first step in the trim and fill method is visual inspection of a funnel plot graphic display of the effect sizes by standard errors of each sample. If visual inspection suggests "missing studies", further statistical analyses are conducted. The subsequent statistical analyses estimate the effect size(s) of the "missing studies" and used this figure to calculate an estimate of the "true effect size" (i.e., what the effect size would be if the "missing studies" were used). Although this method derives a new estimate of effect (i.e., "true effect size"), Duval (2005) cautions against using this new estimate to adjust the results. Rather, Duval suggests, "[the trim and fill] method should be used primarily as a form of sensitivity analysis, to assess the potential impact of missing studies on the meta-analysis," (p. 131).

Homogeneity of the Data

Two statistical analyses were done to test the homogeneity of the data. First, the weighted mean effect sizes were examined using the Q-statistic (Hedges and Olkin 1985). The Q-statistic provides a test of statistical significance indicating whether the differences (variance) in effect sizes are due to subject-level sampling error alone (i.e., homogeneous) or if the differences are greater than would be expected by subject-level sampling alone (i.e., heterogeneous). A statistically significant Q-statistic indicates heterogeneity between studies included in the analysis, and can be used to justify performing cautious moderator analyses (Lipsey and Wilson 2001). Because recent criticism has been raised about the validity of the Q-statistic as a test of homogeneity in meta-analyses (Huedo-Medina et al. 2006), a second test of homogeneity, I^2, was conducted. I^2 estimated the proportion of the variance that was between-studies variance. Like the Q-statistic, I^2 can be analyzed with reference to the appropriateness of moderator analyses.

Moderator Analyses

The effects of selected moderator variables on the weighted mean effect size were examined using an analog to the analysis of variance for categorical variables and a modified weighted regression procedure for continuous and dichotomous variables. To limit the possibility of Type I error, only moderator variables with a priori hypothesized relations to outcome were used in these analyses. All analyses were conducted using the methods of moments random effects model with the SPSS macros provided by Lipsey and Wilson (2001). In the analyses, the effect size was used as the dependent variable and the selected moderator variables as the independent variable.

Findings and Discussion of Descriptive Analyses

Methodological Characteristics

The assessment of research report rigor was conducted to provide an overall assessment of the methodological

qualities of each study. Three studies (23%) received the highest rating (strong), five (38%) received the middle rating (adequate), and five studies (38%) received the lowest rating (weak; see Table 1). In comparison with recent reviews on the state of research in autism (e.g., Lord et al. 2001; Reichow et al. 2007), the research report rigor ratings of the studies reviewed were higher than would be expected. Although the overall rigor appeared to be acceptable, the studies had methodological limitations. Limitations of the remaining four methodological variables are discussed with reference to the results of the descriptive analyses shown in Table 1.

Experimental Design and Assignment to Groups

In the studies reviewed, 2 of 13 used true experimental designs. The most frequently employed design was the quasi-experimental prospective multiple-group comparison—used in 6 of 13 studies. The remaining designs, quasi-experimental prospective pre/post design, quasi-experimental retrospective multiple-group comparison, and quasi-experimental retrospective pre/post design were used in 1, 3, and 1 studies, respectively. With respect to the timing of data collection, 8 of 13 studies used prospective designs (participants were enrolled before intervention began and data were collected at the conclusion of intervention) and 5 studies used retrospective data collection (participants were located after they had concluded or were in the midst of intervention).

Although different designs can demonstrate intervention effectiveness (e.g., single subject research designs), true experimental designs using random assignment to groups is the group design demonstrating the strongest evidence of an intervention's effectiveness (Campbell and Stanley 1963). Of the comparative studies, five used parent selection and three used therapist availability to assign participants to groups; only two studies used random assignment. Thus, the threat of non-equivalent groups is likely present. In studies using random assignment (Sallows and Graupner 2005; Smith et al. 2000), group equivalence was claimed across measured variables, demonstrating the utility of such designs. Random assignment should decrease the threat of participant selection bias if group size is large, which cannot be ruled out when parents choose the condition their children receive. Future studies should employ random assignment to groups, although random assignment with small group sizes does not ensure group equivalence—especially with heterogeneous populations.

An issue related to the method of assigning participants to groups is the choice of the comparison condition (for a discussion on the formation of comparison groups in research on individuals with autism, see Burack et al. 2004). In the comparative studies reviewed, little is known about the comparison conditions, and little uniformity appears to exist across studies. These groups often lacked standardization within the group, were poorly defined, had no measures of procedural fidelity, and had no data on whether participants received supplemental treatments. Some studies of this synthesis described the comparison group as eclectic (e.g., Cohen et al. 2006; Eikeseth et al. 2007; Eldevik et al. 2006). By definition, eclectic treatments can vary across participants within a group, thus creating variability within the comparison group. This variability creates a situation in which the treatment effect might be over- or underestimated, and does not create a situation where treatment components can be compared. Well defined comparison groups are needed if sound generalizations about the effectiveness of EIBI are to be made. Although a no-treatment control group may not be possible with young children (Lord et al. 2005), the choice of an adequate comparison group is important and merits greater attention in future research. Future research should clearly define and quantify the treatment(s) provided to comparison groups.

Procedural Fidelity

The analysis of procedural fidelity revealed mixed results (Table 1). All samples employed procedures and/or measures to ensure or document treatment adherence; 1 of 13 studies used direct measures, 12 studies used indirect measures, and 13 studies used a treatment manual (Lovaas 1981, 2003). Treatment differentiation could only be measured for the ten comparative studies; six of these measured treatment differentiation using indirect measures. Therapist competence was measured in 10 of 13 studies; three studies assessed therapist competence using direct measures and seven studies used indirect measures.

Although all studies contained a reference to a treatment manual (Lovaas 1981, 2003), no verification existed on the use of the manuals. Some studies contained elements to maintain procedural fidelity, but only one study (Anderson et al. 1987) reported direct measures of treatment adherence and no study reported measuring therapy implementation at a level sufficient to draw definitive conclusions about quality and similarity of therapy across participants or within a participant across therapists. Data from studies examining the implementation of EIBI have shown some parents and therapists have difficulty achieving high levels of procedural fidelity (Johnson and Hastings 2002; Symes et al. 2006). In behavioral research, questionable fidelity can limit research conclusions (Bellg et al. 2004). Therefore, future studies should measure procedural fidelity directly across participants, therapists, and conditions. Once methods of measuring procedural fidelity have

been defined and acceptable levels achieved, parametric analyses of different levels of fidelity should be conducted to determine the level of precision necessary for EIBI to be effective.

Measures

Multiple constructs were assessed with multiple measures and measurement methods. The studies measuring each construct are shown in Table 1. Four constructs were measured using standardized tests or assessments. IQ was measured in 12 of 13 studies. Adaptive behavior was measured using the Vineland Adaptive Behavior Scales (Sparrow et al. 1984) in 9 of 13 studies. Language (expressive and receptive language) was measured in 6 of 13 studies using a variety of standardized tests. The remaining constructs were measured using questionnaires or researcher devised measures. Academic placement was the participant's educational placement after treatment (typically after first grade) and was measured in 9 of 13 studies. Psychopathology was a quantification of the degree of each child's impairment (severity or number of autistic symptoms, verbal skills), and was measured in 13 of 13 studies. Severity of autistic symptoms was measured using a quantification of symptoms for autistic disorder or parent interviews and questionnaires. Diagnostic reclassification was defined as the percentage of participants meeting the recovery criteria proposed by Lovaas (1987)—post-intervention IQ in the normal range (i.e., greater than 85) and successful completion of first grade in a regular education classroom or unassisted placement in a regular education setting—and was reported in 8 of 13 studies.

Using multiple types of measures and multiple measurement methods are desirable and should continue in future EIBI studies; however, limitations existed. The original evaluation of EIBI (Lovaas 1987) used limited outcome measures. Although broader measures have been used in some subsequent studies, outcomes were measured narrowly and evidence of functioning in the natural environment was absent. IQ, as measured by standardized tests, was the primary outcome across studies. Its utility can be questioned because the used standardized tests were normed with children who were typically developing, it is not a direct measure of functional outcomes, and intelligence is not a diagnostic marker of autism. Furthermore, multiple tests were used within groups and across individuals over time without reporting the correlations of the scores under such variation. Because of the challenges children with autism present in standardized testing situations (Wolery and Garfinkle 2002), future research on EIBI should supplement standardized measures with observational data of children's functional performance on key variables (e.g., communicative and social behaviors) in natural settings. Researchers have suggested the formation of a standard battery of assessments to evaluate intervention programs (e.g., Charman and Howlin 2003; Schreibman 2000). If such a battery was to be developed, it should include multiple types of measures collected across settings and observers. Examples of measures that should be considered include measures of intelligence, developmental abilities across domains, adaptive behavior, communication, psychopathology (severity of autism), play, social skills, challenging behaviors, rigidity, and other behaviors characteristic of children with autism. After forming an assessment battery, the psychometrics of the measures should be evaluated as they relate to children with autism (e.g., the stability of the measures over time, the validity of the measures as an indication of change in children with autism). Further, the fidelity with which those measures are implemented should be measured and reported.

Two constructs debated since they were used by Lovaas (1987) are academic placement and diagnostic reclassification (i.e., recovery). Both measures were frequently used across the reviewed studies. These measures are inherently intertwined—a certain level of academic placement must be achieved for a participant to be considered to have had a diagnostic reclassification. However, neither measure truly provides the evidence one might infer from the face value of its name. Numerous factors other than child abilities (e.g., district policies, parents' advocacy) can control decisions about children's academic placement (Wolery and Garfinkle 2002). Thus, academic placement is a flawed outcome measure and should not be used as an indication of intervention effectiveness in future studies. It can be reported for informational purposes, but not as an outcome measure. While appealing, diagnostic reclassification requires measures related to the diagnostic criteria for autism. If assessed, diagnostic evaluations should include diagnostic instruments for autism (e.g., Autism Diagnostic Observation Schedule; Lord et al. 1999, Autism Diagnostic Interview-Revised; Rutter et al. 2003). Furthermore, diagnostic assessments should be conducted by qualified individuals who are blind to participants' earlier diagnosis and experimental group membership.

Participant Characteristics

The participants in the studies were assessed on six pre-intervention characteristics (a) diagnosis, (b) chronological age, (c) IQ, (d) adaptive behavior, (e) language, and (f) receipt of additional treatments. Descriptive statistics for participant characteristics of samples receiving EIBI are provided in Table 3. Across studies, 373 children with autism, ASD, PDD, or PDD-NOS participated in the studies. Of the participants, 251 (67%) received EIBI and

Table 3 Description of EIBI participant pre-intervention characteristics by sample

Participant characteristic	Sample ($n = 14$)		
	Identification	*n*	%
Diagnostic breakdown			
100% Autism	a, b, c, j, k, m, n	7	50
>50% Autism (<50% ASD, PDD, or PDD-NOS)	e, g, h, l, p	5	36
<50% Autism (>50% ASD, PDD, or PDD-NOS)	d, f	2	14
Mean pre-treatment chronological age (months)			
<36	a, e, h, j, k, l	6	43
36–42	c, d, f, p	4	29
42–48	b, g	2	14
>48	m, n	2	14
Mean pre-treatment IQ ratio			
<40	d	1	7
40–55	a, c, f, g, j, k, m	7	50
55–70	b, e, g, l, n	4	29
>70	p	1	7
Not measured	h	1	7
Mean pre-treatment composite adaptive behavior (scale score)			
40–54	d, m	2	14
55–70	f, g, j, l, n, p	6	43
>70	k	1	7
Not measured or scale score not reported	a, b, c, e, h	5	36
Mean pre-treatment expressive language (scale score)			
<40	m	1	7
40–54	j, k, l, n	4	29
Not measured or scale score not reported	a, b, c, d, e, f, g, h, p	9	64
Mean pre-treatment receptive language (scale score)			
<40	j, k, m	2	14
40–54	l, n	2	14
Not measured or scale score not reported	a, b, c, d, e, f, g, h, p	9	64
Percentage of participants with verbal language skills at pre-treatment (%)			
100	–	0	0
51–99	b, h, k	3	21
1–50	a, c, g, j, m	5	36
0	d	1	7
Not measured or not reported by group	e, f, l, n, p	5	36
Reported data on other treatments received by participants			
Yes	e, g, h, j, k, p	6	43
No	a, b, c, d, f, l, m, n	8	57

Key: a—Lovaas (1987); b—Anderson et al. (1987); c—Birnbrauer and Leach (1993); d—Smith et al. (1997); e—Sheinkopf and Siegel (1998); f—Smith et al. (2000); g—Bibby et al. (2001); h—Boyd and Corley (2001); j—Sallows and Graupner (2005) clinic-coordinated group; k—Sallows and Graupner (2005) parent-coordinated group; l—Cohen et al. (2006); m—Eldevik et al. (2006); n—Eikeseth et al. (2007); p—Magiati et al., (2007)

122 (33%) were in non-EIBI comparison samples. Of the 251 participants receiving EIBI, 216 (86%) had a diagnosis of autism and 35 (14%) had a diagnosis of ASD, PDD, PDD-NOS. Of the 122 participants in the non-EIBI comparison samples, 96 (79%) had a diagnosis of autism and 26 (21%) had a diagnosis of ASD, PDD, or PDD-NOS.

Overall, most of the samples were comprised of children with autism less than 42 months old. The samples had a wide range of mean IQ scores, 28 (Smith et al. 1997) to 83 (Magiati et al. 2007). The participants' levels of adaptive behavior was typically 2–3 standard deviations below the mean, and the participants' levels of expressive and

receptive language were typically 3–4 standard deviations or more below the mean. When reported (six samples), some samples reported greater than 80% of the participants were receiving supplemental treatments while participating in the research studies.

Given past criticisms concerning the representativeness of the participants of EIBI studies (e.g., Mundy 1993; Schopler et al. 1989), a comparison of the homogeneity of samples within studies and the representativeness of the samples across studies to children with autism were desired. However, research on the sensitivity of the diagnostic criteria shows many individuals diagnosed under one edition of the *Diagnostic and Statistical Manual* (American Psychiatric Association 1980, 1987, 1994, 2000) would not be diagnosed with another edition, and vice-versa (Volkmar 1998). Because of the changing definition of autism between studies and the heterogeneity of children with autism, an overall comparison was not conducted. Nonetheless, the children in the reviewed studies on average had impaired language and adaptive behavior at pre-intervention; they were not exclusively children with mild impairments.

Intervention Characteristics

Intervention characteristics have been hypothesized as being responsible for differences between studies reporting larger and lesser increases on outcome measures (Lovaas 2003). Differences existed in all characteristics of intervention across studies, which are shown in Table 4. Three variables quantified the intensity of the intervention. The range of intervention density was 18.7 (Birnbrauer and Leach 1993) to 40 h per week (Lovaas 1987); 8 of 14 samples had a density of at least 30 h per week. The range of intervention duration was 12 (Anderson et al., 1987) to 48 months (Sallows and Graupner 2005); 9 of 14 samples had a duration of at least 24 months. The range of the total number of hours of therapy was 774 (Anderson et al. 1987) to 7,793 (clinic-coordinated group, Sallows and Graupner 2005); 6 of 14 samples received at least 4,000 h of therapy.

Three variables characterized the organization of intervention services. Most samples (9) provided training to supervisory personnel using the UCLA YAP/MYAP training model (Davis et al. 2002; Lovaas 2003). The service coordination model used most frequently across samples was the clinic-coordination model (7 of 14 samples). Although detailed information concerning the role of parents during intervention was not always provided, reportedly parents provided direct therapy services to their child in 11 of 14 samples.

Multiple therapists per participant were reportedly used with a majority of the samples. Although therapist qualifications varied, parents and undergraduate students were the most frequently used, 11 and 8 samples, respectively. All studies had therapy sessions in the home setting, and 9 of 14 samples used multiple settings. Physical aversives were not frequently employed; only two samples (Lovaas 1987; Smith et al. 2000) reported using physical aversives with any participant.

Descriptive Outcome Analysis

Descriptive analyses were conducted for the samples receiving EIBI[2] for placement, psychopathology, and diagnostic reclassification. Analyses of these data support the conclusion that EIBI is an effective intervention for many children with autism. The results from academic placement and diagnostic reclassification suggest some children will perform well in typical educational settings after intervention. The results for psychopathology suggest, on average, children present fewer or less severe autism symptoms after intervention.

Nine of 13 studies reported data for academic placement at post-intervention or after 1st grade; these data are shown in Table 5. The range of participants placed in regular education classrooms was 23% (Anderson et al. 1987; Boyd and Corley 2001) to 100% (Eldevik et al. 2006; Eikeseth et al. 2007). Across samples, 140 of 217 (65%) participants receiving EIBI had a post-intervention academic placement in regular education classrooms (includes full and partial inclusion). For the academic placement category of "other," (i.e., all placements other than regular education), 77 of 217 (35%) participants receiving EIBI had a class placement at post-intervention in the "other" category.

Post-intervention data on psychopathology was reported for 10 of 13 studies, and are shown in Table 5. These ten samples reported both pre-intervention and post-intervention data, thus, a comparison between the two scores was conducted. All ten samples reported, on average, the participant's presentation of autism was less severe after intervention.

Seven of 13 studies reported data on diagnostic reclassification (i.e., percentage of participants meeting the definition of recovery proposed by Lovaas) and are shown in Table 5. The samples receiving EIBI were compared using the recovery rate of the Lovaas (1987) study (i.e., 47%) as a benchmark for this analysis. One study, (Sallows and Graupner 2005) reported a rate of diagnostic reclassification greater than Lovaas—48% of participants across groups (clinic- and parent-coordinated EIBI) met the criteria of diagnostic reclassification. All other samples reported rates of diagnostic reclassification less than 47%. Of the samples reporting data on diagnostic reclassification, four reported no participants were reclassified. Across samples, 31 of 172 (18%) participants receiving EIBI were reported as meeting the criteria of diagnostic reclassification.

Table 4 Description of EIBI characteristics by sample

Treatment characteristic	Sample ($n = 14$)		
	Identification	*n*	%
Density (hours per week)			
<20	c	1	7
20–29	b, e, f, m, n	5	36
30–39	d, g, h, j, k, l, p	7	50
>39	a	1	7
Duration (months)			
<12	–	0	0
12–23	b, c, e, h, m	5	36
24–36	a, d, f, g, l, n, p	7	50
>36	j, k	2	14
Total hours of therapy			
<2,000	b, c, e, m	4	29
2,000–3,999	f, h, n, p	4	29
4,000–5,999	d, g, l	3	21
>6,000	a, j, k,	3	21
Training model			
UCLA	a, b, d, f, j, k, l, m, n	9	64
Other	c, e, g, h, p	5	36
Service coordination model			
Clinic	a, b, c, d, f, j, m	7	50
Community	e, h, l, n, p	5	36
Parent	g, k	2	14
Parental role			
Conduct therapy	a, b, c, e, f, j, k, l, m, n, p	11	79
Service coordination	g, k	2	14
Assist therapists	d	1	7
Not reported	h, p	2	14
Qualifications of therapist			
Parent	a, b, c, e, f, j, k, l, m, n, p	11	79
Undergraduate college student	a, b, c, d, f, j, k, p	8	57
Lay person (i.e., untrained individual)	e, g, h, l	4	29
Paraprofessional (i.e., teaching assistant)	d, e, m, n	4	29
Location of therapy			
Participant's home	a, b, c, d, e, f, g, h, j, k, l, m, n, p	14	100
School	a, f, h, j, k, l, m, n	9	64
Community	a, d, f, j, k, l, m, n	6	43
Used physical aversives			
Yes	a, f	2	14
No	b, c, e, j, k, n	6	43
Not reported	d, g, h, l, m, p	6	43

Key: a—Lovaas (1987); b—Anderson et al. (1987); c—Birnbrauer and Leach (1993); d—Smith et al. (1997); e—Sheinkopf and Siegel (1998); f—Smith et al. (2000); g—Bibby et al. (2001); h—Boyd and Corley (2001); j—Sallows and Graupner (2005) clinic coordinated group; k—Sallows and Graupner (2005) parent coordinated group; l—Cohen et al. (2006); m—Eldevik et al. (2006); n—Eikeseth et al. (2007); p—Magiati et al. (2007)

Table 5 Descriptive analysis of outcome data for participants receiving EIBI for academic placement, psychopathology, and diagnostic reclassification

Outcome	Sample (n = 13)		
	Identification	n	%
Placement			
Participants in a regular education classroom (%)			
0%	–	0	0
1–25	b, h	2	15
26–50	a, e, f	3	23
51–75	g	1	8
76–99	jk, l, p	3	23
100%	m, n	2	15
Not measured	c, d	2	15
Participants in other classroom settings (%)			
0	m, n	2	15
1–25	jk, l, p	3	23
26–50	g	1	8
51–75	a, e, f	3	23
76–99	b, h	2	15
100	–	0	0
Not measured	c, d	2	15
Psychopathology			
Average change in severity of autistic symptoms			
Less severe	c, d, e, g, h, jk, l, m, n, p	10	77
Equal severity	–	0	0
More severe	–	0	0
Not measured	a, b, f	3	23
Diagnostic Reclassification			
Percentage of participants with diagnostic reclassification (%)			
0	b, d, g, h	4	31
≤47	f, l	2	15
47	a	1	8
>47	jk	1	8
Not measured	c, e, m, n, p	5	38

Key: a—Lovaas (1987); b—Anderson et al. (1987); c—Birnbrauer and Leach (1993); d—Smith et al. (1997); e—Sheinkopf and Siegel (1998); f—Smith et al. (2000); g—Bibby et al. (2001); h—Boyd and Corley (2001); jk—Sallows and Graupner (2005) combined groups; l—Cohen et al. (2006); m—Eldevik et al. (2006); n—Eikeseth et al. (2007); p—Magiati et al. (2007)

Findings and Discussion of Effect Size Analyses

Analyses by Sample

The standardized mean change effect sizes for IQ, adaptive behavior, expressive language, and receptive language for the samples receiving EIBI are shown for each sample, organized by rigor rating, in Fig. 1. The generally positive effect sizes suggest post-intervention performance was on average higher than pre-intervention on multiple dimensions of functioning. More samples (12 of 14) had enough data to calculate effect sizes for IQ than any other measure. The range of the standardized mean change effect sizes for IQ was $g_c = -0.19$ (Magiati et al. 2007) to $g_c = 1.58$ (parent-coordinated group, Sallows and Graupner 2005). One study (Magiati et al. 2007) had a negative ES for IQ, and nine samples had a standardized mean change effect size for IQ equal to or greater than 0.50. Effect sizes for adaptive behavior were calculated for ten samples, with a range of $g_c = -0.25$ (Magiati et al. 2007) to $g_c = 0.86$ (Eikeseth et al. 2007). Five samples had a standardized mean change effect size for adaptive behavior equal to greater than or 0.50; one was positive but less than 0.50; and four were negative, signifying the individual samples had scores on average at post-intervention equal to or lower than pre-intervention. Six samples reported enough data to calculate effect sizes for expressive and receptive language measures. The range of the standardized mean change effect sizes were $g_c = 0.23$ (clinic-coordinated group, Sallows and Graupner 2005) to $g_c = 1.72$ (Smith et al. 2000) and $g_c = 0.45$ (Eldevik et al. 2006) to $g_c = 1.79$ (Smith et al. 2000) for expressive and receptive language, respectively. For each measure, five of six samples had a standardized mean change effect size greater than 0.50.

Between Groups Analyses of Comparative Studies

The standardized mean difference effect size was calculated for the ten comparative studies to analyze the differences between EIBI and comparison groups and is shown for each study, organized by the type of comparison, in Fig. 2. These effect sizes suggest children receiving EIBI made more gains than children receiving minimal behavioral intervention, eclectic treatment, or treatment as usual. While these findings were strong, the lack of adequate comparison groups and the nonrandom assignment of participants to groups limit conclusions about the superiority of EIBI to other treatments.

Comparison of Intensity of Behavioral Intervention

Two studies (Lovaas 1987; Smith et al. 1997) compared different intensity (i.e., density) of behavioral intervention. The standardized mean difference effect sizes for IQ were $g_d = 0.96$ (Smith et al. 1997) and $g_d = 1.19$ (Lovaas 1987) favoring higher density. As sown in Table 3, the pre-intervention IQ scores for the participants in Smith et al. (1997) were lower on average than those in the Lovaas study. No study comparing intensity of behavioral interventions examined adaptive behavior, expressive, or receptive language.

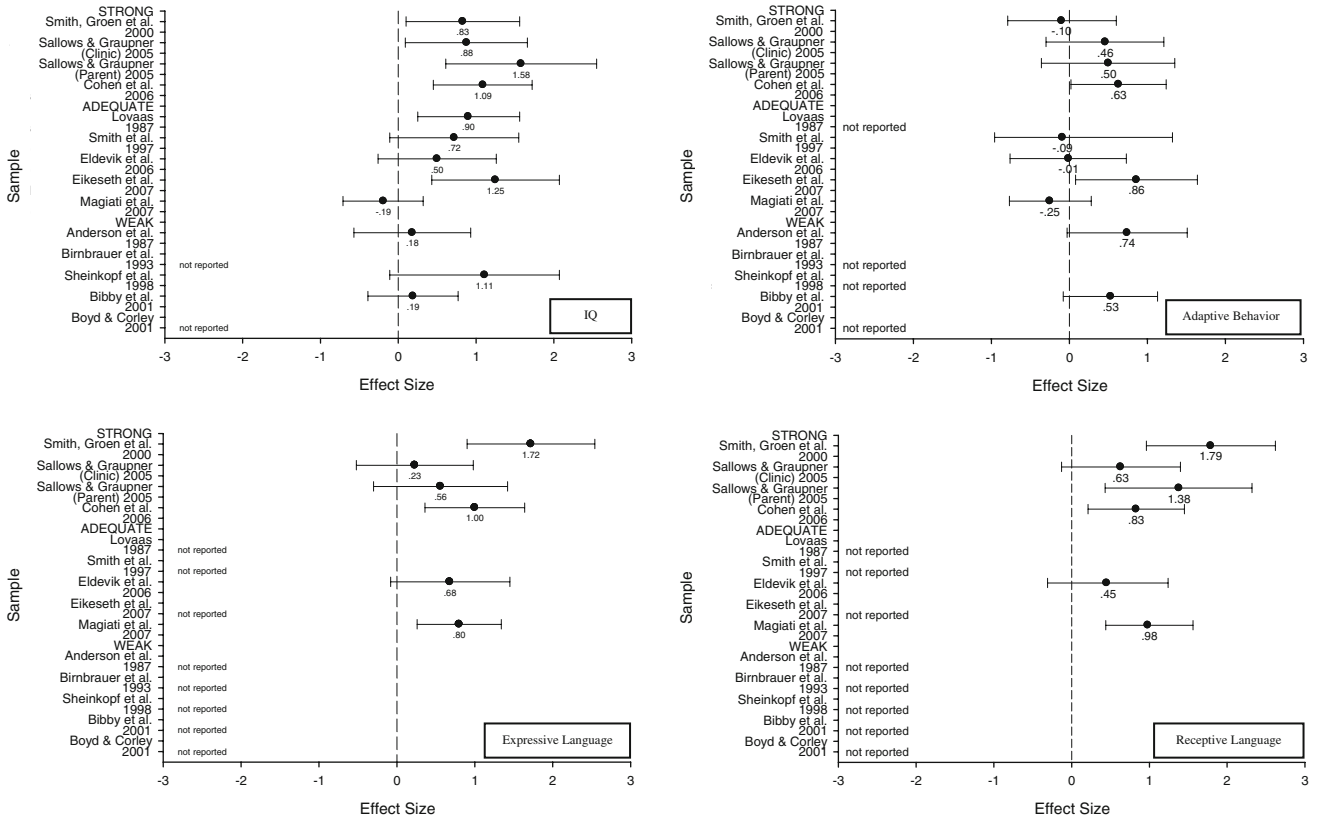

Fig. 1 Forrest plot of standardized mean change effect sizes (g_c) and 95% confidence intervals for samples ordered by rigor

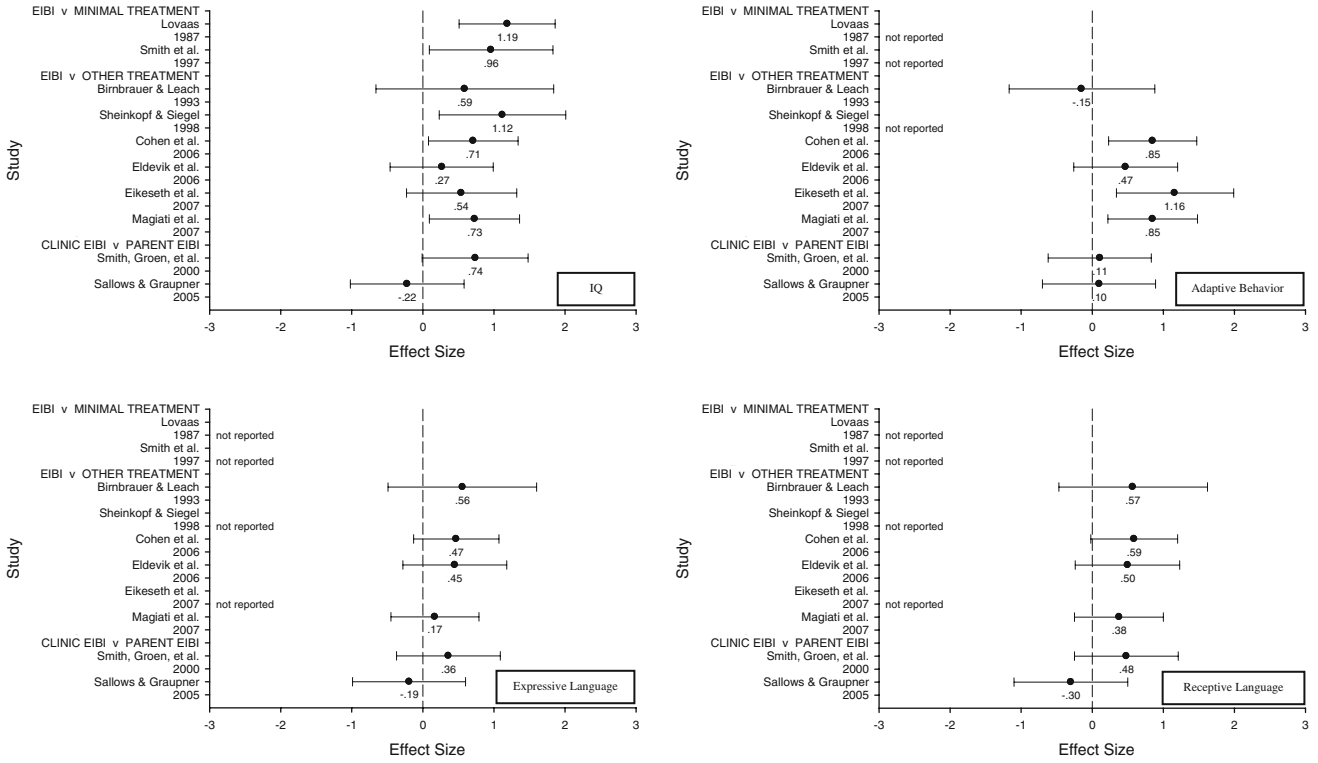

Fig. 2 Forrest plot of standardized mean difference effect sizes (g_d) and 95% confidence intervals for comparative studies ordered by type of comparison

Comparison of EIBI and Other Treatments

Six studies compared EIBI to other treatments. All six studies reported IQ data; the range of the standardized mean difference effect sizes was $g_d = 0.27$ (Eldevik et al. 2006) to $g_d = 1.12$ (Sheinkopf and Siegel 1998), all effects favored EIBI. Five of six studies comparing EIBI to other treatments had adaptive behavior data. One study (Birnbrauer and Leach 1993) had a negative standardized mean difference effect size, $g_d = -0.15$, signifying the treatment as usual group had scores at an equal or slightly higher level as the EIBI group. The remaining four studies showed the opposite effect; the groups receiving EIBI had higher scores than the comparison groups (range $g_d = 0.47 - 1.17$). Four treatment comparison studies reported data for expressive and receptive language. The same patterns were seen for both types of language; the group receiving EIBI had higher scores on both types of language than the groups receiving other treatments (range for expressive language $g_d = 0.17$–0.56, range for receptive language was $g_d = 0.38$–0.59). As shown in Fig. 2, the effect sizes for receptive language tended to be higher than expressive language.

Comparison of EIBI Coordination Models

The result for the comparison of the two coordination models revealed mixed results. In the Smith et al. (2000) study, the clinic-coordinated group had higher standardized mean difference effect sizes for IQ, expressive language, and receptive language than the parent-coordinated group. The opposite occurred for the Sallows and Graupner (2005) study, in which the parent-coordinated group had higher IQ, expressive language, and receptive language effect sizes. For adaptive behavior, both studies reported slightly higher standardized mean difference effect sizes for the clinic-coordinated group.

Findings and Discussion of Meta-Analysis

Meta-analytic techniques (Lipsey and Wilson 2001) provide a quantitative method to determine the average effects across studies. A meta-analysis was conducted on the 12 samples reporting enough data to compute the standardized mean change effect size for IQ. This analysis was conducted on the sample data because the comparison groups were not similar across studies. Although the standardized mean change effect size has been shown to inflate effect sizes (Morris 2000), the relations should remain.

Mean Effect Size

Using a random effects model, the mean effect size was 0.69, which was statistically significant ($p < 0.001$). This is a large effect, and suggests EIBI is, on average, an effective intervention for increasing IQ scores for children with autism.

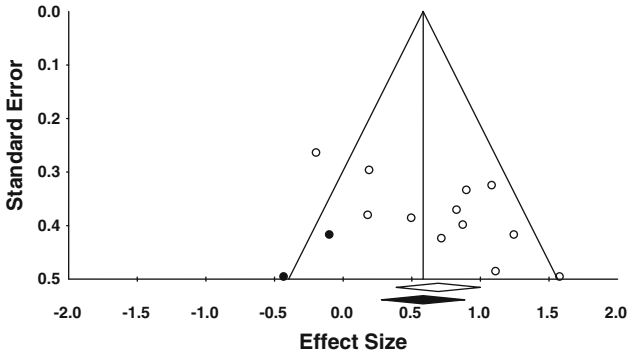

Fig. 3 Funnel plot of effect size by standard error for samples (*open circles* indicate studies in meta-analysis, *closed circles* indicate "missing studies" suggested by publication bias analysis, *open diamond* indicates mean effect size and 95% CI for meta-analysis, *closed diamond* indicates mean effect size and 95% CI when "missing studies" suggested by publication bias analysis are included

Publication Bias

Publication bias often is a problem when conducting research syntheses. A funnel plot of the standard error and effect size for each study was calculated using the trim and fill method of Duval and Tweedie (2000) and is shown in Fig. 3. It suggested the potential absence of two studies to the left of the mean effect size (i.e., 0.69). Thus, the possibility of publication bias exists. Further statistical analyses, conducted using a random effects model, suggested calculating the mean effect size including the two "missing studies" would result in the point estimate of the mean effect size being lowered from 0.69 to 0.53.

Test of Homogeneity

The homogeneity of the data was examined using the Q-statistic and I^2. The Q-statistic was statistically significant, $Q\ (11) = 22.6$, $p = 0.02$, indicating there was greater variability within the standardized mean effect sizes than expected from sampling error alone. Because the utility of the Q-statistic has been questioned (Huedo-Medina et al. 2006), I^2 was calculated. For the current meta-analyses, $I^2 = 51.2$, hence, 51.2% of the variance was between-study variance. Given the statistical significance of the Q-statistic and the large between study variance, as measured by I^2, cautious moderator analyses were warranted (Lipsey and Wilson 2001).

Moderator Analyses

The effects of moderator variables were examined under a method of moments random effects model using a

technique analogous to the analysis of variance for categorical variables and weighted multiple regression for continuous and dichotomous variables (Lipsey and Wilson 2001). The analyses were conducted using the weighted effect sizes for IQ as the dependent variable and variables hypothesized as possible moderators for the independent variables. Due to the small sample size for the dependent variable, moderator variables were analyzed separately.

The relation between two methodological characteristics (rigor and assignment to group) hypothesized as moderators of effect were assessed using a technique analogous to the one-way analysis of variance. The results are shown in Table 6. Neither methodological variable was found to have a statistically significant relation to changes in IQ. An analysis of the relation between experimental design and changes in IQ was desired, however, the small number of studies in each category precluded such an analysis. Although these analyses produced null results between methodological characteristics and outcome, relations have been found in other analyses of early intervention (Casto and Mastropiere 1986; Dunst et al. 1989). Thus, future syntheses of EIBI should continue to monitor the methodological quality (rigor) of studies in relation to outcomes.

Six possible continuous variables and one dichotomous variable hypothesized as moderators of effects were examined using a modified weighted regression procedure. The results of these analyses are shown in Table 7. The only variable with a statistically significant relation to change in IQ was supervisor training model ($B = 0.62$, $p = 0.01$). This suggests studies in which the supervisory personnel were trained according to the UCLA model outlined by Davis et al. (2002) and Lovaas (2003) were more likely to produce larger changes in IQ. This finding must be qualified by the following statements. The statistic was calculated on a small number of samples ($n = 12$). The analysis was conducted using the standardized mean change effect size, which is suspect. The size of the comparative groups, UCLA and other, were unequal and quite small, nine and three, respectively. With such small group sizes, each study contributes substantially to the results and replication may be unlikely. Thus, the finding may not stand when more studies have been completed. However, if this finding is replicated, then potential explanations are (a) an important element of EIBI has not been communicated adequately in the manuals, (b) some aspect of the supervisor training (e.g., identifying when program modifications should occur) may be more important than recognized, or (c) some combination of both of the above suggestions.

Although no other variables were statistically significant, the standardized regression coefficient for the variables of duration and total hours of therapy were large, $B = 0.48$ and 0.40, respectively. These data suggest the number of months of intervention and/or number of hours of therapy participants receive are related to a high probability of achieving a large change in IQ scores. Future research on the minimum number of hours and the minimum length of time needed for participants to achieve desirable outcomes is needed. From the existing data, a minimum number of months and minimum total hours could not be identified. Although evaluating the duration and total number of hours of therapy should continue to be studied, the therapist behaviors within those hours (i.e., fidelity) may be more important. Future analyses of EIBI should continue to quantify intervention characteristics (components) and analyze the differences related to outcome.

Table 6 Results of moderator analysis of categorical variables

Variable	*Q*	*df*	p-value
Rigor	3.39	2	0.184
Method of group assignment	4.61	2	0.100

Table 7 Results of moderator analyses using weighted multiple regression

Variable	*B*	*B*	p-value
Model of supervisor training	−0.624	−0.680	0.011*
Density	0.199	0.017	0.548
Duration	0.483	0.024	0.097
Total hours of treatment	0.402	<0.001	0.186
Pre-treatment chronological age	−0.045	−0.002	0.893
Pre-treatment IQ	−0.276	−0.010	0.376

* Statistically significant

Limitations of Syntheses Methods

The analyses of effect sizes suggest children with autism receiving EIBI made large gains on multiple domains of behavior, and made better progress than children with autism who receive less intense behavioral intervention or other treatments. These results should be interpreted with caution. Although the results appear to suggest most individuals made progress across all domains, this conclusion is not permitted by the data. Individual data typically were not presented, therefore, it is unclear if individuals making change in one domain (e.g., IQ) also made gains in another (e.g., adaptive behavior). Also, the outcomes were measured narrowly; thus, functioning across relevant domains was not measured in depth or breadth. Further, for the mean change effect sizes, no controls existed for maturation. Thus, while the effect sizes were often large, they cannot be attributed to EIBI exclusively.

Although conducting a multi-level synthesis should reduce the limitations of its findings, limitations existed. The inclusion criteria for the synthesis were narrow—more studies were excluded than included. Only research studies using group designs were used, and all were published in English, which increases the possibility of publication bias (Kromrey and Rendina-Gobioff 2006). A greater number and diversity of studies may be needed to draw definitive conclusions. Furthermore, the quality of literature reviews is constrained by the quality of the studies being reviewed. Although such studies are tremendously resource and effort intensive and the authors should be congratulated for attempting them, common threats to internal validity were present.

In the development and initial evaluation of large treatment packages and educational programs, such as EIBI, rigor of method sometimes is sacrificed. This was the case with the initial evaluation of EIBI (Lovaas 1987). As other authors (e.g., Foxx 1993; Gresham and MacMillan 1998; Mesibov 1993; Schopler et al. 1989) have indicated, participants were not selected randomly, they were not assigned randomly to groups, narrow and questionable measures were used, and treatment fidelity data were not reported. These realities make the findings suspect, because a number of threats to internal validity (e.g., maturation, lack of equivalent groups, concurrent history events, treatment drift or infidelity) are potential rival explanations for the obtained effects. Occasionally, a study report is disseminated because of its potential importance and the likelihood of such threats is ignored or minimize through argument (cf., the discussion section of Lovaas' 1987 paper). The assumption is the methodological deficiencies will likely be corrected in subsequent studies. Those studies should examine the package or program with rigorous research practices, including use of experimental (group or single subject) or causal-comparative designs rather than with quasi-experimental designs. With rigorous designs and research practices, threats to internal validity can be detected (controlled if present) and be thought of as absent if not detected. Unfortunately, some of the subsequent studies of EIBI have presented their own methodological deficiencies rather than correcting those of earlier studies; 2 of 13 used quasi-experimental designs (e.g., one-group, pre- and post-test designs; or non-equivalent group designs). This review included a collection of studies; many contained an array of methodological inadequacies (e.g., use of quasi-experimental designs, lack of equivalent groups, lack of adequate fidelity measures, unknown characteristics of comparison conditions). An alternative would have been to conduct a best evidence synthesis (Slavin 1986), which would have retained only highly rigorous studies. We concluded using the inclusion criteria specified earlier allowed a description of the state of the evidence as it currently exists. This tactic allows commentary on the weaknesses of the research methods used in evaluating EIBI with the hope that improvement will occur in subsequent studies.

Methodologically rigorous studies of educational programs, however, are complex and costly. To illustrate the complexity of such studies, a common recommendation is to use random assignment to groups in the context of a true experimental design (Campbell and Stanley 1963). Random assignment is thought to ensure equivalence of the compared groups—lack of which is a serious threat to internal validity. However, while random assignment is recommended, it is only successful when the sample size is large and the population is generally homogeneous. The sample sizes in the reviewed studies were small. The population of interest, young children with ASD, is by all accounts highly heterogeneous. Thus, random assignment, while a desirable research practice, is unlikely to guarantee equivalence of groups with the sample sizes reported, given the population's heterogeneity. The heterogeneity also makes it unclear what variables should be used to (a) document equivalence of the groups upon entry into the study, or (b) match participants to obtain group equivalence. Researchers evaluating EIBI must grapple with these issues, which often occur in the context of logistical realities and costs.

Although the use of meta-analytic techniques provides an unbiased evaluation of the data contained within the studies, this technique also has limitations. First, the sample size of this meta-analysis was small. Although conducting a meta-analysis using a small number of studies is possible (Lipsey and Wilson 2001), such an analysis can produce results which might not be representative of the population effect (i.e., be unlikely to be replicated). Second, interpretation of the magnitude of effect based solely on the mean effect size may be misleading and is not recommended. Furthermore, the standardized mean change effect size, which was used for the meta-analysis, is calculated without reference to a comparison or control group. Thus, the threats to internal validity of history, maturation, lack of procedural fidelity, and instrumentation threats cannot be eliminated. Although simulation studies have shown the standardized mean change effect size produces an overestimation of the true effect (Morris 2000), the relations should remain. Finally, the results of the moderator analyses should be interpreted cautiously due to the use of the standardized mean change effect size and the small number of studies used for the analyses.

Conclusions

This review is not the definitive statement about the effectiveness of EIBI; rather, it is a portrayal of selected

effects from the relatively small number of available studies of the intervention. Some of these studies are ongoing and other studies are likely to be initiated. Thus, the goals of this synthesis are to describe what appears to be known to date and to suggest some avenues for future research on EIBI. This research synthesis was conceptualized as a multi-level analysis of the research on EIBI incorporating both descriptive and statistical analyses. The descriptive analysis provided an overview of the studies' experimental methods, participant characteristics, and intervention characteristics, as well as outcomes for variables for which the calculation of an effect sizes was not appropriate. The effect size analyses for the samples and between groups in comparative studies provided estimates of the magnitude of participants' change after receiving EIBI. Finally, the meta-analysis provided an unbiased, quantitative synthesis of the studies examining the effects of EIBI on IQ scores, and an analysis of moderator variables across the studies; thus, providing one of the first attempts to quantify relations within the EIBI research. When conducting research syntheses on debated interventions, quantitative and unbiased methods for summarizing findings are desired.

In sum, the findings of the current synthesis were mixed. Although the data and findings of this synthesis can be used to make claims about the effectiveness of EIBI (particularly in relation to IQ scores), the synthesis also exposed many knowledge gaps. Care must be taken to keep the findings of this synthesis in context and not make findings appear larger than they are. Although some of the findings of this synthesis were robust, all of the findings had limitations. These limitations provide areas for future research to explore, and ensure the debate about the effectiveness of EIBI will continue. The large effects demonstrated in the studies comparing EIBI with other treatments (e.g., eclectic treatment or treatment as usual) might be an artifact of comparison groups with poorly defined (i.e., organized) treatments that have yet to be empirically validated. No comparisons between EIBI and other widely recognized treatment programs have been published. Without comparisons between EIBI and empirically validated treatment programs, it is not possible to determine if EIBI is more or less effective than other treatment options.

The findings of the moderator analysis suggest the greatest results on IQ change might be seen when supervisory staff were trained using the UCLA model, duration of intervention was long, and the total hours of therapy were high. However, the small number of studies, the use of the standardized mean change effect size rather than the standardized difference effect sizes and the measured dependent variable (IQ) are qualifying factors.

Finally, the data analyzed for this synthesis demonstrated EIBI has been shown as an intervention capable of producing strong effects, suggesting EIBI can be an effective intervention for some children with autism. However, the intervention has not worked for all children. A review of the data from studies reporting data individually for each participant revealed at least one participant receiving EIBI in each study did not improve or regressed on at least one outcome variable. While this may be an artifact of measures that are insensitive to changes within participants, it also suggests EIBI is not an intervention that fits the needs of all children with autism. Because receiving early intervention is thought to be critical in the determination of future outcomes of children with autism (Lord et al. 2001), it is imperative children not responding to intervention are identified early so additional and/or different treatments can begin. Therefore, practitioners must continuously monitor the progress being made by all individuals receiving EIBI. Further analyses of measures of progress, such as the Early Learning Measure (Smith et al. 2000), are needed and is of high priority.

Acknowledgments The authors would like to thank Pat Snyder for her statistical consultation, Ann Kaiser, Craig Kennedy, and Mark Lipsey for their review of earlier drafts of this manuscript, and Erin Barton and Matthew Busick for their assistance with interobserver agreement. This project was supported by the US Department of Education Office of Special Education through an ESCE Doctoral Leadership Training Grant (H325D030012).

References

References marked with an asterisk indicate studies included in the review.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed. ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd Rev. ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

*Anderson, S. R., Avery, D. L., DiPietro, E. K., Edwards, G. L., & Christian, W. P. (1987). Intensive home-based early intervention with autistic children. *Education & Treatment of Children, 10*, 353–366.

Bellg, A. J., Resnick, B., Minicucci, D. S., Ogedegbe, G., Ernst, D., Borrelli, B., et al. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*, 443–451.

*Bibby, P., Eikeseth, S., Martin, N. T., Mudford, O. C., & Reeves, D. (2001). Progress and outcomes for children with autism receiving parent-managed intensive interventions. *Research in Developmental Disabilities, 22*, 425–447.

Billingsley, F. F., White, O. R., & Munson, R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment, 2*, 229–241.

*Birnbrauer, J. S., & Leach, D. J. (1993). The Murdoch early intervention program after 2 years. *Behaviour Change, 10*(2), 63–74.

*Boyd, R. D., & Corley, M. J. (2001). Outcome survey of early intensive behavioral intervention for young children with autism in a community setting. *Autism, 5*, 430–441.

Burack, J. A., Iarocci, G., Flanagan, R. D., & Bowler, D. M. (2004). On mosaics and melting pots: Conceptual considerations of comparison and matching strategies. *Journal of Autism and Developmental Disorders, 34*, 65–73.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Casto, G., & Mastropiere, M. A. (1986). The efficacy of early intervention programs: A meta-analysis. *Exceptional Children, 52*, 417–424.

Charman, T., & Howlin, P. (2003). Research into early intervention for children with autism and related disorders: Methodological and design issues. *Autism, 7*, 217–225.

*Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Journal of Developmental and Behavioral Pediatrics, 27*(2), S145–S155.

Davis, B. J., Smith, T., & Donahoe, P. (2002). Evaluating supervisors in the UCLA treatment model for children with autism: Validation of an assessment procedure. *Behavior Therapy, 33*, 601–614.

Dawson, G., & Osterling, J. (1997). Early intervention in autism. In M. J. Guarlnick (Ed.), *The effectiveness of early intervention* (pp. 307–326). Baltimore: Brookes.

Dunst, C. J., Synder, S. W., & Mankinen, M. (1989). Efficacy of early intervention. In H. J. Walberg, M. C. Wang, & M. C. Reynolds (Eds.), *Handbook of special education: Research and practice* (Vol. 3, low incidence conditions; pp. 259–294). Elmsford, NY: Pergamon.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 127–144). Hoboken, NJ: John Wiley & Sons.

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

*Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2007). Outcome for children with autism who began intensive behavioral treatment between ages 4 and 7: A comparison study. *Behavior Modification, 31*, 264–278.

*Eldevik, S., Eikeseth, S., Jahr, E., & Smith, T. (2006). Effects of low-intensity behavioral treatment for children with autism and mental retardation. *Journal of Autism and Developmental Disorders, 36*, 211–224.

Foxx, R. M. (1993). Sapid effects awaiting independent replication. *American Journal on Mental Retardation, 97*, 375–376.

Green, V. A., Pituch, K. A., Itchon, J., Choi, A., O'Reilly, M., & Sigafoos, J. (2006). Internet survey of treatments used by parents of children with autism. *Research in Developmental Disabilities, 27*, 70–84.

Gresham, F. M. (2005). Treatment integrity and therapeutic change: Commentary on Perepletchikova and Kazdin. *Clinical Psychology: Science and Practice, 12*, 391–394.

Gresham, F. M., & MacMillan, D. L. (1998). Early intervention project: Can its claims be substantiated and its effects replicated? *Journal of Autism and Developmental Disorders, 28*, 5–13.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hedges, L., & Olkin, I. (1985). *Statistical models for meta-analysis*. New York: Academic Press.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Marínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*, 193–206.

Iovannone, R., Dunlap, G., Huber, H., & Kincaid, D. (2003). Effective educational practices for students with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 18*, 150–165.

Johnson, E., & Hastings, R. P. (2002). Facilitating factors and barriers to the implementation of intensive home-based behavioural intervention for young children with autism. *Child: Care, Health and Development, 28*, 123–129.

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement, 66*, 357–373.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lord, C., Bristol-Power, M., Cafiero, J. M., Filipek, P. A., Gallagher, J. J., Harris, S. L., et al. (2001). *Educating children with autism*. Washington, DC: National Academy Press.

Lord, C., Rutter, M., DiLavore, P. C., & Lisi, S. (1999). *Autism diagnostic observation schedule*. Los Angeles: Western Psychological Services.

Lord, C., Wagner, A., Rogers, S., Szatmari, P., Aman, M., Charman, T., et al. (2005). Challenges in evaluating psychosocial interventions for autistic spectrum disorders. *Journal of Autism and Developmental Disorders, 35*, 695–708.

Lovaas, O. I. (1981). *Teaching developmentally disabled children: The me book*. Baltimore: University Park.

*Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55*, 3–9.

Lovaas, O. I. (2003). *Teaching individuals with developmental delays: Basic intervention techniques*. Austin, TX: Pro-Ed.

*Magiati, I., Charman, T., & Howlin, P. (2007). A two-year prospective follow-up study of community-based early intensive behavioural intervention and specialist nursery provision for children with autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 48*, 803–812.

McEachin, J. J., Smith, T., & Lovaas, O. I. (1993). Long-term outcome for children with autism who received early intensive behavioral treatment. *American Journal on Mental Retardation, 97*, 359–372.

Mesibov, G. B. (1993). Sapid effects awaiting independent replication. *American Journal on Mental Retardation, 97*, 379–380.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *The British Journal of Mathematical and Statistical Psychology, 53*, 17–29.

Mundy, P. (1993). Normal versus high functioning status in children with autism. *American Journal on Mental Retardation, 97*, 381–384.

Perepletchikova, R., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12*, 365–383.

Reichow, B., Barton, E. E., Volkmar, F. R., & Cicchetti, D. V. (2007, May). *The status of research on interventions for young children with autism spectrum disorders*. Poster presented at the International Meeting for Autism Research, Seattle, WA.

Reichow, B., Volkmar, F. R., & Cicchetti, D. V. (in press). Development of an evaluative method for determining the strength of research evidence in autism. *Journal of Autism and Developmental Disorders*. doi: 10.1007/s10803-007-0517-7.

Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised*. Los Angeles: Western Psychological Services.

*Sallows, G. O., & Graupner, T. D. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. *American Journal on Mental Retardation, 110*, 417–438.

Sandall, S. R., Hemmeter, M. L., Smith, B. J., & McLean, M. E. (2005). *DEC recommended practices: A comprehensive guide for practical application in early intervention/early childhood special education*. Longmont, CO: Sopris West.

Schopler, E., Short, A., & Mesibov, G. (1989). Relation of behavioral treatment to "normal functioning": Comment on Lovaas. *Journal of Consulting and Clinical Psychology, 57*, 162–164.

Schreibman, L. (2000). Intensive behavioral/psychoeducational treatments for autism: Research needs and future directions. *Journal of Autism and Developmental Disorders, 30*, 373–378.

*Sheinkopf, S. J., & Siegel, B. (1998). Home-based behavioral treatment of young children with autism. *Journal of Autism and Developmental Disorders, 28*, 15–23.

Slavin, R. E. (1986). Best evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*(9), 5–11.

Smith, T., Buch, G. A., & Gamby, T. E. (2000). Parent-directed intensive early intervention for children with pervasive developmental disorder. *Research in Developmental Disabilities, 21*, 297–309.

*Smith, T., Eikeseth, S., Klevstrand, M., & Lovaas, O. I. (1997). Intensive behavioral treatment for preschoolers with severed mental retardation and pervasive developmental disorder. *American Journal on Mental Retardation, 102*, 238–249.

*Smith, T., Groen, A. D., & Wynn, J. W. (2000). Randomized trial of intensive early intervention for children with pervasive developmental disorder. *American Journal on Mental Retardation, 105*, 269–285.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales*. Circle Pines, MN: American Guidance Service.

Stahmer, A. C., Collings, N. M., & Palinkas, L. A. (2005). Early intervention practices for children with autism: Descriptions from community providers. *Focus on Autism and Other Developmental Disabilities, 20*, 66–79.

Symes, M. D., Remington, B., Brown, T., & Hastings, R. P. (2006). Early intensive behavioral intervention for children with autism: Therapists' perspectives on achieving procedural fidelity. *Research in Developmental Disabilities, 27*, 30–42.

Volkmar, F. R. (1998). Categorical approaches to the diagnosis of autism. *Autism, 2*, 45–59.

Volkmar, F. R., Cook, E. H., Jr., Pomeroy, J., Realmuto, G., & Tanguay, P. (1999). Practice parameters for the assessment and treatment of children, adolescents, and adults with autism and other pervasive developmental disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 38*(Suppl. 12), 32S–54S.

Wolery, M., & Garfinkle, A. N. (2002). Measures in intervention research with young children who have autism. *Journal of Autism and Developmental Disorders, 32*, 463–478.