Using Participant Data to Extend the Evidence Base for Intensive Behavioral Intervention for Children With Autism

Sigmund Eldevik
Akershus University College (Lillestrom, Norway)

Richard P. Hastings and J. Carl Hughes
Bangor University (Bangor, Wales)

Erik Jahr
Akershus University Hospital (Lorenskog, Norway)

Svein Eikeseth
Akershus University College

Scott Cross
Lovaas Institute for Early Intervention (Culver City, CA)

Abstract

We gathered individual participant data from 16 group design studies on behavioral intervention for children with autism. In these studies, 309 children received behavioral intervention, 39 received comparison interventions, and 105 were in a control group. More children who underwent behavioral intervention achieved reliable change in IQ (29.8%) compared with 2.6% and 8.7% for comparison and control groups, respectively, and reliable change in adaptive behavior was achieved for 20.6% versus 5.7% and 5.1%, respectively. These results equated to a number needed to treat of 5 for IQ and 7 for adaptive behavior and absolute risk reduction of 23% and 16%, respectively. Within the behavioral intervention sample, IQ and adaptive behavior at intake predicted gains in adaptive behavior. Intensity of intervention predicted gains in both IQ and adaptive behavior.

There is a growing body of evidence that intensive behavioral intervention can result in significant improvement in the intellectual, social, adaptive, and language functioning of young children with autism spectrum disorders (Cohen, Amerine-Dickens, & Smith, 2006; Eikeseth, Smith, Jahr, & Eldevik, 2007; Hayward, Eikeseth, Gale, & Morgan, 2009; Howard, Sparkman, Cohen, Green, & Stanislaw, 2005; Lovaas, 1987; Remington et al., 2007; Sallows & Graupner, 2005; Smith, Groen, & Wynn, 2000). A sizeable minority of children might even reach the average to superior range within one or more of these areas of functioning following intervention (Cohen et al., 2006; Eikeseth et al., 2007; Hayward et al., 2009; Howard et al., 2005; Lovaas, 1987; Remington et al., 2007; Sallows & Graupner, 2005; Smith et al., 2000). There is also promising, although limited, evidence that these outcomes may maintain over the long term into adolescence

following the cessation of intervention (McEachin, Smith, & Lovaas, 1993). Such data have led to positive conclusions about the evidence base for intensive behavioral intervention in two recent narrative reviews (Eikeseth, 2009; Rogers & Vismara, 2008). According to Rogers and Vismara, clinic-based intensive behavioral intervention (or what they call the Lovaas treatment approach) can be considered well-established based on formal criteria (Chambless et al., 1996, 1998; Chambless & Hollon, 1998).

Although there are statistically significant group differences in controlled studies, a more thorough analysis of what the results mean in clinical terms is also required. Such an analysis can be done in several ways. One approach is to examine outcome using meta-analysis of aggregated data that are typically reported in published studies, such as the mean pre- and posttest scores in the experimental and control groups (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, Hedges, & Valentine, 2009). Reichow and Wolery (2009) recently conducted a synthesis of the research, including an aggregated data meta-analysis on the effects of intensive behavioral intervention for children with autism. Based on 12 studies, they found a weighted mean change (i.e., pre–post change in intervention groups only) effect size for IQ of .69 following intensive behavioral intervention. Such an effect size would normally be considered clinically meaningful. In a second aggregated data meta-analysis of 9 controlled studies of intensive behavioral intervention, using a weighted mean difference effect size, Eldevik et al. (2009) found a large effect for IQ change in favor of intensive behavioral intervention, Hedges' $g = 1.10$, 95% CI = .87, 1.34, and a smaller, although still statistically significant, effect for change in adaptive behavior composite (ABC) scores, Hedges' $g = .66$, 95% CI = .41, .90.

An especially significant feature of the Eldevik et al., (2009) analysis is that individual participant data were obtained from the authors of studies selected for the review. Thus, the aggregated data meta-analysis was based on individual study effect sizes calculated using the same method, for similar evaluation periods, and following the removal of children whose data appeared in more than one report. An aggregated data meta-analysis of individual study effect sizes derived from individual participant data is a recommended first step in any analysis of

evidence for an intervention using individual level data (Cooper & Patall, 2009). A second step is to conduct an individual participant data meta-analysis proper. Such an analysis is likely to have important benefits over aggregated data meta-analysis, including the possibility of dividing the individual participants into new subgroups and applying different statistical methods (Cooper & Patall, 2009). This form of meta-analysis (sometimes also called *mega-analysis*) involves the combination of data across studies into a single intervention and comparison/control group(s).

Given that the outcome for individual children in intensive behavioral intervention studies varies considerably (Howlin, Magiati, & Charman, 2009), an important step when examining the evidence base for this intervention is to evaluate meaningful changes at the level of individual children. To date, the method for assessing which children achieve meaningful change (best outcome) has not been consistent in existing research. Lovaas (1987) defined *best outcome* as intellectual functioning (IQ) scores within the normal range and successful first grade performance in public schools. Sallows and Graupner (2005) used the terms *rapid learners* and *moderate learners* to define similar outcomes. A more objective method for establishing meaningful change at the level of the individual child is needed.

Remington et al. (2007) used the Reliable Change Index (N. Jacobson & Truax, 1991), a construct borrowed from psychotherapy outcome research, to examine meaningful change in their intensive behavioral intervention controlled study. *Reliable change* is the amount by which an outcome measure needs to change before one can be 95% certain that the change cannot be accounted for by the variability of scores in the sample and/or measurement error. The reliable change index is computed by subtracting the pretest scores from the posttest scores and then dividing by the standard error of difference. The standard error of difference is, in turn, computed directly from the standard error of measurement and describes the distribution of change scores that would be expected if no change occurred (N. Jacobson & Truax, 1991, p. 14).

Using N. Jacobson and Truax's formula, Remington et al. (2007) found that 6 out of 23 children (26%) in their intensive behavioral intervention group achieved positive reliable change in IQ after 2 years, whereas 3 out of 21

(14%) in the treatment as usual group achieved this level of change, and the IQs of 3 children in this group also decreased to a reliable extent. To date, no other published intensive behavioral intervention study has used this objective criterion to identify best outcome children, and Remington et al. only reported this analysis for IQ and not other domains of outcome.

One advantage of establishing a dichotomous outcome variable for change in intensive behavioral intervention at the level of individual participants (i.e., achieved reliable change or not) is that effect size statistics commonly used to evaluate the potency of health interventions can be generated. Such statistics include the number needed to treat and absolute risk reduction (Straus & Sackett, 2005). These statistics are particularly helpful as simple ways to communicate information about interventions to policymakers. The number needed to treat represents the number of children who would need to be treated with a specified intervention to obtain one additional success over the success rate in a comparison intervention. For example, number needed to treat = 4 means that for every four children who are treated with intervention X, one additional child will respond to this intervention who would not have responded to a comparison intervention. A result of number needed to treat = 1 means that all children receiving an intervention succeed when they would not have done so following a comparison intervention. In other words, the larger the number needed to treat, the less effective the treatment relative to the comparison (Kraemer et al., 2003).

Absolute risk reduction is computed in a similar way as number needed to treat but expressed as a measure of the difference in percentage response between two interventions (Pinson & Gray, 2003). When the absolute risk reduction is used as a measure of intervention effectiveness, the results are usually given in negative outcome. This means that an effective intervention will reduce negative outcome or, put another way, reduce the risk of having bad outcome. For example, if in intervention A, 50% of patients do not respond to intervention and in intervention B, 90% do not respond to intervention, the absolute risk reduction (also called risk difference) is 40% in favor of intervention A.

A further advantage of establishing an objective criterion for meaningful outcome for individual children with autism receiving intensive behavioral intervention is that the search for correlates or predictors of intensive behavioral intervention outcome can become more consistent. For example, the 6 children who achieved reliable change following intensive behavioral intervention in the Remington et al. (2007) study were compared with the 3 children in the intensive behavioral intervention group whose IQs decreased. The children who met reliable change criteria had higher IQ, mental age (MA), Vineland Adaptive Behavior Scales–VABS (Sparrow, Balla, & Cicchetti, 1984) Composite scores, along with higher VABS Communication and Socialization scores at intake. In addition, these best outcome children at intake had lower VABS Motor scores, more behavior problems on the Developmental Behavior Checklist (Einfeld & Tonge, 1995), and more autistic symptoms on the Developmental Behavior Checklist autism algorithm (Einfeld & Tonge, 2002), but also had fewer treatment hours in their second year of intensive behavioral intervention.

Apart from the Lovaas (1987) intensity comparison (40 vs. ≤10 hr), intensive behavioral intervention studies have not been explicitly designed to explore moderators of outcome. Rather, as in the Remington et al. (2007) study, various methods to examine correlates of outcome have been adopted. Correlates of outcome explored in existing research include rates of learning early in intervention or initial skill acquisition (Sallows & Graupner, 2005; Weiss, 1999), age at intake (Harris & Handleman, 2000), IQ at intake (Ben-Itzchak & Zachor, 2007; Harris & Handleman, 2000), initial social skills (Ben-Itzchak & Zachor, 2007; Eikeseth, Smith, Jahr, & Eldevik, 2007), toy play and socially avoidant behavior at intake (Sherer & Schreibman, 2005), and autism subtype (Beglinger & Smith, 2005). Notably, despite its potential significance to the intensive behavioral intervention debate, the intensity of intervention has been shown to relate to outcomes only in Lovaas' (1987) original experimental comparison. However, most salient in the current context is that given there is no consistency in the definition of *meaningful outcome* in intensive behavioral intervention, there is currently no evidence base that can be used to identify children at intake who are likely to achieve best outcome, let alone to prescribe a certain intensity (or duration) of intervention.

In the present study we collected individual participant data by contacting authors and from

published intensive behavioral intervention outcome studies identified via a systematic review. We then used all of these data to establish whether each child met reliable change criteria for changes in IQ or adaptive behavior after approximately 2 years of intervention. These data were then used to address two aims. First, we conducted an individual participant data meta-analysis of intensive behavioral intervention outcomes against those of control/comparison interventions. This extended the work of Eldevik et al. (2009) and Reichow and Wolery (2009) because both controlled and uncontrolled studies could be included in the analysis, the data were at a different level of analysis than these authors' aggregated data meta-analyses, and effect size statistics based on dichotomous outcomes were adopted. Our second aim was to explore predictors of outcome in children who had received intensive behavioral intervention. Using this analysis we were able to extend beyond the small n analyses from individual published studies and to facilitate a more sophisticated analysis of outcome prediction in one important respect. We were able to explore both main effects as well as interactions between key variables (e.g., age at intake combined with IQ at intake) as potential predictors. Such analyses were not possible in previous research because participant numbers were too small.

Method

Searching Strategy and Data Collection

We conducted a comprehensive literature search using PsycINFO, Pubmed, and ERIC databases (up to March 2008) using a combination of the following terms: *behavior analytic*, *behavioral*, *early*, *intervention*, and *autism* and/or *pervasive developmental disorder -not otherwise specified* (PDD-NOS). The first author read the titles and abstracts of all papers collected from this initial search; studies that contained standardized outcome data on the effects of behavioral intervention for young children with autism were obtained for more detailed coding. The first author manually browsed the reference section of each study in an attempt to locate other studies that might have been missed during the electronic search.

Following this selection process, we developed a coding scheme (available from the first author) and coded the selected studies in two main ways. First, we coded whether the children

had received behavioral intervention that adhered to the common elements described by Green, Brennan, and Fein (2002, p. 70); that is, (a) intervention was individualized and comprehensive, addressing all skill domains; (b) many behavior analytic procedures were used to build new repertoires and reduce interfering behavior (e.g., differential reinforcement, prompting, discrete-trial instruction, incidental teaching, activity-embedded trials, task analysis, and others); (c) one or more individuals with advanced training in applied behavior analysis and experience with young children who had autism directed the intervention; (d) typical developmental sequences guided selection of intervention goals and short-term objectives; (e) parents served as active cotherapists for their children; (f) intervention was delivered in one-to-one fashion initially, with gradual transitions to small- and large-group formats when warranted; (g) intervention typically began in the home and was carried over into other environments (e.g., community settings), with gradual, systematic transitions to preschool, kindergarten, and elementary school classrooms when children developed the skills required to learn in those settings; (h) programming was intensive, including 20 to 30 hr of structured sessions per week plus informal instruction and practice throughout most of the children's other waking hours, year round; (i) in most cases, the duration of intervention was 2 or more years; and (j) most children started intervention in the preschool years, when they were 3 to 4 years of age.

The second way we coded the selected studies was by applying a series of true/false scores using the following criteria: (a) the participants were, on average, between 2 and 7 years old when intervention started; (b) the children were independently diagnosed with autism or PDD-NOS; (c) a full-scale measure of intelligence and/or a standardized measure of adaptive behavior, such as the VABS, was conducted at intake and after intervention–we excluded studies in which the researchers had primarily administered a nonverbal intelligence measure, such as the Leiter International Performance Scale–Revised (Roid & Miller, 1997) or the Merrill-Palmer Scale of Mental Tests (Stutsman, 1948) because the results of such assessments may differ substantially from those of full scale intelligence tests (Scheuffgen, Happe, Anderson, & Frith, 2000); (d) the duration of intervention was between 12 and 36 months; (e) the study was not a case study (or series of case

studies); and (f) the results had been published in a peer-reviewed journal. In addition, if data on control or comparison groups were reported, these were included and grouped according to the criteria given below. If all the above criteria were met, the authors of the study were approached and asked to provide data on individual children, if this was not already available in the published paper.

Data on other groups included in intensive behavioral intervention evaluation studies were coded as either comparison group data, which meant that another form or forms of intervention of similar intensity (in terms of 1:1 hours) was specified, or control group data, which meant that no or a considerably less intensive alternative intervention was specified, often merely described as "treatment as usual." Although it would probably be impossible to determine whether the children in the comparison groups had a specific common provision (even within a single study), classifying the studies in this way could yield useful information. For example, it is important to establish whether intensive behavioral intervention might be efficacious when compared to other similarly intensive interventions or only when compared against an ill-defined treatment as usual.

The initial electronic and manual searches resulted in 2,150 potential hits in total across the databases. Through the screening process, we selected 33 papers for closer examination and detailed coding. We also chose one of the database searches that had resulted in 607 potential hits for a reliability check. The screening results from the first author were compared to those of a second coder (another author) using the same decision criteria. Agreement was high overall in terms of whether to select a paper for further coding, Cohen's Kappa = .85. Notably, disagreements only occurred because the second screener included fewer studies than did the first author. Thus, there were no instances of the second screener including a study for further coding that was not already included by the first author.

The remaining 33 studies were then coded by the first author and two independent scorers (master's level students in behavior analysis) using the true/false criteria described above. Agreement was calculated between the first author and each of the independent scorers separately by dividing the total number of agreements by the total number of agreements plus disagreements and multiplying by 100. Initial agreement was high in both cases (91% and 94%, respectively), and the few disagreements that occurred were resolved after brief discussions. We excluded 18 out of the 33 studies for one or more of the following reasons: (a) 7 had inadequate intake and/or outcome data, most often reporting primarily Performance IQ instead of Full Scale IQ (Bibby, Eikeseth, Martin, Mudford, & Reeves, 2002; Drew et al., 2002; Fenske, Zalenski, Krantz, & McClannahan, 1985; Luiselli, Cannon, Ellis, & Sisson, 2000; Magiati, Charman, & Howlin, 2007; Sheinkopf & Siegel, 1998; Solomon, Necheles, Ferch, & Bruckman, 2007); (b) in 5 of the studies, the duration of intervention was too short to meet inclusion criteria (Harris, Handleman, Gordon, Kristoff, & Fuentes, 1991; Ingersoll, Schreibman, & Stahmer, 2001; Reed, Osborne, & Corness, 2007a, 2007b; Stahmer & Ingersoll, 2004); (c) in 2 papers the researchers reported data from case studies only (Butter, Mulick, & Metz, 2006; Green et al., 2002); (d) in 3 of the studies, investigators reported data that were already included in other studies (Beglinger & Smith, 2005; Eikeseth et al., 2007; McEachin et al., 1993); and (e) upon closer scrutiny one of the studies provided intervention that did not meet the definition of behavioral intervention (Gabriels, Hill, Pierce, Rogers, & Wehner, 2001).

In only 4 of the 15 remaining studies did researchers report individual outcome data in the original published paper. The authors of the 11 remaining studies were contacted and asked to provide data on individual children; all of them agreed. However, individual data from Control Group 2 (n = 21) in the Lovaas (1987) study were not available. Furthermore, data from 4 children in the comparison group of one study (Eldevik, Eikeseth, Jahr, & Smith, 2006) were extracted because they were also in the comparison group of another study included in the analysis (Eikeseth, Smith, Jahr, & Eldevik, 2002). One of the authors whom we contacted also volunteered an additional study (Hayward et al., 2009); because this study had been subject to peer review and met all other criteria, it was also included in the present analysis. Figure 1 presents a flowchart of the search and selection procedure.

Table 1 summarizes the main characteristics of the studies included in this analysis, the mean age of participants at intake, and their mean IQ and adaptive behavior scores at intake and postintervention. Furthermore, the mean intensi-
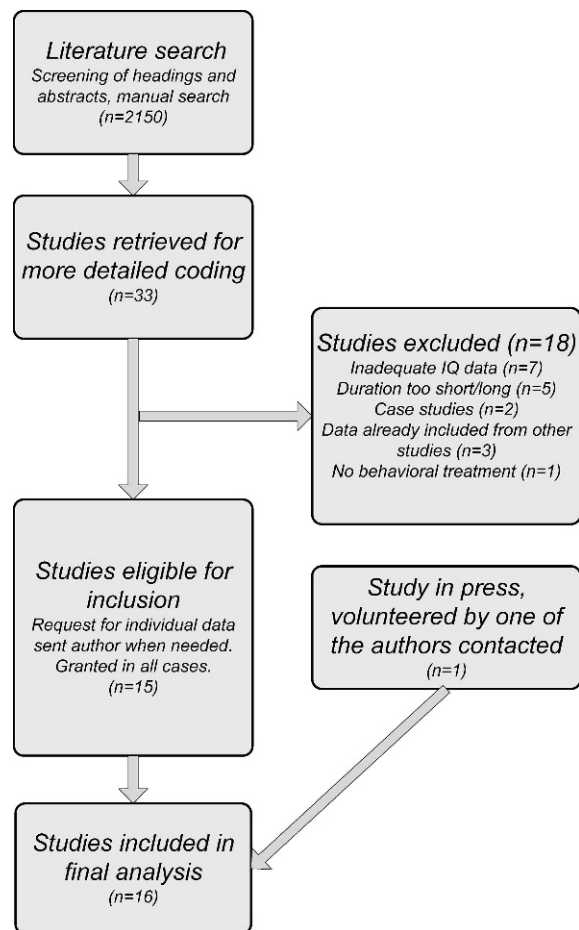
Figure 1. Flowchart on the procedure for selecting studies.

ty in terms of weekly hours and duration is provided, and the research design and assignment procedures employed are briefly described along with any inclusion criteria employed in the original paper. If the researchers reported outcome data at more than one point in time, we chose the point that was closest to a 2-year duration of intervention.

Participants

Individual data were available for 453 participants, including 309 who had received intensive behavioral intervention, 105 in control groups, and 39 in comparison groups. Due mainly to different assessment protocols (i.e., because the measures were not included in the research), some IQ data (1 study) and adaptive behavior data (2 studies) are missing (see below). A one-way ANOVA showed that the three groups were similar on intelligence measures at intake. Chil-

dren in the comparison intervention group were older than those in the other two groups at intake, and children in the control group had higher VABS Composite scores (see Table 2). However, the total sample, as well as the separate subgroups with the sample, are generally representative of the autism population (Volkmar & Klin, 2005). Because not all authors were able to provide data on the gender of each child (38.5% missing), these data were not included in the present analysis. Within the intensive behavioral intervention group, the number of weekly intervention hours for each child was only available for 75 out of 309 children (24.3%). To include intensity as a variable, we decided to create a median split of the overall data using a hierarchy of evidence. First, we used the data provided by the author on intervention intensity for each child if these data were available. Second, we used the mean weekly hours of intervention for the intensive behavioral intervention group that the child was in. Data on the group means for the intensive behavioral intervention studies were typically based on reports that all children had been exposed to at least the relevant number of weekly hours specified in the intervention. In total, 152 children (49.5%) received 36 or more hours of intervention on a weekly basis, and 155 children (50.5%) received fewer than 36 hr of weekly intervention.

Child Measures

Intelligence. The Bayley Scales of Infant Development (BSID), either the first or second edition (Bayley, 1969, 1993), were most often used for the youngest children or the children who scored below the basal on other intelligence tests. The BSID is a measure of mental developmental level for children up to 42 months. It yields a Mental Developmental Index, which is considered broadly equivalent to an IQ. For the older and higher functioning children, the most frequently used measures of intelligence were the Stanford-Binet Intelligence Scale: Fourth Edition (Thorndike, Hagen, & Sattler, 1986), the Wechsler Preschool and Primary Scale Intelligence-Revised (Wechsler, 1989), the Wechsler Intelligence Scale for Children-Revised (Wechsler, 1974), or the Wechsler Intelligence Scales for Children-Third Edition (Wechsler, 1993). All of these tests have been validated and used extensively for children with developmental delays and autism (Newsom

& Hovanitz, 1997). If the child scored below the norms on a test, researchers generally computed a ratio IQ by dividing the obtained MA with chronological age and multiplying by 100. Unfortunately, we did not have data regarding which tests were used for each child at what point nor information on whom a ratio IQ was used. IQ outcome data were obtained from a total of 422 children (31 missing). These were divided as follows: 279 children in the intensive behavioral intervention groups (30 missing), 104 children in the control groups (1 missing), and 39 children in the comparison groups (0 missing).

Adaptive behavior. The VABS, which was the measure for adaptive skills in all studies included in this research, provides standard scores for communication, daily living skills, and socialization; and for children under 6 years old, motor skills. It also yields a total ABC. In the present study we only used the ABC scores because we did not have access to the various domain scores. The VABS is widely regarded as the best interview for assessing adaptive levels for children with autism (Klin, Saulnier, Tsatsanis, & Volkmar, 2005). Data on adaptive behaviors were obtained from a total of 357 children (96 missing): 248 children in the intensive behavioral intervention groups (61 missing), 70 children in the control groups (35 missing), and 39 children in the comparison groups (0 missing).

Data Analysis Procedure

To evaluate effectiveness of behavioral intervention at the level of individual children, we applied the statistical approach outlined by N. Jacobson and Truax (1991). The formula for computing reliable change requires that one is able to determine the stability and distribution of the test scores (in this case IQ and ABC scores). Because neither of these are well-established for young children with autism, we decided to use our relatively large sample to generate suitable information (following Remington et al., 2007). We estimated the stability of test scores over 2 years by finding the correlation between pre- and postscores in the control group, where no identified intervention had been applied and, thus, where stability might be better estimated than from groups receiving active interventions. We used intake data to calculate the *SD* for test scores from the whole sample of 453 children. Using the formula reported in N. Jacobson and

Truax (1991, p. 14), we established the absolute change in scores required to achieve a reliable change index score of 1.96 (95% certainty).

In some intensive behavioral intervention studies, investigators excluded children with intake IQs at or below 35 (Cohen et al., 2006; Sallows & Graupner, 2005; Smith et al., 2000). Given this practice, we conducted analyses on the whole sample and also repeated them for the sample (*n* = 387) whose intake IQs were 35 or above. Thus, we calculated change scores above which reliable change was indicated for the whole sample and for the 35+ IQ sample. To be considered reliable, the change in IQ had to be at least 27.4 points, rounded to 27 for the purposes of this analysis (26.6 for the subset of children with IQ > 35 at intake); for the ABC the change had to be at least 21.0 points (21.3 for the subset of children with IQ > 35 at intake). The more lenient criterion on the VABS mainly reflected a smaller *SD* in the test scores at intake. None of the analyses reported here revealed a different pattern of results when the children with intake IQs below 35 were excluded; thus, no further results excluding those children are reported.

After classifying each child in terms of whether his or her intellectual functioning and adaptive levels changed to a reliable extent, we computed number needed to treat and absolute risk reduction (Laupacis, Sackett, & Roberts, 1988). This was done for the total sample (i.e., an individual participant data meta-analysis) and, when possible, for the individual studies (i.e., studies that had a control or comparison group). The latter were included to illustrate the degree of variability across studies. To conduct the number needed to treat and absolute risk reduction calculations, we used readily available free access online calculators (Straus, Newton, & Tomlinson, 2004).

To explore predictors of intensive behavioral intervention outcomes, we conducted a multiple regression analyses for the behavioral intervention group (*n* = 309). The dependent variables were absolute change scores for IQ and ABC. We used absolute change scores rather than a dichotomous outcome variable for ease of analysis and to ensure the maximum possible variability in the dependent variable given the difficulties inherent in searching for moderated effects in multiple regression analysis (McClelland & Judd, 1993). The variables we investigated as possible predictors were age at intake, IQ at intake, ABC at

Table 1. Characteristics of the Studies Included in the Present Analyses

| | Age | | Pretest | | | | Posttest | | | |
| | | | IQ | | ABC | | IQ | | ABC | |
Country/Study/Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
United States										
Lovaas (1987)										
IBA[c] (clinic)	34.6	8.9	62.9	13.7	–		83.3	28.6	–	
Control	40.9	10.3	57.1	14.5	–		50.1	22.4	–	
Anderson et al. (1987)										
IBI (clinic)	42.8	11.8	55.0	18.9	47.7	8.2	60.6	25.1	57.7	15.3
Smith et al. (1997)										
IBI (clinic)	36.0	6.9	27.8	4.9	50.3	9.1	35.8	14.3	51.7	17.9
Control	38.0	5.4	27.3	5.4			24.0	8.2	–	
Weiss (1999)										
IBI (parent)	42.0		–		49.9	7.8	–		83.6	28.3
Harris & Handleman (2000)										
IBI (clinic)	49.0	8.8	59.3	24.2	–		77.6	28.6	–	

(Table 1 continued)

Table 1. Extended

Intensity			Gender		Design/Assignment/	Comments[b]
Hr	Mon.	*n*	M	F	Inclusion[a]	
40	24–36	19	16	3	QCT/staff availability and archives	Five subjects deemed untestable at intake, 3 in experimental group and 2 in Control Group 1
<10	24	19	11	8	Included if CA < 40 months if mute or CA < 46 months if echolaic and prorated MA of > 11 months at CA 30 months	Intelligence scores based on MA score from Vineland Social Maturity Scale (Doll, 1953) were used in these cases INDA[c]
20	12–24	14	11	3	UCT/parent willingness and geographical Included if CA < 72 months	Ratio IQ and VABS computed on basis of tables in original paper Intensity set to 20 but was reported to be flexible between 15 and 25 hr per week One child was 18 months and 1 child was 23 months at intake Duration was either 1 or 2 years
30	24	11	11	0	QCT/archival data Included if CA ≤ 46 months and IQ < 35	Postmeasures conducted 3– 4 years after treatment in some cases
<10	24	10	8	2		VABS data available only for 6 of 11 children in the IBI group Control group received minimal treatment
40	24	20	19	1	UCT/enrollment center	Workshops were done every 4– 6 weeks Mix of clinic and parent managed programs Entire caseload of clinician from 80 children enrolled at center
40	12–36	27	23	4	UCT/enrollment center	No. of hr per week between 35 and 45 Follow-up testing done 4–6 years after treatment Duration of treatment 1, 2, or 3 years INDA

Table 1. Continued

| Country/Study/Group | Age | | Pretest | | | | Posttest | | | |
| | | | IQ | | ABC | | IQ | | ABC | |
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Smith et al. (2000)										
IBI (clinic)	36.1	6.0	50.5	11.2	63.7	9.6	66.5	24.1	61.3	28.7
Control	35.7	5.4	50.7	13.9	65.2	9.0	50.5	20.4	59.9	16.7
Smith et al. (2000)										
IBI (parent)	35.8	4.8		11.8	54.8	4.5	58.3	19.3	60.2	16.7
Sallows & Graupner (2005)										
IBI (clinic)	33.7	3.9	48.8	8.8	59.8	5.7	70.8	24.6	63.4	23.6
IBI (parent)	30.2	3.9	44.4	8.2	54.4	5.3	63.8	23.5	58.9	21.8

(Table 1 continued)

Table 1. Extended Continued

Intensity			Gender		Design/Assignment/ Inclusion[a]	Comments[b]
Hr	Mon.	n	M	F		
24.5	24	15	12	3	RCT/matched-pair random	Clinic directed group: no. of hr for IBI group for first year in treatment
<10	~24	13	11	2	Included if CA < 42 months and ratio IQ between 35 and 75	Gradual reductions in Year 2
						Treatment phased out after 18 months for children responding slowly
						Average duration 33 months
						Parent managed group: 5 hr a week of parent training for first 3–9 months, parents asked to do 5 hr a week in between sessions: total < 10 hr per week of ABA + 12.5 hr of special education classes per week
						ABA treatment hr second year presumed to be gradually decreasing, school hr presumed to be the same
						Follow-up testing at CA 7–8 years
						Duration between testing on average 54 months
						Autism and PDD-NOS lumped together in the present analysis
						INDA
26	24	6	6	0	UCT/Consecutive referrals Included if CA < 48 months	Two boys deemed untestable and IQ set to 30
						Posttreatment after 2–3 years
						Children had to be under 48 months at intake
						Average of 26.2 hr a week the first 5 months, after that 30 hr for 5 of the children (1 dropped out)
						Supervision monthly
38	24	13	11	2	RCT/matched-pair random	In order to keep as many variables as possible constant, Year 2 outcome data were obtained from the authors
31	24	11	8	3	Included if CA < 42 months and ratio IQ ≥ 35	Intensity data from Sallows & Graupner (2005)
						INDA

Table 1. Continued

Country/Study/Group	Age		Pretest				Posttest			
			IQ		ABC		IQ		ABC	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Howard et al. (2005)										
IBI (clinic)	30.9	5.2	58.5	18.2	70.5	11.9	89.9	20.9	81.3	11.1
Comparison (AP)[d]	37.4	5.7	53.7	13.5	69.8	10.5	62.1	19.6	69.1	12.9
Control (GP)[e]	34.6	6.5	59.9	14.8	71.6	10.5	68.8	15.3	68.3	9.9
Cohen et al. (2006)										
IBI (clinic)	34.4	5.4	62.0	16.4	64.0	8.4	81.1	21.8	79.5	13.4
Control	33.2	3.7	59.4	14.7	71.9	11.5	65.9	16.5	70.7	13.3

(Table 1 continued)

© American Association on Intellectual and Developmental Disabilities

Table 1. Extended Continued

Intensity			Gender		Design/Assignment/	
Hr	Mon.	*n*	M	F	Inclusion[a]	Comments[b]

25–40
25
15
	14
	13
	15
		29
		16
		16
			25
			13
			16
				4
				3
				0
					QCT/parental preference and IEP teams
					Included if CA < 48 months
						IBI: Multiple settings (home, school, and community)
						  25–30 hr per week under 3 years of age
						  35–40 hr per week over 3 years of age
						Autism educational programming: public classroom for children with autism
						1:1 or 1:2 staff:child ratio
						25–30 hr per week of intervention, supervision by special education teacher
						Intervention eclectic (PECS, SIT, TEACCH, DTT)
						7 children received 1–2 session per week of speech therapy
						Generic educational programming: local community special education classrooms
						Average of 15 hr per week intervention, 1:6 staff:child ratio
						13 children received speech and language therapy 1–2 times per week
						INDA

35–40		24 24	18 17	3 4	QCT/parental preference Included if CA < 48 months and ratio IQ > 35	Community-nonuniversity setting Community services selected by family In Control Group 1, child had an Early Start Autism Intervention Program 9 hr a week 2 children home-based development program 1–4 hr a week 17 special day class eclectic, ratio 1:1 to 3:1, 3–5 days a week for up to 5 hr Speech, behavioral, and occupational therapies 0–5 hr per week 3 where mainstreamed for up to 45 minutes a day INDA

Table 1. Continued

Country/Study/Group	Age Mean	SD	Pretest IQ Mean	SD	Pretest ABC Mean	SD	Posttest IQ Mean	SD	Posttest ABC Mean	SD
Australia										
Birnbrauer & Leach (1993)										
IBI (clinic)	38.1	7.1	45.3	17.9	47.5		57.6	18.7	41.0	5.1
Control	33.2	10.3	45.0	9.4	51.5		43.2	15.0	42.5	4.9
Israel										
Ben-Itzchak & Zachor (2007)										
IBI (clinic)	25.9	3.2	71.4	18.8	65.9	7.1	82.9	23.2	90.7	12.3
Norway										
Eikeseth et al. (2002)										
IBI (clinic)	66.3	11.3	61.9	11.3	55.8	9.0	79.1	18.1	67.0	16.3
Comparison	64.8	9.9	65.2	15.0	60.0	13.2	68.9	18.8	60.2	11.7
Eldevik et al. (2006)										
IBI (clinic)	53.1	9.5	41.0	15.2	52.5	3.9	49.2	16.6	52.4	9.2
Comparison	45.1	16.5	42.8	13.0	50.1	9.2	38.5	15.5	44.6	7.5
United Kingdom										
Hayward et al. (2009)										
IBI (clinic)	35.7	6.2	53.5	15.1	62.3	6.8	72.0	19.6	68.3	14.5
IBI (parent)	34.4	5.7	54.7	15.3	65.1	10.4	69.7	22.9	72.5	17.3
Remington et al. (2007)										
IBI (clinic)	35.7	4.0	61.4	16.7	60.2	5.8	73.5	27.3	61.5	15.4
Control	38.4	4.4	62.3	16.6	57.0	6.8	60.1	27.8	54.6	13.1

[a]UCT = uncontrolled clinical trial, QCT = quasiexperimental controlled clinical trial, RCT = randomized controlled clinical trial. INDA = individual data obtained from author. VABS = Vineland Adaptive Behavior Scales, ABC = adaptive behavior composite. [c]Intensive behavioral intervention.

intake, and intensity of treatment (median split of intensity at 36 hr per week). To protect against some errors of statistical inference, we centered all variables following the guidelines suggested by Kraemer and Blasey (2004). Thus, the binary independent variable (high or low intensity of treatment) was recoded as either $+\frac{1}{2}$ or $-\frac{1}{2}$ and all other independent variables (age, IQ, and ABC scores at intake), by subtracting the median value. In addition to the main predictor variables, we added an interaction analysis between the main predictors. This was done by generating product terms from the centered variables. For IQ change, we included interaction terms for age and IQ at

Table 1. Extended Continued

Intensity			Gender		Design/Assignment/	
Hr	Mon.	*n*	M	F	Inclusion[a]	Comments[b]
19	22	9	5	4	QCT/parent willingness and geographical	Untestable subjects set to IQ of 30
–	24	5	5	0	Included if CA between 24 and 48 months	Ratio scores computed for the rest of subjects
						Scores posttreatment are deviation IQ used where available
						Ratio VABS scores calculated both pre and post
						INDA
35	12	21	20	1	UCT/enrollment center	Only children available for Year 2 follow-ups included here
						5 children 23 months and 2 children 21 months at intake
						INDA
28	12	13	8	5	QCT/staff availability Included if CA between 48 and 84 months and	Comparison received eclectic treatment of similar intensity
29	14	12	11	1	IQ ≥ 50	INDA
13	20	13	10	3	QCT/archival date Included if CA < 72	Comparison received eclectic treatment of similar intensity
12	23	11	10	1	months	4 subjects from comparison group taken out here because included in Eikeseth et al. (2002)
						INDA
37	13	23	19	4	QCT/geographical Included if CA	INDA
34	14	21	15	6	< 42 months	
26	24	23	18	5	QCT/parent preference Included if CA between	Control group received TAU, special school, mainstream or mix, but little or no 1:1, speech therapy, TEACCH etc. INDA
16	24	21	18	3	30 and 42 months	

intake, age at intake and intensity of intervention, and IQ at intake and intensity. For change in ABC, we included interaction terms for age and ABC scores at intake, age at intake and intensity of intervention, and ABC at intake and intensity of intervention.

Results

The proportion of children in intensive behavioral intervention, control, and comparison groups achieving reliable change in IQ and ABC is displayed in Figure 2. Each bar on the graph in

Table 2. Available Number of Subjects and Demographics of Entire Sample and Subgroups

Group	Age				Intelligence				Adaptive behavior			
	n	Mean	*SD*	Range	*n*	Mean	*SD*	Range	*n*	Mean	*SD*	Range
Behavioral treatment (*n*=309)	278	38.0	11.4	16–84	286	55.6	18.2	17–120	252	60.3	10.9	26–95
Control (*n*=309)	95	36.5	7.1	18–72	105	54.8	17.1	19–97	73	65.0	11/6	45–113
Comparison (*n*=39)	39	47.6	15.9	21–84	39	47.6	15.9	21–84	39	61.2	13.7	37–96
Total (*N*=453)	412	38.5	11.5	16–84	430	55.3	17.7	17–120	362	61.3	11.5	26–113

Figure 2 represents an individual child's change in test score. These have been sorted from the highest negative to the highest positive change. A reference line on the y-axis shows the criterion for reliable change. Overall, 83 of the 279 children in the intensive behavioral intervention group (29.8%) achieved reliable change in IQ and 51 of 248 achieved reliable change in ABC scores (20.6%). In the control group, 9 of 104 achieved reliable change in IQ (8.7%), and 4 of 70 achieved reliable change in ABC scores (5.1%). In the comparison interventions group, 1 of 39 children achieved reliable change in IQ (2.6%), and 2 of 39 achieved reliable change in ABC scores (5.7%). We compared the proportions in the three groups statistically using 3×2 chi-square tests. There was a significant difference in the proportions achieving reliable change for IQ, $\chi^2(2, N = 422) = 29.11$, $p < .001$, and for ABC, $\chi^2(2, N = 357) = 11.81$, $p = .003$. Examination of the standardized residuals in the six cells of these two analyses revealed that there were more children than expected achieving reliable change in the intensive behavioral intervention group, and fewer children than expected achieving this change in the two other groups. Exploratory 2×2 chi-square comparisons between the control and comparison group for IQ and ABC change revealed no difference between these two groups, $\chi^2(1, N = 130) = 2.06$, $p < .151$ and $\chi^2(1, N = 96) = .141$, $p = .707$, respectively.

Because the chi-square comparisons showed that there were no significant differences in outcome between the control and comparison groups, we combined them to carry out the individual participant data meta-analysis focusing on the number needed to treat and absolute risk reduction for intensive behavioral intervention. The number needed to treat was computed to be 5, 95% CI = 3.4, 6.3, for achieving a reliable change in IQ and 7, 95% CI = 4.5, 9.8, for achieving reliable change in ABC scores, which translates to an absolute risk reduction of 23%, 95% CI = 16.0%, 29.6%, and 16%, 95% CI = 10.2%, 22.3%, respectively, in favor of the intensive behavioral intervention group. The number needed to treat and absolute risk reduction for IQ and ABC, along with the 95% confidence intervals for the individual studies (i.e., the controlled studies in which there is a comparison or control group against which to calculate an effect size) are shown in Tables 3 and 4. At the level of individual studies, there is considerable variability in effect sizes, and many of the individual studies were focused on small samples and, therefore, were underpowered.

The multiple regression analyses for prediction of IQ and ABC change are summarized in Table 5. A graphical analysis of residuals showed the assumptions of normality and equal variance approximately held. Overall, the models explained a statistically significant, though small, proportion of the variance for both IQ change, $F(4, 211) = 5.22$, $p < .001$, $R^2 = .090$, adjusted $R^2 = .073$, and ABC change, $F(4, 213) = 14.45$, $p < .001$, $R^2 = .213$, adjusted $R^2 = .199$. The results from the regression analyses showed that high intervention intensity was the only variable that independently and positively predicted both IQ and ABC gain. In addition, ABC at intake and IQ at intake predicted gains in ABC. Those children with lower ABC scores at intake had larger ABC change over 2 years, whereas higher IQ at intake predicted larger ABC gains. No interaction terms were statistically significant independent predictors of IQ or ABC change.

Discussion

Despite the recognized difficulties of obtaining individual participant data over a long time period (20+ years of research) (Cooper & Patall,
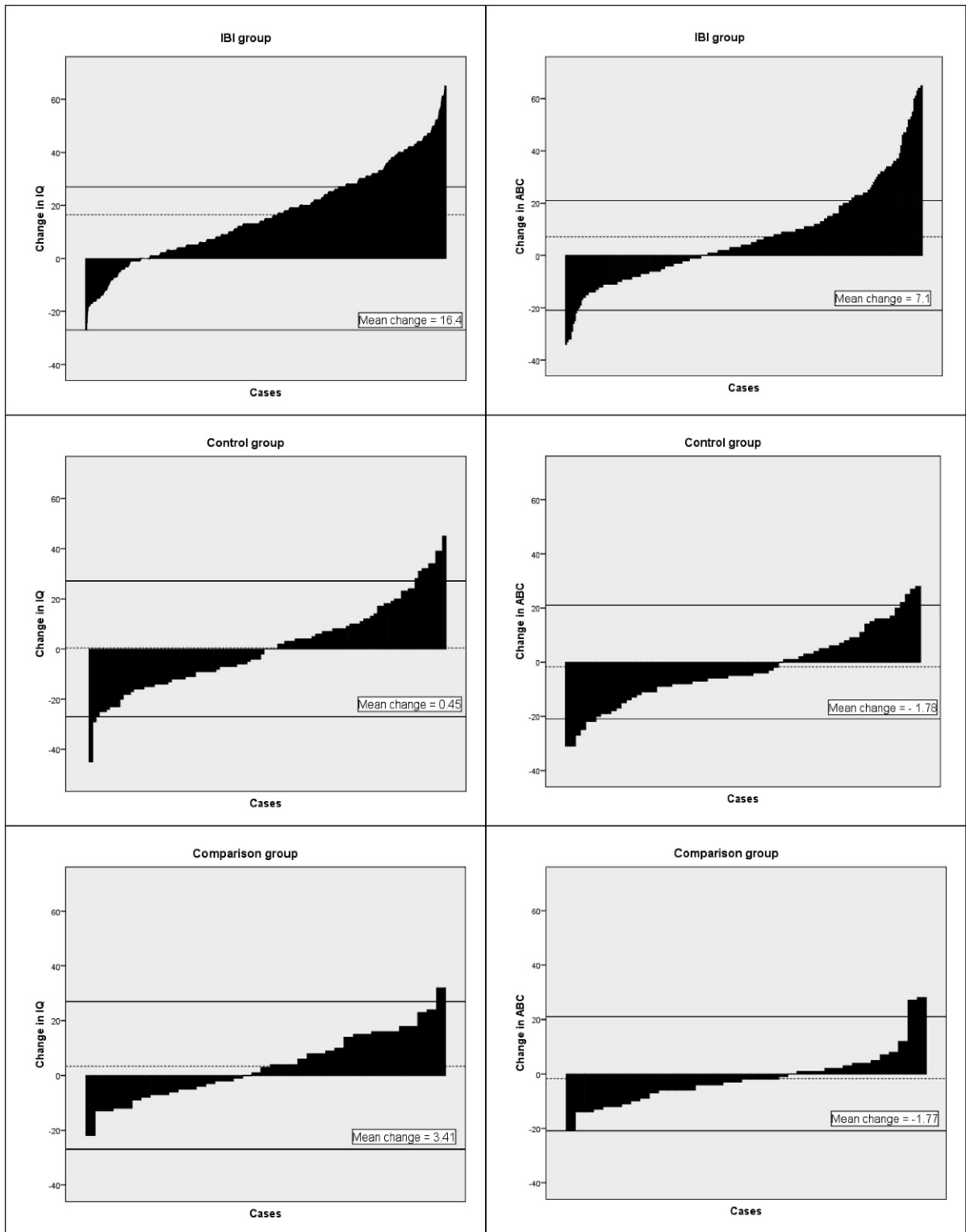
Figure 2. Bars indicate changes in IQ and ABC scores for children in the IBI, control, and comparison groups. The lines at ±27 IQ points and ±21 Adaptive Behavior Composite (ABC) points show the criteria for reliable change. The dotted line shows the mean change for the group.

Table 3. Number of Children Meeting Reliable Change Criteria: Outcome Intelligence

Study/Group	Outcome intelligence[a]					
	RCI+	RCI−	NNT	95% CI NNT/NHH	ARR (%)	95% CI (%)
Lovaas (1987)						
IBI[b]	9	10	3	1,5–5,7	42.0	17.5–66.7
Control	1	18				
Eikeseth et al. (2008)						
IBI	0	10	5	2,2–575.3	23.1	0.2–46.0
Comparison	3	12				
Birnbrauer & Leach (1983)						
IBI	2	7	5	NNT 2,0 to NNH 20,2[d]	22.2	−4.9–49.4
Control	0	5				
Smith et al. (2000) (RCT)[c]						
IBI	6	9	3	1,5–6,6	40.0	15.2–64.9
Control	0	13				
Eldevik et al. (2006)						
IBI	0	13		—	0.0	0.0
Comparison	0	11				
Smith et al. (1997)						
IBI	1	10	11	NNT 3,8 to NNH 12.7[d]	9.1	−7.9–26.1
Control	0	10				
Howard et al. (2005)						
IBI	14	11	3	1,5–10	37.3	10.0–64.5
Control	3	13				
Howard et al. (2005)						
IBI	14	11	3	1,4–3,7	49.8	27.0–72.5
Comparison	1	15				
Cohen et al. (2006)						
IBI	9	12	4	1,7–12,6	32.9	7.9–57.8
Control	2	18				
Remington et al. (2007)						
IBI	5	18	14	NNT 3,3 – NNH 6,6[d]	7.5	−15.1–30.0
Control	3	18				

[a]RCI = reliable change index, the plus sign signifies that criterion for reliable change was met; the minus sign means that criterion was not met. NNT = number needed to treat, NNH = number needed to harm, ARR = absolute risk reduction, CI = confidence interval. [b]Intensive behavioral interventions. [c]Randomized controlled clinical trial. [d]Because the 95% CI for the absolute risk reduction extends from a negative number where treatment may harm (NNH) to a positive number where treatment may benefit, it is hard to compute a 95% CI for the NNT. This means that we cannot say with 95% certainty whether the intervention is harmful, has no effect, or is helpful compared to control. What we can say in this instance is that we can be 95% certain that one of these statements is true: The experimental treatment is harmful (compared to control), and the NNH is greater than x. The experimental treatment is helpful (compared to control), and the NNT is greater than y. Expressed as NNT y to ∞ (indefinitely) to NNH x (adapted from Altman, 1998).

Table 4. Number of Children Meeting Reliable Change Criteria: Outcome Adaptive Behavior

Study/Group	Outcome adaptive behavior[a]					
	RCI+	RCI−	NNT	95% CI NNT/NHH	ARR (%)	95% CI (%)
Eikeseth et al. (2002)						
IBI[b]	4	9	4	1.8–17.6	30.8	5.7–55.9
Comparison	0	12				
Bimbrauer & Leach (1993)						
IBI	0	9	−5	NNT 6.6 ~ NNH 1,8[c]	−20/0	−15.1–55.1
Control	1	4				
Smith et al. (2000) RCT[c]						
IBI	4	11	4	2,0–23.3	26.6	15.2–64.8
Control	0	13				
Eldevik et al. (2006)						
IBI	1	12	13	NNT 4,5 ~ NNH 14,7[c]	7.7	−6.8–22.2
Comparison	0	11				
Howard et al. (2005)						
IBI	5	18	5	2,6–20,5	21.7	4.9–38.6
Control	0	13				
Howard et al. (2005)						
IBI	5	18	11	NNT 3.1 ~ NNH 7,1[c]	9.2	−14.1–32.6
Comparison	2	14				
Cohen et al. (2006)						
IBI	4	15	11	NNT 3,0 ~ NNH 7,4[c]	9.9	−13/4–33.3
Control	2	16				
Remington et al. (2007)						
IBI	2	21	26	NNT 5,4 ~ NNH 9,3[c]	3.9	−10.8–18.6
Control	1	20				

[a]RCI = reliable change index, the plus sign signifies that criterion for reliable change was met; the minus sign means that criterion was not met. NNT = number needed to treat, NNH = number needed to harm, ARR = absolute risk reduction, CI = confidence interval. [b]Intensive behavioral interventions. [c]Randomized controlled clinical trial. [d]Because the 95% CI for the absolute risk reduction extends from a negative number where treatment may harm (NNH) to a positive number where treatment may benefit, it is hard to compute a 95% CI for the NNT. This means that we cannot say with 95% certainty whether the intervention is harmful, has no effect, or is helpful compared to control. What we can say in this instance is that we can be 95% certain that one of these statements is true: The experimental treatment is harmful (compared to control), and the NNH is greater than x. The experimental treatment is helpful (compared to control), and the NNT is greater than y. Expressed as NNT y to ∞ (indefinitely) to NNH x (adapted from Altman, 1998).

2009), we were able to gather such data for each of the 16 evaluation studies of intensive behavioral intervention identified via a systematic review. Only data from one of Lovaas' (1987) original control groups were unavailable. When we compared the intensive behavioral intervention group with control and comparison groups, an individual participant data meta-analysis showed meaningful differences in outcomes for children with autism in favor of intensive behavioral intervention. For IQ, the number needed to treat was 5 (absolute risk reduction = 23%), and for the ABC, the number needed to treat was 7 (absolute risk reduction = 16%). Given that the data for this individual participant data meta-analysis were identified via a systematic review, they might be considered a benchmark against which to evaluate future intensive behavioral intervention outcome

Table 5. Regression Analysis of Predictors of IQ and Adaptive Behavior Composite (ABC) Gain in the Intensive Behavioral Intervention

	IQ gain	
Predictor	β	p
IQ		
Main effects		
IQ at intake	−.135	.031
ABC[a] at intake	.128	.054
Age at intake	−.069	.282
Intensity	.266	.000
Interactions		
Age at intake × IQ at Intake	−.021	.718
Age at Intake × Intensity	−.049	.382
IQ at Intake × Intensity	.014	.811
ABC		
Main effects		
IQ at intake	.363	.031
ABC at intake	−.342	.054
Age at intake	−.038	.282
Intensity	.217	.000
Interactions		
Age at intake × ABC at Intake	.102	.132
Age at Intake × Intensity	.058	.365
ABC at Intake × Intensity	.190	.005

[a]Adaptive Behavior Composite.

studies as well as to audit the outcomes achieved in clinical practice. Such data have not been previously available in the field.

The effect sizes obtained from the individual participant data meta-analysis compare favorably to psychological and medical treatments for common disorders such as major depression (number needed to treat between 3 and 5), obsessive compulsive disorders (number needed to treat between 4 and 5), and bulimia nervosa (number needed to treat = 9) (Pinson & Gray, 2003). We have not been able to locate published number needed to treat or absolute risk reduction data for other interventions for autism. The decision to offer interventions cannot be made by looking at the number needed to treat score in isolation; one would also need to know the intervention costs, long-term economic and social savings, and resources required. Also, any side effects of intervention would be important to document. Full data on these variables are not currently available in the field of autism. Howev-

er, it is informative to note that there appears to be no additional negative psychological impact on family members associated with intensive behavioral intervention (Hastings, 2003; Hastings & Johnson, 2001; Remington et al., 2007). Furthermore, autism-specific eclectic preschool services may cost no less than home-based intensive behavioral intervention (Magiati et al., 2007).

The present analysis provides evidence that intensive behavioral intervention is an evidence-based intervention for children with autism. According to the criteria developed by the Oxford Centre for Evidence Based Medicine (2009), the evidence for intensive behavioral intervention for young children with autism is at Level 1b. This level requires evidence from at least one well-designed randomized controlled study and evidence from systematic reviews. Level 1a (the highest level of evidence) would require a systematic review of several randomized controlled trials showing homogeneity in results. Similarly, the intensive behavioral intervention evidence base meets the criteria for the evidence-based practices in special education proposed by Gersten et al. (2005). These criteria require at least four acceptable quality studies or two high quality studies supporting the practice and a weighted effect size significantly greater than zero (e.g., Eikeseth, 2009), one high quality study (Smith et al., 2000) and four acceptable quality studies (Cohen et al., 2006; Eikeseth et al., 2002; Howard et al., 2005; Remington et al., 2007). Eldevik et al. (2009) found that all of these studies had a weighted effect size significantly greater than zero.

Combined with the earlier meta-analysis of controlled studies reported by Eldevik et al. (2009) based on effect sizes calculated using individual participant data, the present individual participant data meta-analysis completes the two meta-analysis steps advocated by Cooper and Patall (2009). The evidence from the present study also extends the number of studies included in the Reichow and Wolery (2009) aggregated data mean change effect size meta-analysis and, like the Eldevik et al. study, adds a quantitative dimension to earlier systematic reviews (Howlin et al., 2009; Reichow & Wolery, 2009).

An individual participant data analysis vastly increases the power to detect intervention effects (Cooper & Patall, 2009), establishing estimates with reduced error. However, it is clear from Tables 3 and 4 that there is considerable variability in the estimates of effect sizes (number needed

to treat and absolute risk reduction) at the level of individual studies. These tables only include controlled studies that could be used to generate study level effect sizes (i.e., pretest–posttest single group designs are excluded from these tables). In addition, several studies include only very small samples within which one or two children reaching, or not quite reaching, criteria for reliable change on either IQ or ABC can have a large impact on the computed effect sizes. In several individual studies (especially for ABC outcomes), the confidence intervals obtained for the effect sizes precluded any conclusion of likely positive gain or harm for the children in that study. These data have been provided for information purposes and to allow researchers to draw their own conclusions about the variability in outcomes within individual studies. However, these data also confirm the importance of carrying out individual participant data meta-analysis across studies in drawing conclusions about the evidence base of an intervention.

In addition to the variability summarized in Tables 3 and 4, in applying the general common elements of intensive behavioral intervention defined by Green et al. (2002), we may risk combining quite different interventions. For example, we made no distinction between center-based, community-based, or home-based programs. We know that the level and frequency of supervision will have varied between studies, although we did not have access to relevant data. Furthermore, separate intensive behavioral intervention programs are likely to stress the use of techniques differently; some may be based heavily on discrete trial training; others, on incidental teaching; others, on pivotal response training; and still others, on verbal behavior and natural environment teaching. As the field develops, it will be important to complete further meta-analyses based on evaluation studies of interventions sharing a more restricted set of features. At the present time, too few studies are available to enable this task.

We also conducted a large sample analysis of the correlates of outcome within the intensive behavioral intervention group of 309 children. The results from these regression analyses show that high intervention intensity was the only variable that independently predicted both IQ and ABC gain. In both cases, high intensity (36+ weekly intervention hours) was associated with larger gains. In addition, ABC at intake and IQ at intake predicted gains in ABC. Those children with lower ABC scores at intake had larger ABC change over 2 years (perhaps indicating ceiling effects for those who start with higher ABC scores at intake), whereas higher IQ at intake predicted larger ABC gains. No interaction terms were statistically significant independent predictors of IQ or ABC change. These findings generally confirm those of previous research that suggest intensity and intake ability may be associated with outcome in intensive behavioral intervention (Eikeseth et al., 2007; Harris & Handleman, 2000; Lovaas, 1987; Remington et al., 2007). Interestingly, despite the considerable sample size, no hypothesized interactions between variables predicted outcome. It is still likely to be important to explore interactions between predictors of outcome in future research where sample size permits because such interactions may tell us a great deal about the ideal conditions for positive outcomes for intensive behavioral intervention. Our conclusions are limited by the lack of available data on correlates of outcome and also the likely lack of validity of the coding of intervention intensity. There is no substitute for the systematic exploration of moderator effects built into the design of intervention studies (Kraemer, Frank, & Kupfer, 2006), and this is a priority for future intensive behavioral intervention research.

One potential difficulty with our research is that the criteria used to calculate whether an individual child's changes in test scores were reliable might be considered conservative. The reliable change criteria that were computed in the present study required a substantial change in IQ (27 points) and ABC (21 points), arguably representing a significant practical gain, reflecting improvements in the potential for independent living, improved quality of life, a reduced need for professional support, and a reduced economic cost for long-term care and habilitation (J. Jacobson, Mulick, & Green, 1998; Jarbrink & Knapp, 2001). Under many circumstances, a change equivalent to one SD would be considered substantial, especially when using standardized and norm-referenced instruments, such as intelligence scales and the VABS (Weinberg, 1989). Our approach emphasizes the importance of data specific to young children with autism in considering change as a result of intervention. In fact, making the assumption that data from normative samples will apply for children with

autism may lead to overestimates of the impact of an intervention.

Perhaps the most significant limitation of the present individual participant data meta-analysis is the quality of the studies entering the review. We applied several important quality control criteria (e.g., definition of intensive behavioral intervention used, quality of outcome measurement), but we did not exclude studies on the basis of research design (apart from case studies). Specifically, there is a lack of true random assignment to groups (except for two studies), the use of different assessment instruments both within and across studies, and the lack of measures of intervention fidelity. Furthermore, there is considerable variability in the duration of treatment (although we standardized that to a greater degree than would have been possible relying only on published aggregated data from each study). Thus, our results should be viewed as preliminary, and future researchers conducting meta-analyses will need to incorporate research quality selection criteria when the body of randomized studies available for analysis is larger.

References

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 and Treatment of Children*, *10*, 352–366.

Bayley, N. (1969). *Bayley Scales of Infant Development*. New York: Psychological Corp.

Bayley, N. (1993). *Bayley Scales of Infant Development* (2nd ed.). San Antonio: Psychological Corp.

Beglinger, L., & Smith, T. (2005). Concurrent validity of social subtype and IQ after early intensive behavioral intervention in children with autism: A preliminary investigation. *Journal of Autism and Developmental Disorders*, *35*, 295–303.

Ben-Itzchak, E., & Zachor, D. A. (2007). The effects of intellectual functioning and autism severity on outcome of early behavioral intervention for children with autism. *Research in Developmental Disabilities*, *28*, 287–303.

Bibby, P., Eikeseth, S., Martin, N. T., Mudford, O. C., & Reeves, D. (2002). "Progress and outcomes for children with autism receiving parent-managed intensive interventions": Er-

ratum. *Research in Developmental Disabilities*, *23*, 79–104.

Birnbrauer, J. S., & Leach, D. J. (1993). The Murdoch Early Intervention Program after 2 years. *Behaviour Change*, *10*, 63–74.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Butter, E. M., Mulick, J. A., & Metz, B. (2006). Eight case reports of learning recovery in children with pervasive developmental disorders after early intervention. *Behavioral Interventions*, *21*, 227–243.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., & Crits-Christoph, P. (1998). Update on empirically validated therapies, II. *Clinical Psychologist*, *51*, 3–16.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., & Crits-Christoph, P. (1996). An update on empirically validated therapies. *Clinical Psychologist*, *49*, 5–18.

Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Developmental and Behavioral Pediatrics*, *27*, 145–155.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*, 165–176.

Drew, A., Baird, G., Baron-Cohen, S., Cox, A., Slonims, V., Wheelwright, S., et al. (2002). A pilot randomised control trial of a parent training intervention for pre-school children with autism: Preliminary findings and methodological challenges. *European Child and Adolescent Psychiatry*, *11*, 266–272.

Eikeseth, S. (2009). Outcome of comprehensive psycho–educational interventions for young children with autism. *Research in Developmental Disabilities*, *30*, 158–178.

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2002). Intensive behavioral treatment at

school for 4- to 7-year-old children with autism: A 1-year comparison controlled study. *Behavior Modification*, *26*, 49–68.

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 controlled study. *Behavior Modification*, *31*, 264–278.

Einfeld, S. L., & Tonge, B. J. (1995). The Developmental Behavior Checklist: The development and validation of an instrument to assess behavioral and emotional disturbance in children with mental retardation. *Journal of Autism and Developmental Disorders*, *25*, 81–104.

Einfeld, S. L., & Tonge, B. J. (2002). *Manual for the Developmental Behavior Checklist* (2nd ed.). Melbourne, Australia: Centre for Developmental Psychiatry, Monash University and Clayton, Victoria: University of New South Wales, School of Psychiatry.

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.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2009). Meta-analysis of early intensive behavioral intervention for children with autism. *Journal of Clinical Child and Adolescent Psychology*, *38*, 439–450.

Fenske, E. C., Zalenski, S., Krantz, P. J., & McClannahan, L. E. (1985). Age at intervention and treatment outcome for autistic children in a comprehensive intervention program. *Analysis and Intervention in Developmental Disabilities*, *5*, 49–58.

Gabriels, R. L., Hill, D. E., Pierce, R. A., Rogers, S. J., & Wehner, B. (2001). Predictors of treatment outcome in young children with autism. *Autism*, *5*, 407–429.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Green, G., Brennan, L. C., & Fein, D. (2002). Intensive behavioral treatment for a toddler at high risk for autism. *Behavior Modification*, *26*, 69–102.

Harris, S. L., & Handleman, J. S. (2000). Age and IQ at intake as predictors of placement for

young children with autism: A four- to six-year follow-up. *Journal of Autism and Developmental Disorders*, *30*, 137–142.

Harris, S. L., Handleman, J. S., Gordon, R., Kristoff, B., & Fuentes, F. (1991). Changes in cognitive and language functioning of preschool children with autism. *Journal of Autism and Developmental Disorders*, *21*, 281–290.

Hastings, R. P. (2003). Behavioral adjustment of siblings of children with autism engaged in applied behavior analysis early intervention programs: The moderating role of social support. *Journal of Autism and Developmental Disorders*, *33*, 141–150.

Hastings, R. P., & Johnson, E. (2001). Stress in UK families conducting intensive home-based behavioral intervention for their young child with autism. *Journal of Autism and Developmental Disorders*, *31*, 327–336.

Hayward, D. W., Eikeseth, S., Gale, C., & Morgan, S. (2009). Assessing progress during treatment for young children with autism receiving intensive behavioural interventions. *Autism*, *13*, 613–633,

Howard, J. S., Sparkman, C. R., Cohen, H. G., Green, G., & Stanislaw, H. (2005). A comparison of intensive behavior analytic and eclectic treatments for young children with autism. *Research in Developmental Disabilities*, *26*, 359–383.

Howlin, P., Magiati, I., & Charman, T. (2009). A systematic review of early intensive behavioral interventions for children with autism. *American Journal on Intellectual and Developmental Disabilities*, *114*, 23–41.

Ingersoll, B., Schreibman, L., & Stahmer, A. (2001). Brief report: Differential treatment outcomes for children with autistic spectrum disorder based on level of peer social avoidance. *Journal of Autism and Developmental Disorders*, *31*, 343–349.

Jacobson, J. W., Mulick, J. A., & Green, G. (1998). Cost-benefit estimates for early intensive behavioral intervention for young children with autism–General model and single state case. *Behavioral Interventions*, *13*, 201–226.

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

Jarbrink, K., & Knapp, M. (2001). The economic impact of autism in Britain. *Autism*, *5*, 7–22.

Klin, A., Saulnier, C., Tsatsanis, K., & Volkmar, F. R. (2005). *Clinical evaluation in autism spectrum disorders: Psychological assessment within a transdisciplinary framework*. Hoboken, NJ: Wiley.

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, *13*, 141–151.

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes: Clinical, research, and policy importance. *Journal of the American Medical Association*, *296*, 1286–1289.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 1524–1529.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*, *318*, 1728–1733.

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.

Luiselli, J. K., Cannon, B. O. M., Ellis, J. T., & Sisson, R. W. (2000). Home-based behavioral intervention for young children with autism/pervasive developmental disorder: A preliminary evaluation of outcome in relation to child age and intensity of service delivery. *Autism*, *4*, 426–438.

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 disorders. *Journal of Child Psychology and Psychiatry*, *48*, 803–812.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390.

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.

Newsom, C., & Hovanitz, C. A. (1997). Autistic disorder. In E. J. Mash & L. G. Terdal (Eds.), *Assessment of childhood disorders* (3rd ed., pp. 408–452). New York: Guilford Press.

Oxford Centre for Evidence Based Medicine. (2009). Levels of evidence. Retrieved from http://www.cebm.net/index.aspx?o=1025

Pinson, L., & Gray, G. E. (2003). Number needed to treat: An underused measure of treatment effect. *Psychiatric Services*, *54*, 145–146.

Reed, P., Osborne, L. A., & Corness, M. (2007a). Brief report: Relative effectiveness of different home-based behavioral approaches to early teaching intervention. *Journal of Autism and Developmental Disorders*, *37*, 1815–1821.

Reed, P., Osborne, L. A., & Corness, M. (2007b). The real-world effectiveness of early teaching interventions for children with autism spectrum disorder. *Exceptional Children*, *73*, 417–433.

Reichow, B., & Wolery, M. (2009). Comprehensive synthesis of early intensive behavioral interventions for young children with autism based on the UCLA Young Autism Project Model. *Journal of Autism and Developmental Disorders*, *39*, 23–41.

Remington, B., Hastings, R. P., Kovshoff, H., Espinosa, F., Jahr, E., Brown, T., et al. (2007). Early intensive behavioral intervention: Outcomes for children with autism and their parents after two years. *American Journal on Mental Retardation*, *112*, 418–438.

Rogers, S. J., & Vismara, L. A. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child and Adolescent Psychology*, *37*, 8–38.

Roid, G. M., & Miller, L. M. (1997). *Leiter International Performance Scale–Revised: Examiners manual*. Wood Dale, IL.: Stoelting.

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

Scheuffgen, K., Happe, F., Anderson, M., & Frith, U. (2000). High "intelligence," low "IQ"? Speed of processing and measured IQ in children with autism. *Development and Psychopathology*, *12*, 83–90.

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

Sherer, M. R., & Schreibman, L. (2005). Individual behavioral profiles and predictors of

treatment effectiveness for children with autism. *Journal of Consulting and Clinical Psychology*, 73, 525–538.

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 severe 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.

Solomon, R., Necheles, J., Ferch, C., & Bruckman, D. (2007). Pilot study of a parent training program for young children with autism: The PLAY Project Home Consultation Program. *Autism*, 11, 205–224.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Service.

Stahmer, A. C., & Ingersoll, B. (2004). Inclusive programming for toddlers with autism spectrum disorders: Outcome from the children's toddler school. *Journal of Positive Behavior Interventions*, 6, 67–82.

Straus, S. E., Newton, D., & Tomlinson, G. (2004). *Centre for Evidence-Based Medicine. Stats calculator*. Retrieved from http://cebm.utoronto.ca/practise/ca/statscal/

Straus, S. E., & Sackett, D. L. (2005). *Evidence-based medicine: How to practice and teach EBM*. Edinburgh: Elsevier Churchill Livingstone.

Stutsman, R. (1948). *Guide for administering the Merrill-Palmer Scale of Mental Tests*. New York: Harcourt, Brace & World.

Thorndike, R. L., Hagen, E. R., & Sattler, J. M. (1986). *The Stanford-Binet Intelligence Scale* (4th ed.). Chicago: Riverside.

Volkmar, F. R., & Klin, A. (2005). Issues in the classification of autism and related conditions. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders, Vol. 1: Diagnosis, development, neurobiology, and behavior* (3rd ed., pp. 5–41). Hoboken, NJ: Wiley.

Wechsler, D. (1974). *The Wechsler Intelligence Test for Children–Revised*. San Antonio, TX: Psychological Corp.

Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence–Revised*. San Antonio, TX: Psychological Corp.

Wechsler, D. (1993). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: Psychological Corp.

Weinberg, R. A. (1989). Intelligence and IQ: Landmark issues and great debates. *American Psychologist*, 44, 1998–1104.

Weiss, M. J. (1999). Differential rates of skill acquisition and outcomes of early intensive behavioral intervention for autism. *Behavioral Interventions*, 14, 3–22.

Correspondence regarding this article should be sent to Sigmund Eldevik, Akershus University College. BP. 423, Lillestrom, 2002, Norway. E-mail: eldevik@online.no