

Bangor University

DOCTOR OF PHILOSOPHY

Behavioural early interventions for children with autism or intellectual disabilities

Eldevik, Sigmund

Award date:
2009

Awarding institution:
Bangor University

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

***BEHAVIOURAL EARLY INTERVENTIONS FOR CHILDREN
WITH AUTISM OR INTELLECTUAL DISABILITIES***

**THESIS SUBMITTED IN ACCORDANCE WITH THE REQUIREMENTS OF THE UNIVERSITY OF
BANGOR FOR THE DEGREE OF DOCTOR IN PHILOSOPHY**

SIGMUND ELDEVIK

OCTOBER, 2009



BU Deiniol Thesis
2009:P7
30110007985832

Summary

Behavioural Early Interventions for Children with Autism or Intellectual Disabilities

A meta-analysis was conducted to evaluate the effects of early behavioural intervention for children with autism. A review of the literature identified nine controlled studies that met criteria for inclusion. The results of the analysis yielded an average effect size of 1.10 for gain in IQ, and .66 for adaptive behaviour gains. These effect sizes would generally be considered as large and moderate, respectively. Following on this an individual participant data meta-analysis was done. Here, participants from 16 studies were divided into three groups depending on the type of intervention they had received; 309 children had received behavioural intervention, 39 comparison interventions, and 105 were in a control group. Results showed that more children receiving behavioural intervention achieved reliable change in IQ, and adaptive behaviour. This equated to a Number Needed to Treat of 5 for IQ and 7 for adaptive behaviour. Regression analysis showed of this larger sample showed that IQ and adaptive behaviour at intake predicted gains in adaptive behaviour and that intensity of intervention predicted gains in both IQ and adaptive behaviour. Next, I evaluated the utility of behavioural intervention implemented in local mainstream pre-schools. Outcome, particularly on IQ, appeared satisfactory when compared against suggested benchmarks, even though it proved difficult to achieve the recommended intensity of intervention. Finally, I evaluated if behavioural intervention could benefit children with intellectual disabilities. A group of children (n=11) with intellectual disabilities received approximately 10 hours per week of behavioural intervention and another group (n=14) received treatment as usual. After one year, changes in intelligence and adaptive behaviour were significantly in favour of the behavioural

intervention group. In general my¹ results support the clinical implication that at present behavioural intervention should be the intervention of choice for children with autism, and also appears promising for children with intellectual disabilities.

(Thesis contains 44931 words excluding tables and figures)

¹ In the empirical chapters I have used “we” and “ours” instead of “I” and “mine” etc, to denote the active form. This is to reflect the fact that these papers are based on published articles with multiple authors.

Summary.....	3
Behavioural Early Interventions for Children with Autism or Intellectual Disabilities....	3
List of Tables and Figures (page numbers need update when everything else is done).....	8
Preface.....	9
Acknowledgments	11
Declarations	12
Chapter 1: Introduction.....	13
Autism	14
What is Autism?.....	14
Identification, Assessment and Diagnosis	18
Behaviour Analysis.....	20
Historical Background of Behavioural Interventions	22
Comprehensive Interventions	23
Behavioural Intervention Programs.....	28
Overview of Thesis.....	30
Chapter 2: Meta-analysis	33
Methods and Rationale of Meta-analysis	34
Individual Participant Data Meta-analysis.....	38
Chapter 3: Meta-analysis of Behavioural Intervention for Children with Autism	39
Introduction	40
Methods	43
Search Strategy and Data Collection.....	43
Child Measures.....	47
Tests of Homogeneity and Publication Bias	49
Effect Size Measures.....	49
Results	49
Homogeneity	49
Effect Size Measures.....	50
Publication Bias	50
Discussion.....	50
Implications for Research, Policy, and Practice.....	55
Chapter 4: Using Individual Participant Data to Extend the Evidence Base for Intensive Behavioural Intervention for Children with Autism.....	56
Introduction	57
Methods	61
Searching Strategy and Data Collection.....	61
Participants.....	65

Child Measures.....	65
Data Analysis Procedure	67
Results	72
Discussion.....	73
Chapter 5: Behavioural Intervention for Children with Autism in Local Mainstream	
Pre-school Settings	80
Introduction	81
Methods	83
Participants.....	83
Pre-School Setting.....	84
Staff Training and Supervision.....	85
Behavioural Intervention	86
Outcome Measures	87
Results	88
Discussion.....	93
Chapter 6: Cognitive and Adaptive Behaviour Outcomes of Behavioural Intervention for Young Children with Intellectual Disability.....	
Introduction	96
Method	99
Participants.....	99
Child Measures.....	100
Setting	101
Interventions.....	102
Data Collection Procedure.....	105
Results	106
Data Analysis.....	106
Main Findings	108
Discussion.....	109
Chapter 7: General Discussion	
Aggregated Data Meta-analysis	114
Can One Number Summarize Intervention Effects?.....	115
Are we Compiling Different Interventions?.....	116
Is the Quality of the Studies Good Enough?.....	118
Are All Relevant Studies Included?.....	121
Individual Participant Meta-analysis	121
Intervention in Other Settings and with Other Diagnostic Groups	123
General Implications.....	124
How Can Intervention Methods be Further Improved?	124
How Can We Improve The Implementation of EIBI?	126
Measuring Outcome	128
Research on Child Development and Other Populations.....	130

Scaling and Economical Implications.....	132
Improving the Evidence Base Standard.....	132
Summary and Conclusions	133
References.....	135

List of Tables and Figures

Chapter 3: Meta-analysis of Behavioural Intervention for Children with Autism

Figure 3.1: Selection procedure.....	46
Table 3.1: Overview studies	48
Figure 3.2: Forrest plots.....	52
Figure 3.3: Funnel plots.....	53

Chapter 4: Using Individual Participant Data to Extend the Evidence Base for Intensive Behavioural Intervention for Children with Autism

Figure 4.1: Selection procedure.....	66
Table 4.1: Overview studies	69
Table 4.2: Demographics.....	71
Figure 4.2: Bar graphs IQ and ABC change	74
Table 4.3: Reliable change and IQ.	75
Table 4.4: Reliable change and ABC.....	75
Table 4.5: Regression analysis and IQ change.	76
Table 4.6: Regression analysis and ABC change.....	76

Chapter 5: Behavioural Intervention for Children with Autism in Local Mainstream Pre-school Settings

Table 5.1: Description of sample	84
Table 5.2: Outcome	90
Figure 5.1: Bar graphs IQ and ABC change	91
Table 5.3: Correlations of outcome.....	92

Chapter 6: Cognitive and Adaptive Behaviour Outcomes of Behavioural Intervention for Young Children with Intellectual Disability

Table 6.1: Outcome	110
Figure 6.1: Bar graphs IQ and ABC change	111

Preface

There has been an accumulating evidence base for the benefits of behavioural intervention² for autistic children over the last 40 years. However, the intervention remains controversial. There are many questions to which there are no clear answers. Some of the questions that are commonly raised are: Is it the intervention methodology and techniques that make a difference, or is it the intensity of intervention that is important, as it often involves 30–40 hours per week of instruction? How is it, if the results of behavioural interventions are so good, that it is not more of an established standard intervention? How does behavioural intervention compare to other commonly provided interventions? Is there evidence of children requiring less care when they reach adulthood, of living independent lives, or even achieving complete recovery? Is it possible to provide the intervention in a typical community setting on a large scale, without the resources from a research grant or nearby University? Do any characteristics, such as intelligence or age at intake, predict outcome for individual children? Is it the case that only the very young and higher functioning children that are provided high intensity intervention will benefit? Is there perhaps a dose-response relationship between weekly hours of intervention and outcome? Is there a threshold in terms of intensity for when intervention has any effect? If so, is the threshold the same for all children? Should the intervention be stopped when children start school or should it continue, and if so, for how long, and in what form? If this intervention can work for children with autism, is there evidence that it can also benefit children with other diagnoses and perhaps less severe special needs? Do the procedures and techniques have anything to contribute in the typical mainstream classroom for the average child? Does the intervention place unreasonable stress on the child or on the parents? These are some of the questions that are frequently raised when interacting with parents, other professionals, and policy makers. Do we have any solid data that could shed light on any of these questions? The present thesis is an attempt to find out.

² The terms behavioural intervention, behavioural early intervention, intensive behavioural interventions (IBI), and early intensive behavioural interventions (EIBI) are used interchangeably throughout this thesis.

In chapter 1 I will provide a review of the current knowledge on the diagnosis of autism and other pervasive developmental disorders. I will then provide some historical background on behavioural interventions and describe how it is related to behaviour analysis and its applied branch. Then, I will briefly describe how a contemporary a behavioural intervention program may be implemented.

The main questions that I will attempt to answer in this thesis are: What is the current evidence base for behavioural intervention? How does it hold up when you apply common criteria from the field of evidence based practices? And, what may predict outcome of intervention? Also, I ask, is it possible to provide this intervention in a typical mainstream community setting? And finally, could children with intellectual disabilities (not autism) benefit from this intervention?

Acknowledgments

I wish to thank my supervisory committee Richard Hastings and Carl Hughes at the University of Bangor. Without their close and competent guidance this thesis would not have been possible. Also, I need to thank my colleagues and co-authors Scott Cross, Svein Eikeseth and Erik Jahr, for help with compiling data and useful discussions and comments along the way. I would also like to thank Michael Borenstein for advice on carrying out the meta-analysis in chapter 3 and Chris Whitaker for assistance with the statistical analysis in chapter 4. I also thank the authors who kindly contributed individual participant data for the analyses that were conducted in chapter 3 and 4, along with Alison Sharland and Iwona Ondire for coding the papers included in these chapters. A number of supervisors have worked diligently on the intervention programs described in chapter 5 and 6, both at Nordvoll School, department for Early Intervention; Silje Haugland, Birgitte Kaldhussæter, Are Karlsen, Rannveig Nasset, Cathrine Olson, Alfred Steinvik, Hege Tryggestad, Elisabeth Ulvestad, Astri Valmo, Tone Åker and Hege Aarlie; and at the Child Habilitation unit at Akershus University Hospital; Alvdís Roulund and Janne Line Østern. Finally, I would like to thank Alison and Charles Sharland who through the Sharland Foundation have provided the University of Bangor a research grant so that the present investigation could be completed.

Chapter 1: Introduction³

³ Parts of this chapter are based on a manuscript for a book chapter: Eikeseth, S., Jahr, E., & Eldevik, S. (2010). Autisme (Autism). In S. Eikeseth & F. Svartdal (Eds.), *Atferdsanalyse: Teori og praksis (Behaviour Analysis: Theory and Practice)*. Oslo: Gyldendal Fagbokforlaget.

Autism

What is Autism?

Autism was first described by the Austrian psychiatrist Leo Kanner in 1943 (Kanner, 1943). The clinical presentation of autism varies substantially from child to child. The picture depends on, amongst other things, the age of the child, its gender, intelligence and language functioning. Although, this variation (and indeed the disorder itself) most likely has neurobiological explanations, at present these have not been identified. In the diagnostic manuals (DSM-IV; American Psychiatric Association, 1994; ICD-10; World Health Organization, 1992) autism is placed under the more general heading of pervasive developmental disorders along with atypical autism [or what is called Pervasive Developmental Disorder – Not Otherwise Specified (PDD-NOS) in DSM-IV] and Asperger syndrome. Characteristic of all these diagnoses (often referred to as Autism Spectrum Disorders), are impairments in social interaction and communication, and the presence of stereotypical and repetitive behaviours.

Reduced social interaction. Children with autism show little or no interest in play or interaction with other children, including siblings. They will have limited understanding for the needs of others and have trouble understanding how other people feel. In addition to this they will often not react adequately to praise or attention from adults and seldom show social pointing or joint attention. They will avoid eye-contact, and may instead focus on the lower part of the face of others or stare out in the open air (Gillberg, 2007).

Impairment of social communication. About half of the children with autism will not develop functional spoken language (Bryson, Clark, & Smith, 1988), and about 75 % of the children who do develop spoken language will be echolalic (Baltaxe & Simmons, 1981). Most often this will manifest itself as repetition of words and sentences that the child does not understand; as the child acquires more understanding of language the echoed speech will also get more advanced. For those children that develop some functional speech, speech will often appear with abnormal intensity, pitch and/or intonation. A limited use of conventional body

language and gestures, like shaking and nodding of head, waving good bye and pointing are also characteristic (Baltaxe, 1981; Baltaxe & Simmons, 1975; Fay & Schuler, 1980; Schreibman, Kohlenberg, & Britten, 1986). Already before the child is two years of age one can observe delayed (or no) understanding of language, limited mimicking and use of social gestures. Similarly, social imitation will be lacking (such as imitating stirring of food, vacuuming, shaving etc.), and the child may show little or no interest in early interactional play, such as Peek-a-boo.

Circumscribed, stereotypical and repetitive behaviours. Many children with autism will rock back and forth with their torso, wave or shake their hands in front of their eyes, twirl objects, sort objects into long rows, or make the same puzzle over and over again. Many children will be interested in toys, but the toy may not be used in the way it was intended to be used. Some children will be attracted to particular objects or parts of objects. Examples of this include rocks, threads, straws, sand, hair locks and plastic bottles. Glittering, shining or blinking toys often seem particularly fascinating. Objects may be chosen because of how they look, feel, sound, smell or taste. Often times the children will get very distressed if you attempt to prevent or limit these activities. Some children may have a low tolerance for changes in routines, insisting on the same route when going for a walk, or that the meal follows exactly the same sequence every time. Also trivial changes like hanging a new picture on the wall, redecorating or the use of gloves in wintertime may result in despair and temper tantrums.

Early onset. In order to get the diagnosis autistic disorder the child must have impairments in all of these three areas (communication, social interaction and repetitive behaviours). However, the nature and severity will vary from child to child, and no one child will show all of the characteristics mentioned above. In addition, all of the impairments must be observed before the child turns three years of age. At the same time the child has to be at least two years of age before the diagnosis can be made. Some children that are diagnosed may have notable impairments during the first years of life, while others (20–35 %) may have a close to normal development before a developmental arrest or regression is observed (Chawarska, et al., 2007). Hence, parents may sometimes report that the child appeared to have normal development until about 18 months of age, including the

use of single words (mummy, daddy, light etc) but that the use of words gradually disappeared along with other social skills before the child reached two years of age (Chawarska, et al., 2007).

The atypical autism diagnosis should be considered when the child has milder symptoms in one or more of the aforementioned areas, or a later debut. The Asperger syndrome is a more debated diagnosis. According to the diagnostic manuals (ICD-10 and DSM-IV), this diagnoses may be used when the child has had normal language and intellectual development until approximately 6 years of age. There are however at least five different definitions of Asperger syndrome in the published literature, hence it is difficult for professionals to agree on exactly what Asperger syndrome is and how it may be separated from (high-functioning) autism (Klin, McPartland, & Volkmar, 2005). This has led to many professionals avoiding the use of the diagnosis Asperger syndrome altogether, and instead apply the diagnostic term *high-functioning autism* (high functioning autism may by convention be used to describe children with an IQ >70).

Intellectual disability. Epidemiological research shows that between 50 % and 80 % of children with autism also have an intellectual disability (Baird, et al., 2006; Fombonne, 1999). The numbers are lower if children with atypical autism and Aspergers syndrome are included, probably closer to 20 %. The average IQ of children with autistic disorder is reported to be between 55 and 60, while the average for autism spectrum disorders is estimated to be between 75 and 80. Intellectual disability is a condition that is diagnosed before 18 years of age. It is marked by below-average general intellectual functioning and impairment in the ability to acquire the skills necessary for daily living. In more objective terms it is defined as the child scoring more than two standard deviations below the mean of the population (i.e., below 70) on a standardized measure of intelligence and on a measure for general adaptive behaviours. It affects about 1 to 3 % of the population. The causes of intellectual disability are numerous, but a specific reason is found in only approximately 25 % of the cases. The most common causes include exposures to toxins or trauma before or during birth, chromosomal, genetic or metabolic abnormalities and malnutrition.

In spite of intelligence scores in the intellectual disability range many children with autism have well developed skills in circumscribed or limited areas, such as memory for certain facts, numerical skills, music, drawing, and visuospatial skills (e.g., puzzles) (Gillberg, 2007). This will be reflected in their scoring profile on intelligence tests. On such tests many children will achieve higher scores on subtests that contain visuospatial and performance skills and lower scores on subtests that tap into language skills (Ghaziuddin & Mountain-Kimchi, 2004). Most children with autism will be motorically clumsy, especially as they come into school age. Clumsiness is indeed proposed by some as a diagnostic criterion for Asperger Syndrome (Klin, McPartland, et al., 2005).

Other common co-morbid disorders. Autism is associated with a host of other problems and co-morbid disorders. Although it presents special diagnostic challenges it has been found repeatedly that around 70 % of the children with autism also fulfil criteria for another psychiatric disorder (Leyfer, et al., 2006; Simonoff, et al., 2008). Attention Deficit Hyperactivity Disorder (ADHD) and anxiety disorders are amongst the most common; however the percentages vary a great deal from study to study. According to the current diagnostic manuals one cannot get the diagnosis of ADHD (and a number of other psychiatric diagnoses) and autism at the same time. However, many researchers and clinicians will still apply two (or more) diagnoses, arguing that this will better describe the child's problems.

Prevalence. Autism spectrum disorders are among the most common developmental disorders. The rate is estimated to be between 3 and 11 per 1000, whereas autistic disorder is estimated to be between 1 and 4 per 1000 (Baird, et al., 2006; Fombonne, 2003). But, the numbers are uncertain, and vary from report to report. There is no doubt however; that many more children are receiving a diagnosis now than compared to what was usual up until 10–15 years ago. There is no evidence that there are actually more autistic children now than previously, it is more likely that the increase is due to added awareness and professional expertise in making the diagnoses (Gillberg, 2007). The ratio between boys and girls also varies from study to study, but the average is estimated to be approximately 4:1 (Fombonne, 2005). The rates differ however with IQ; in children with lower IQs the ratio is 2:1.

Causes. Researchers and clinicians have moved away from fruitless psychodynamic explanations and started looking for biological causes. Research shows that autism may have a number of causal factors, and that these are neurological in nature. As neurobiological research develops, it is likely that autism will be divided into subgroups based on the underlying causes. In the future we will thus no longer talk about autism referring to one group of children (Volkmar & Klin, 2005). The specific causes of autism spectrum disorders are presently not known, but it is clear that genes and conditions during pregnancy, birth and early life are related to these diagnoses. One relatively robust finding is that approximately 30 % of children with autism have elevated levels of serotonin (Buitelaar & Willemsen-Swinkels, 2000). Furthermore, research has shown that autism may have a genetic component (Bailey, Le Couteur, Gottesman, & Bolton, 1995). Also, a number of medical conditions are associated with autism. It is estimated that almost 20 % have some form of epilepsy, and conservative estimates suggest that approximately 5.5 % have some other clear cut medical diagnosis, for example, cerebral palsy, fragile x, tuberous sclerosis or Down Syndrome (Fombonne, 2005).

If you already have one child with autism, the likelihood of your next child also having autism is increased 50–100 times (or to about 5 %) (Bolton, Macdonald, Pickles, Rios, & et al., 1994; Piven, Gayle, Chase, Fink, & et al., 1990). The likelihood of a sibling with atypical autism increases to about 25 %, and the likelihood of other types of problems (e.g., problems with language, learning and social behaviour) will also increase.

Identification, Assessment and Diagnosis

Researchers and clinicians have not yet been able to find biological markers or laboratory tests that can identify autism. The screening and diagnosis is thus only based on observations and measures of the child's behaviour. Screening can start when children are approximately one year old. Screening involves short observations using checklists where one can identify those children who should be offered a more comprehensive assessment. There are several good screening instruments that may be used by personnel with no particular training or experience. The most widely used and validated is probably the Checklist for Autism In Toddlers (CHAT) (Baird, et al.,

2000; Baron-Cohen, Allen, & Gillberg, 1993). CHAT may be used when children are approximately 18 months old, and can be conducted in 15–20 minutes. It consists of nine questions to the parents, and a further five items that are based on observations of the child. At this screening it is expected that the child should point (with the index finger) to show another person something and also to follow where another person is looking and/or pointing. Furthermore, the child should have started to engage in pretend play (e.g., pretend to pour coffee into a cup). Failure to show one or more of these milestones may indicate that the child is at risk, and further assessments are recommended. The same is true, if the child independent of age should lose language or social skills. To identify these children as early as possible, screening will need to be done routinely as part of the check ups at about 18 months (Coonrod & Stone, 2005).

The further assessments of the child need to be done by professionals with special training and experience in diagnosing children with developmental disorders. This assessment is based on several components; observations of the child, interviews with the parents, standardized assessments, medical and neurological check ups (Filipek, et al., 1999). The Autism Diagnostic Interview-Revised (ADI-R) (Lord, Rutter, & Le Couteur, 1994) is considered to be the corner stone of the assessment. The ADI-R follows strictly the ICD-10 criteria for autism. It is a semi-structured interview with parents or caregivers that takes about 2 hours. In addition to the ADI-R it is recommended to use the Autism Diagnostic Observation Schedule (ADOS) (Lord, et al., 2000) for direct systematic measures of communication and social skills. The ADOS has standards for children at various developmental levels, and is particularly helpful as a supplement to the ADI-R for diagnosing atypical autism.

Intellectual development for children under three and a half years of age (or children with a mental age below this level) can be measured with the Bayley Scales of Infant Development-III (Bayley, 2006). For children above this age and that have some spoken language, the Wechsler scales (WISC-III, WPPSI-III) (Wechsler, 1993, 2002) or the Stanford-Binet (SBV) (Roid, 2003) are most widely used and recommended (Klin, Saulnier, Tsatsanis, & Volkmar, 2005). To assess general adaptive behaviour in daily life the Vineland Adaptive Behavior Scales (VABS); (Sparrow, Cicchetti, & Balla, 2005) are considered to be the most useful (Klin, Saulnier, et al., 2005;

Chawarska & Bearss, 2008). This is a semi-structured interview with parents that measures skills within communication, socialization and daily life. It also contains a section for measuring motor development for children less than 6 years of age and a brief measure of maladaptive behaviours. The VABS is standardized in the same way as most intelligence tests, having an average score of 100 and a standard deviation of 15. The VABS will also yield age-equivalent scores. Generally it is important to diagnose children with autism as early as possible so that an intervention could be started. A diagnosis will usually have great influence in determining the resources the child will receive for intervention and other purposes.

Behaviour Analysis

Behaviour analysis consists of four main branches: radical behaviourism, the experimental analysis of behaviour (EAB), applied behaviour analysis (ABA), and service delivery (Cooper, Heron, & Heward, 2007). Radical behaviourism is the philosophy of the science of behaviour. Skinner originally used the term *radical* to convey that this version of behaviourism encompasses all behavioural phenomena. It is not limited to the study of overt behaviour, but includes private events such as feelings and thinking (Skinner, 1945). Skinner established EAB as a field within the natural sciences with the goal of discovering relations between behaviour and environmental variables. Through thousands of laboratory experiments the basic principles of operant conditioning has been described. One of the most basic principles in operant conditioning is that behaviour is selected by consequences that follow the behaviour. This process is similar to the natural selection of traits that Darwin described as one of the basic principles of evolution. The principles of operant conditioning are the foundation of ABA. ABA may be defined as the science in which tactics derived from the principles of behaviour discovered largely in the EAB are applied to improve socially significant behaviour, and experimentation is used to identify the variables responsible for behaviour change (Cooper, et al., 2007)

Contemporary ABA began with the founding of the Journal of Applied Behavior Analysis in 1968. The first issue contained a paper recommending what dimensions applied behaviour analysis should be based on (Baer, Wolf, & Risley, 1968). It

describes the following seven dimensions; application, behavioural, analytic, technological, conceptual systems, effective, and generality.

Applied. This dimension addresses the question: How immediately important is this behaviour or skills to this subject? This must be determined by the interest in which society shows for solving the problem and not by the research procedure or some theory. In other words ABA should focus on areas that are of social significance to individuals.

Behavioural. ABA is pragmatic. It addresses how it is possible to get someone to do something better or more effectively. This can mean increasing behaviour, decreasing behaviours or teaching new skills. This implies that we have to study and measure the behaviour of interest and not how we talk about that behaviour. It also implies that we must be able to demonstrate believable control over the behaviour that is targeted for change. By applying single-subject research designs such as reversal designs or multiple baseline designs, such control may be shown. This dimension also speaks to Skinner's believe that behaviour, and not mind or cognition, is the subject matter of interest.

Technological. Technological within behaviour analysis refers to the description and verification of procedures and interventions such that they become part of the evidence base, and are described in enough detail so that they may be replicated by other professionals facing similar clinical issues.

Conceptual Systems. All tactics and procedures within ABA are based on the principles of behaviour. It is therefore recommended that procedures are described in terms of the relevant principles from which they were derived.

Effective. An application behaviour analysis must result in practical changes. How much change that needs to occur for it to have practical significance is often best determined by society and not the behaviour analyst.

Generality. Any changes produced needs to last over time, in different environments, and spread to other behaviours not directly treated by the intervention. Also, continued change in the behaviour after intervention has ceased is a criterion for generality.

Behavioural intervention for children with autism and intellectual disabilities is thus part of a larger scientific endeavour. At its basis is the application of behaviour principles to socially significant behaviour and as such it is solidly placed within the applied branch of behaviour analysis. Even though behavioural intervention is provided in many forms and under many names, the intervention must entail a common set of core dimensions in order for it to be called behavioural. In addition to the general dimensions mentioned above more specific elements have been proposed (Green, et al., 2002, Eikeseth, 2009). What these more specific elements should be is a matter of some debate but most agree that further research is needed to clarify this. True to the dimensions of applied behaviour analysis the field of early intervention has steadily accumulated knowledge and evidence for almost 50 years. It may be useful next to consider this history in more detail and some of the most important findings.

Historical Background of Behavioural Interventions

Early in the 1960s behavioural psychologists started to study this group of children, and since then have published hundreds of scientific papers investigating how operant principles can be utilized to help persons with autism (Cunningham & Schreibman, 2008; Matson & LoVullo, 2008; Matson, Matson, & Rivet, 2007; Matson & Smith, 2008). Within the psychoanalytic tradition that was dominant at the time, parents and in particular the mother, was blamed for the child's problems. The mother was said to be cold and uninterested in the child, and treatment would often include such elements as forced holding in an attempt to build a bond with the child, mimicking the child's behaviour to get into its world and psychoanalysis of the mother. Both the theory and interventions that sprung out of it have provided more harm than good, often adding to the despair of parents that were already deeply worried about their child (Cipani, 2008).

Another approach was taken by behavioural psychologists who started to consider autism as a behavioural phenomenon and not a symptom of some underlying emotional psychological disorder. One of the assumptions was that autism involved a problem in learning, and that lack of development could be related to the fact that

social stimuli, like intimacy, praise and social attention did not reinforce the behaviour of these children in the same way that it reinforces behaviour in other children.

In the first study systematically applying reinforcement contingencies it was shown that children with autism could be taught arbitrary behaviours (Ferster & Demyer, 1962). Then, came studies showing that it was possible to condition social stimuli as reinforcers (Lovaas, Schaffer, & Simmons, 1965). Building on these early studies, new studies were conducted demonstrating that through the systematic application of operant principles, previously mute children could learn to imitate language (Lovaas, Berberich, Perloff, & Schaeffer, 1966) and acquire some functional speech (Hewitt, 1965; Lovaas, 1977; Risley & Wolf, 1967). Also, it was shown that aggressive behaviour and stereotypical and ritual behaviours could be understood and modified using operant principles. When investigating the behaviour of children with autism in more detail it was found that many had a special handicap with learning in that they often responded only to irrelevant or incomplete aspects of stimuli (this phenomenon has been called stimulus overselectivity; Lovaas, Koegel, & Schreibman, 1979). In the last decades the focus of research has shifted gradually to more complex areas such as social interaction, advanced language, play and integration into mainstream settings (Jahr, Eldevik, & Eikeseth, 2000; Matson, et al., 2007).

Comprehensive Interventions

In 1973 Lovaas and his colleagues presented the first evaluation of a comprehensive behavioural intervention (Lovaas, Koegel, Simmons, & Long, 1973). This study is by many considered the start of the cumulating evidence base for broader intervention principles and programmes, and it identified a number of problems and challenges. First, one saw little transfer of skills learned in the intervention setting to new settings and persons. Second, there was little or no transfer of skills from one skill area (e.g., language) to other skill areas (e.g., social interaction). Third, skills taught and established during the intervention were not always maintained after intervention ceased. And fourth, the youngest children and those who were treated at home rather than in an institution or clinic made better progress.

In an attempt to address these shortcomings intervention was changed so as to: (a) cover all important areas in the child's life, (b) take place in the child's natural environment and not in institution or clinics, (c) include parents and other caregivers so that they could be actively involved, (d) be more comprehensive in terms of intensity and duration and to include most of the child's waking hours, and (e) start when the children were very young. These elements were incorporated in the Young Autism Project conducted through UCLA and the results were published in 1987 (Lovaas, 1987). This project included an intensive intervention group of 19 children who received 40 hours per week of intervention for at least 2 years, a control group of 19 children who received less than 10 hours a week of intervention, and another control group of 21 children that was recruited from a nearby hospital clinic that received treatment as usual. All children were below four years of age when intervention started, and the three groups were similar on all measures at intake such as developmental level, language, play skills and level of stereotypy. Assessment of the children after an intervention period of 2–3 years showed substantial differences between the groups. Almost half of the children (9 out of 19) in the intensive behavioural intervention group had normal intellectual functioning, and had completed first grade in mainstream school without any special assistance. Only one child from the two control groups achieved similar outcome. Most children from the control groups were in special classes for children with autism and/or intellectual disabilities. A follow-up study conducted when the children were on average 13 years, showed that the effects had endured in the long term, and that 8 out of the 9 children that were in the best outcome group were indistinguishable from their peers on a series of standardized tests and clinical interviews (WISC-R, Vineland Adaptive Behavior Scales, Personality Inventory for Children) (McEachin, Smith, & Lovaas, 1993).

The results of this project were quite extraordinary, and naturally led to calls for replications from the scientific community. Moreover, the study was criticized on several grounds (Gresham & MacMillan, 1997; Schopler, Short, & Mesibov, 1989), and in particular with respect to the methodology. Lovaas and colleagues have rebutted the criticism on several occasions (Eikeseth, 2001; Lovaas, Smith, & McEachin, 1989; Smith & Lovaas, 1997) but have agreed and indeed themselves pointed out some of the shortcomings. In particular this pertains to threats to internal

validity due to the lack of true random assignment and the limitation of the various assessment instruments that were employed (Smith, Mruzek, & Peyton, 2008).

In spite of the critique it was widely accepted that the children who received behavioural intervention in this project showed significant improvement in their level of functioning, and that this progress was explained by the intervention (Baer, 1993; Schopler, et al., 1989). Moreover, there was widespread agreement to the importance of replicating the study (Foxy, 1993; Kazdin, 1993; Smith, McEachin, & Lovaas, 1993). Following the publication of this study, the number of studies evaluating early intensive behavioural intervention (some directly replicating Lovaas' study), has gradually increased, particularly in the last ten years (Cohen, Amerine-Dickens, & Smith, 2006; Eikeseth, Smith, Jahr, & Eldevik, 2002; Eikeseth, Smith, Jahr, & Eldevik, 2007; Eldevik, Eikeseth, Jahr, & Smith, 2006; Hayward, Eikeseth, Gale, & Morgan, 2009; Howard, Sparkman, Cohen, Green, & Stanislaw, 2005; Remington, Hastings, Kovshoff, Espinosa, Jahr, Brown, Alsford, Lemaic, & Ward, 2007; Sallows & Graupner, 2005; Smith, Groen, & Wynn, 2000). The research may now be in a phase where there is a need to take stock and also attempt to specify variables that can influence the outcome of behavioural intervention. These include variables associated with the intervention itself such as the intensity and duration, but also child characteristics such as age and IQ.

Intensity of intervention is often mentioned as one of the key variables related to outcome of intervention (Eldevik, et al., 2006). However, a widely adopted operational definition of intensity does not exist at present. Usually, intensity is reported as the number of weekly intervention hours. This is a rather crude measure, and other measures have been suggested and applied in some studies such as the average number of learning opportunities per hour. In one study children were reported to have 50–100 opportunities per hour presented as discrete trials, incidental teaching or other behavioural analytic procedures (Howard, et al., 2005). A relationship has also been reported between monthly intervention hours and the numbers of mastered behavioural objectives for children under the age of seven. This finding however, was not evident for children between 7 and 12 years of age (Granpeesheh, Dixon, Tarbox, Kaplan, & Wilke, 2009).

Thus, there may be a dose-response relationship between weekly hours and gains made on outcome variables, but at the same time there is most likely also some threshold value for when the intervention can result in clinically significant effects (Eikeseth, 2001). Complicating the issue further is the large individual variation amongst children with autism. Thus, the necessary intensity of intervention is likely to vary across children. After a review of the literature Green (1996) concluded that to be safe we should recommend at least 30 hours per week, and that further research was needed before more precise recommendations could be given. Some researchers have suggested that it is the intensity itself that is important irrespective of the type of intervention (Spreckley & Boyd, 2009). However, results of studies comparing behavioural intervention to eclectic intervention of similar intensity suggest that the type of intervention and not intensity in itself is important (Eikeseth, et al., 2002; Eldevik, et al., 2006; Howard, et al., 2005). Even though the intensity seems to be related to the effects of intervention, further specifications are needed. Green (1996) also suggested that other measures of intensity need to be developed such as the number of opportunities to respond (as was reported by Howard et al., 2005). A further specification could be to identify the amount and quality of learn units (instruction-response-consequence) occurring in the intervention (Warren, Fey, & Yoder, 2007).

The duration of treatment is presumed to be important with respect to early intervention in general (Guralnick, 1998) and behavioural intervention in particular (Lovaas & Smith, 1988b). In Lovaas' pioneer study the duration was 2–3 years. In one study it was found that children that showed good progress the first year also had good progress over the next years, whereas children with limited or no progress the first year, continued this trend (Sallows & Graupner, 2005). Even if you can achieve significant gains in some areas after the first year of intervention a more holistic progress, including the whole spectre of social skills may require a substantially longer intervention period.

The child's level of functioning may also be related to effects of intervention. Some studies have found that children with higher scores on intelligence tests benefit more from intervention (Harris & Handleman, 2000; Lovaas, 1988). Another, variable that could influence outcome is the child's age at intake (Fenske, Zalenski, Krantz, &

McClannahan, 1985; Harris & Handleman, 2000). So far however the results reported in the literature are not clear, with some studies reporting that children below the age of three at intake may have the best results while others have reported as good results with children up to 7 year of age at intake (Eikeseth, et al., 2002; Eikeseth, et al., 2007). The impact of early intervention has been linked to early plasticity in the nervous system (Lovaas & Smith, 1989); this rests on the assumption that neuro-developmental disorders or damages to a certain degree may be compensated through intensive stimulation and intervention (Stein, Brailowsky, & Will, 1995). Even if it is clear that environmental experiences lead to neurological changes (Fox, Calkins, & Bell, 1994), there has been no research investigating how intervention may affect neurological development in children with autism.

In summary the findings seems to be in concert with a behavioural systems approach to child development (Novak & Pelaez, 2004). In this approach both phylogenic and ontogenic factors interact to determine development. One example of this may be how the mother responds to her baby's babbling with baby talk (mainly result of phylogenic interaction), and this in turn may change the baby's babbling into words (mainly result of ontogenic interaction), which in turn changes the mothers baby talk into more advanced language. Another example has recently been described in how genes and environment interactions may be related to the development of challenging behaviours (Langthorne & McGill, 2008). Similarly, Lovaas and Smith (1989) have proposed a theory where autistic behaviour is viewed as a result of a mismatch between the child's nervous system and the typical environment. Development may be put on the right track if the environment is changed in special ways so that the child can acquire certain skills.

Development in all these examples is seen as a transactional relation between genetic, historical and current environments. In early intervention personal factors will interact with the (special) environment to produce a certain outcome. Given the vast variability in these factors it should come as no surprise if the outcome is quite variable as well.

Behavioural Intervention Programs

A thorough assessment of the child's cognitive, language and adaptive level should be done before commencing on the intervention program. This assessment will provide a baseline for measuring effects of intervention, and information that will make it easier to tailor the program to meet the child specific strengths and weaknesses. The assessment process will often also yield useful information about the child's attention span and motivational preferences.

Most children with autism will initially gain little benefit of a looser teaching and interaction form such as group teaching, circle games and assemblies. It has proven much more efficient to establish a 1:1 teaching format. Furthermore, teaching will work better in a separate room, where there is less distraction from others. These conditions will make it easier for the child to understand the relation between his or her own behaviour and its consequences. The long-term goal however, is to teach the child the necessary skills to partake and benefit from more natural teaching and interaction forms and a regular classroom.

When beginning the intervention it is important to first establish a good, positive and predictive relationship platform with the child. Often, it is necessary to first do just fun activities with the child, and then gradually introduce the teaching format. Usually this is done with tasks that are easy to do for the child such as completing an inset puzzle or building a tower. This is done in very short sessions, perhaps only one or two trials. Eventually programs are made for all important areas of functioning, such as communication, play, social interaction and independence. Each skill is broken down into simple manageable steps which later can be chained together or built upon. These methodologies are referred to as discrete trial teaching (DTT) and task analysis. In DTT, the instruction is separated into learn units where first a clear and precise instruction/material is presented, the child's response to this is then defined along with procedures for how to help and fade help, and what consequence the child's responding should have. The intervention consists of many such discrete learn units. The individual variations with respect to acquisition rate and potential for development are vast. Also the skill profile varies and makes it necessary to tailor the program for each individual child. There are several intervention manuals where general outlines and plans for progression are provided (Leaf & McEachin, 1999;

Lovaas, 2003; Maurice, Green, & Luce, 1996; Sundberg & Partington, 1998). Most children entering the program need to start with basic skills such as imitation, matching to sample, language comprehension, and working on expressing wants and needs either by talking or through signs or pictures. Progression will move from imitation of gross motor movements (e.g., clapping hands, waving, touching head) to fine motor (e.g., hand and finger movements) to oral motor (e.g., blow, open mouth, tongue out). The child will then be better prepared for imitation of play in natural settings (e.g., push the train on a track, hug a doll, and pretend to drink from cup). These play responses may in turn be chained into imitation of play sequences (e.g., tea party, train and track, and doll play).

Matching to sample is a skill area that is usually started in parallel with imitation in the introductory phase. This program in its most basic form consists of putting similar things together (e.g., putting one block with another identical block) and can involve both objects and pictures. Basic receptive language skills will often be learned quicker once the child has mastered matching to sample. In receptive language the child is usually first taught to point to (or get, or give) objects or body parts. A child that has mastered gross motor imitations will be better prepared for imitation of oral motor movements and building on that, imitation of sounds and words. These sounds and words are then made functional so they can be used to express basic wants and needs, labelling objects or characteristics of the objects (e.g., colour, shape, size category etc). Based on what the child masters in vocal imitation and labelling, one can progress to teach the child to use phrases to describe events or objects in the environment. This increased understanding and use of sentences makes it possible to work on more advanced skills in play and communication. Examples of this may include asking and answering basic questions and taking part in conversations and games. In play the child could be taught turn-taking, and to build on topics in cooperative play, role-play and symbolic play.

When imitation and instruction following have reached a certain level, it is usually much easier to teach various skills that make the child more independent, such as eating with knife and fork, dressing (zippers, tying shoe laces), toileting and sleeping in his or her bed for the entire night. An important dimension of this intervention is to arrange so that the child can practice and use skills in natural settings immediately.

For instance one can do circle games where the child has to use the newly acquired skills with other normal children and/or adults (e.g., imitation games, labelling of objects, introduce oneself and asking/answering questions).

The above is an outline, variations will occur both in the order and general progress. The main principle of this model is that one can gradually build more complex skills in a step by step fashion. Some skills can be taught in parallel because they are developmentally independent and do not interfere with each other (e.g., imitation and matching). Some skills established at one level will facilitate skills at another and more advanced level (e.g., oral motor imitation facilitates learning of vocal imitation). Progress in these early programs may give some clue for what sort of progress to expect over the next 2–3 years of intervention (Sallows & Graupner, 2005). The progress over the first weeks and months of intervention will also give clues as to what sort of communication to choose for the child. For instance a child that readily acquires oral motor imitation will move on to vocal imitation. For this child spoken language would be the primary goal. Another child could struggle mightily with basic vocal imitation even after extensive teaching. In this case a picture based communication system or sign language may work better initially.

Overview of Thesis

The remainder of this thesis is organised in to six chapters. The next chapter will provide a brief introduction to the methodology of meta-analysis. The next four chapters after that are empirical in nature (chapters 3 to 6); these are followed by a final general discussion chapter.

Although the evidence base for behavioural intervention is expanding and reviews have recently been published including two meta-analyses, they all have some obvious limitations. Chapter 3 is an attempt to address these shortcomings and conduct an updated quantitative review of the evidence base, using state of the art meta-analytic techniques. Hedges' *g* mean difference effect size measures are computed for changes in intelligence and adaptive behaviour, and the outcome data is also analysed to see if it is homogeneous and if any likely publication bias exists in the literature. To increase the strength of the analysis over previous published

reviews, only studies that had a control or a comparison group were included and a more precise definition of behavioural intervention is employed in the study selection procedure.

In chapter 4 a more in-depth analysis of the outcome is conducted. This was done by compiling individual participant data from the same studies included in the meta-analysis and some additional uncontrolled studies. Thus, I expanded the meta-analysis to what is sometimes referred to as a mega-analysis or an integrative data analysis (Cooper & Patel, 2009). Compilation of individual participant data opens the possibility for application of other (and better) means of statistically analyzing the data and dividing children into new groups. Hence, first I assigned the individual children to one of three groups; behavioural intervention, comparison or control. Then I applied a recently suggested measure of clinical significance; the reliable change index. This index will tell you at which level you can be 95 % certain the changes are due to the intervention and not measurement error or variability in the sample. Individual children could then be divided into two groups, those who met these criteria, and those who did not. This approach made it possible to apply state of the art effect size measures such as the Number Needed to Treat and Absolute Risk Reduction. This large sample of individual data also made it possible to apply solid statistical methods to see if any of any the commonly proposed variables really predicted outcome. Furthermore, I could look for possible interaction effects amongst these. While the outcome of these exercises are in themselves interesting, the main goal of this was to provide a better and more simple way of communicating the potential benefits of behavioural intervention to policy makers and other funding agencies.

In chapter 5 I ask the question of whether the intervention can be implemented in a typical mainstream community setting. Although the outcome of intervention has been replicated in a number of different settings, the multitude of local adaptations needed is probably endless. A program in Oslo, Norway was set up in order to provide behavioural intervention in local mainstream pre-schools. The program provided supervision and training to staff in mainstream pre-schools who had enrolled a child with autism. The staff was in turn responsible for running the day-to-

day intervention. The outcome and strengths and weaknesses of this model are discussed.

A number of children without a diagnosis of autism, but who had a diagnosis of an intellectual disability, were also referred to the program evaluated in chapter 5 and to a nearby Child Habilitation Unit. Through collaboration with them, I was able to compare a group of children with intellectual disability who were provided behavioural intervention to a group of children who were provided standard special education. To our knowledge this is the first attempt to evaluate outcome of behavioural intervention for children that are not on the autism spectrum. The promising results from children with autism suggest that the intervention may also benefit other populations and that the methods may indeed also be adopted for the general education of all children. This will constitute chapter 6 of this thesis. Chapter 7 offers a discussion of the main themes while returning to the central questions and aims of this thesis.

Chapter 2: Meta-analysis

Meta-analysis was created out of the need to provide a quantitative summary of the results from several individual studies in a user friendly way (Cooper & Patall, 2009). The most usual form of meta-analysis is to analyse aggregated data from individual studies. Aggregated simply means that the unit of analyses is the average group scores (preferably both from the experimental group and the control group) reported in the studies included in the review. A more recent development has been individual participant data meta-analysis. Here, the unit of analyses is data from the individual participants included in the reviewed studies. Again, it is preferable to have individual data both from participants in an experimental group and a control group, and this type of review opens up possibilities for more sound analysis. In general, meta-analysis has some notable advantages compared to the more traditional narrative review. The selection of which studies to include in the review are more transparent and more objective quantitative criteria are applied to evaluate the effects of the intervention. The results of a meta-analysis are commonly reported as the effect-size. Effect sizes are often considered on a scale where 0.2 to 0.3 might be a "small" effect, around 0.5 a "medium" effect and 0.8 to infinity, a "large" effect.

Methods and Rationale of Meta-analysis

Meta-analysis is in many ways parallel to the classic scientific method. First a problem is described and a research hypothesis is formulated such as: Is intervention X better than intervention Y?, or; Is intervention Z effective?. The next step is making transparent objective guidelines for which studies to include in the review. The aim is to include a population of studies rather than a selected sample that might be biased. The sampling process can include electronic searches of databases, browsing abstracts and reference lists, contacting experts in the field and so on. As of today there is no standardized reliable technique for selecting studies or determining whether a set of studies represents a population. However, procedures have been developed in an attempt to better determine if publication bias exists. Perhaps the easiest way is to place the individual studies in a scatter plot with effect size on the horizontal axis and the standard error (study size) on the vertical axis. Publication bias can be detected if the studies are distributed symmetrically around the average effect size. Also, a funnel is drawn consisting of straight lines defining the region within which 95% of the studies will lie when there is no publication bias. If more

studies are outside of the funnel or studies are asymmetrically placed around the mean, this may indicate publication bias and that a meta-analysis may not be appropriate.

When all studies are located the next step is to code them on the relevant variables. This can be done on methodological aspects (e.g. random assignment or not) and on characteristics of the intervention (e.g. does this study report effects of intervention A or intervention B). The goal of the meta-analysis is then to see if any of these variables are related to effects. The basic statistic in meta-analysis is the effect size (ES). The effect size transforms results into z scores that represent the difference between groups. The basic formula used for computing the ES is to subtract the average score of the control group from the average score of the intervention group and then divide by the standard deviation of the control group. There are however, a number of different formulae proposed, including formulae that can control for small sample sizes and also formulae to compute effect sizes where no control group is available. Typically the ES's of the individual studies are displayed in a forest plot, where the weight given to it, the confidence intervals and the p -values are reported. Often the weight will be represented by the size of the square (or other such symbol representing each study) on the plot; the larger the square the more weight the study carries. Study weight is normally assigned based on the studies precision and precision in turn is based on the inverse of the variance which in turn will be an indirect measure of sample size. Hence, precision and the corresponding confidence intervals will be more narrow the greater the weight of the study. There is also a correspondence between the p -value and the confidence interval. The p -value will fall under the .05 value only if the 95% confidence interval does not include the null value. In other words, we can see which of the studies that represent statistical significant effects by scanning the confidence intervals. At the bottom of the forest plot the ES from the individual studies are combined in a summary effect size. The summary effects size is usually represented by a diamond. The summary effect size is the weighted mean of the individual studies included in the analysis. There are two different ways the summary effect size is computed based on whether one assumes the studies share a common true effect size or not. If we assume they share a common effect size a fixed model is used, if we assume it varies, a random model is used.

As with the individual study effect sizes the summary effect size will also come with a confidence interval and a p-value. If the effect sizes vary substantially from study to study our attention will shift to explore this further, if the variability is modest this would have to be reported, and if we have a consistent effect size we may focus on the summary effect itself and use this to answer our original hypothesis.

While meta-analysis is now widely used and adopted the methodology is still somewhat controversial. Borenstein, et al., (2009) have recently listed some of the most common criticisms and I will go through them point for point

1) One number cannot accurately summarize the research in a field, since it only focuses on a combined effect ignoring the fact that intervention effects may vary from study to study. This criticism appears somewhat misguided as an analysis of heterogeneity of outcome should be part of any meta-analysis. If there is substantial dispersion of results, focus should be shifted to analyse this further rather than the combined effects, and modest dispersion should serve to place the mean effect in context. Thus, the main goal of a meta-analysis is to synthesize the effects and not simply to report a summary effect.

2) The file drawer problem invalidates meta-analysis. Several lines of evidence show those studies reporting good effects are more likely to get published; hence a meta-analysis will inherit this publication bias. This is a legitimate concern, however, it is not limited to meta-analysis, also narrative reviews will suffer from this, but in a narrative review this may be easier to ignore. Furthermore, methods have been developed to determine which meta-analyses should be considered suspect and which that should be considered robust to the impact of publication bias.

3) One risks mixing apples and oranges (i.e., different interventions) and the combined effect size may mask this problem. Studies will inevitably differ and it may be difficult to decide on how similar they need to be. Some meta-analysts can make questionable judgement in this regards whilst at the same time, critics may make unreasonable demands for similarity. However, in meta-analysis one always has the option of dividing the studies into subgroups and then comparing them empirically.

4) Garbage in makes for garbage out, referring to the notion that low-quality studies may be included in the analysis. Fundamental errors may then be transferred from individual studies to the meta-analysis where they are harder to detect. Meta-analysis will be especially vulnerable if all studies included are biased in the same way. However, through careful management of the studies included in the analysis one could take the opportunity to compare outcome of different level of quality studies. For instance outcome of studies with random assignment could be compared to those with quasi-random assignment to determine if effects sizes differ.

5) Important studies are often not included or ignored. In the selection procedure thousands of studies may be reviewed but only very few studies are eventually included in the analysis. Eligibility criteria should always be decided before the search is conducted but there will always be room for judgement in terms of which studies to include. For this reason the inclusion criteria need to be clearly stated and transparent and described as part of the report.

6) Meta-analysis may disagree with the results of single well-designed randomized trials. Some researchers have compared outcome of large scale randomized trials with the outcome of meta-analysis for the same intervention and found that these are in agreement only approximately 65 % of the time (LeLorier, Gregoire, Benhaddad, Lapiere, & Derderian, 1997). However, other research shows that also individual randomized trials are in agreement with each other only approximately 65 % of the time (Borenstein, et al., 2009). Often, reasons for this may be found in the patient population being studied and how the methods were applied. Hence, rather than being a problem for meta-analysis this should encourage researchers to look for finer differences in individuals and the intervention employed.

7) Meta-analysis is performed poorly and likely to contain any number of mistakes outlined above. The mistakes are flaws in application and not in the methodology and steps need to be taken by the research community to avoid them.

Individual Participant Data Meta-analysis

An individual participant data (IPD) meta-analysis provides some important benefits over a traditional aggregated data (AD) meta-analysis (Cooper & Patall, 2009). Among these is the possibility to perform subgroup analysis not conducted in the original studies, to add new information to the data sets and to apply different and more refined statistical methods. For these (and other) reasons Cooper and Patall (2009) conclude that when IPD and AD are both achievable, meta-analysis of IPD is superior to meta-analysis of AD. However, in general an IPD meta-analysis is often both costly and time consuming, and often data sets are simply not available. So at present it is suggested that the two analyses should be complementary with the first step always being to conduct an AD meta-analysis. While an IPD is stronger methodologically, it has its own potential pitfalls (Cooper & Patall, 2009).

- 1) The Ecological fallacy (Robinson, 1950). This pertains to the researcher comparing effects or relationships based on IPD that ignore group memberships. Within the current thesis this could be a problem if I tried to make assumptions for the individual child without knowing what sort of intervention they had received.
- 2) Simpson's Paradox refers to an situation in which the successes of groups seem reversed when the groups are combined (Simpson, 1951). One of the best real life examples of this occurred at the University of California, Berkeley which was sued for bias against women applying to graduate school (Bickel, Hammel, & O'Connell, 1975). Actually the University was slightly biased against men; however women tended to apply for the more competitive departments and thus had an overall lower admission rate. With respect to the present thesis I could get a similar paradox if I for example concluded that girls had significantly less benefit from IBI than boys since fewer of them achieved IQ's within the normal range following intervention. However, it may be that girls on average still gained more than boys in IQ points, but that girls had significantly lower IQ's at intake.

Chapter 3: Meta-analysis of Behavioural Intervention for Children with Autism⁴

A systematic literature search for studies reporting effects of Early Intensive Behavioural Intervention (EIBI) identified 34 studies, nine of which were controlled designs having either a comparison or a control group. We completed a meta-analysis yielding a standardized mean difference effect size for two available outcome measures: change in full-scale intelligence (IQ), and/or adaptive behaviour composite (ABC). Effect sizes were computed using Hedges' g . The average effect size was 1.10 for change in IQ (95 % confidence interval (CI) = .87, 1.34); and .66 (95 % CI .41, .90) for change in ABC. These effect sizes are generally considered to be large and moderate, respectively. Our results support the clinical implication that at present, and in the absence of other interventions with established efficacy, EIBI should be an intervention of choice for children with autism.

⁴ This chapter has been published as: 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.

Introduction

There is a developing evidence base for the positive effects of comprehensive interventions for children with autism spectrum disorders (ASD). Two recent narrative reviews have focused on a range of comprehensive interventions for children with autism (Eikeseth, 2009; Rogers & Vismara, 2008). The conclusion from both of these reviews is that Early Intensive Behavioural intervention (EIBI) is an effective intervention when compared against no intervention controls or eclectic/autism-specific special education interventions. When applying more formal criteria (Chambless, et al., 1998; Chambless & Hollon, 1998; Chambless, et al., 1996), Rogers and Vismara found that EIBI (or what they call the “Lovaas treatment approach”) should be considered “well established” and that no other intervention presently qualifies for this status.

EIBI programs (including the Lovaas treatment approach) have been described by Green, Brennan, and Fein (2002, p. 70, see also Eikeseth, 2009 for a similar definition) as having the following common elements: (a) intervention is individualized and comprehensive, addressing all skill domains; (b) many behaviour analytic procedures are used to build new repertoires and reduce interfering behaviour (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 behaviour analysis and experience with young children with autism directs the intervention; (d) normal developmental sequences guides the selection of intervention goals and short-term objectives; (e) parents serve as active co-therapists for their children; (f) intervention is delivered in one-to-one fashion initially, with gradual transitions to small-group and large-group formats when warranted; (g) intervention typically begins in the home and is carried over into other environments (e.g., community settings), with gradual, systematic transitions to preschool, kindergarten, and elementary school classrooms when children develop the skills required to learn in those settings; (h) programming is intensive, year round, and includes 20 to 30 hours of structured sessions per week plus informal instruction and practice throughout most of the children’s other waking hours; (i) in most cases,

the duration of intervention is 2 or more years; and (j) most children start intervention in the preschool years, when they are 3 to 4 years of age.

In addition to narrative reviews, there have been two recent systematic reviews of outcome research on EIBI. The first review presented a systematic description of the research published to date and pointed to challenges for future research (Howlin, Magiati, & Charman, 2009). Eleven studies were identified using the following inclusion criteria: the study had to have a control or comparison group with a minimum of 10 participants in each group, age at intake had to be less than 6 years, and intervention had to be provided for at least 12 hours a week for 12 months. Howlin et al. discussed a number of problems associated with drawing conclusions about the efficacy of EIBI. First, while not accurately reported in some of the studies, they estimated that the EIBI groups on average received significantly more hours of intervention than did control groups. Second, a variety of assessment instruments were used across children and studies that made it difficult to compare results across studies and may have led to results being spuriously positive. Third, in some studies it was unclear at what points in time the assessments were conducted, particularly at post treatment when in some cases assessments were undertaken years after intervention had ended. Fourth, the studies reported test scores in different ways that included standard scores, age equivalents, and raw scores.

Howlin et al. (2009) concluded that in general the average effects of EIBI were favourable compared to controls, but that the variability across individual children in the EIBI studies was substantial. Howlin et al. could not identify any reliable predictors of outcome. Intake IQ was found by some researchers to be related to better outcomes but others found no such relationship. Furthermore, age at intake was not found to be related to outcome in any of the studies. However, age range was limited with all children being under 7 years of age. Initial language ability was identified as a possible predictor only in some of the studies that explored this, and autism symptomatology was found to be related both with better and with worse outcomes, in two different studies. Given these problems they concluded that conducting a meta-analysis of the evidence was not appropriate.

The second recent systematic review conducted by Reichow and Wolery (2009) addressed similar questions to Howlin et al. (2009) and drew similar conclusions. However, unlike Howlin et al., they included a meta-analysis. The authors argued that a meta-analysis of EIBI is feasible, but that it had to be limited to change in intelligence scores, and that in order to have enough studies, they would have to include studies that were not controlled. Thus, the meta-analysis used standardized mean change effect sizes, and not the more methodologically rigorous standardized mean difference effect size. The mean change effect size is computed without comparison or control group data and, as the authors point out, any conclusions are limited by threats to validity such as maturation. In addition, the standardized mean change effect size may inflate effect size estimates (Morris, 2000). Based on 12 studies, Reichow and Wolery reported a weighted mean effect-size for change in intellectual functioning following EIBI of .69.

A second aspect of the Reichow and Wolery (2009) analysis that may affect the validity of conclusions was that studies using a variety of outcome measures for intelligence were included. For example, studies which primarily relied on performance based non-verbal measures of intelligence such as the Merrill-Palmer (Stutsman, 1948) and the Leiter-R (Roid & Miller, 1997) were treated as equivalent to studies which reported full scale IQ measures. Because the performance-based tests measure areas where children with autism often are relatively strong (e.g., visual-spatial tasks), scores tend to be higher than on full-scale IQ-tests (Lord et al., 2006). Not separating these tests in an analysis on the effects of EIBI may affect the conclusions drawn, especially in cases where these different measures are used interchangeably pre and post treatment. A minor methodological problem with the Reichow and Wolery (2009) analysis was that they reported reliability estimates only on the coding of information from the selected studies (research methods, participants, and intervention characteristics); no such data were provided for the initial procedure for selecting studies to be included in the review.

The purpose of the present study was to provide a replication and extension of the Reichow and Wolery (2009) meta-analysis, with a focus on methodological improvements. First, we selected studies with comparison/control groups only, whilst employing a more precise definition of EIBI (Green et al., 2002). This makes it

possible to apply more methodologically rigorous mean difference effect size measures. Second, we required more uniformity in outcome measurement and included only full-scale measures of intelligence. Third, we were able to add a meta-analysis of changes in adaptive behaviour. Fourth, because our literature search was conducted later in time and employed a somewhat different definition of EIBI, some additional recent published studies were included. Fifth, we included inter-rater reliability for our literature search and initial selection procedure for studies in the review. Sixth, we based our analysis on individual raw data gathered from authors rather than group average data reported in the original papers. This made it possible to prevent children from being represented more than once if they were included in more than one published outcome study and also to ensure the selection of evaluation time periods as similar as possible across studies.

Methods

Search Strategy and Data Collection

We searched the PsycINFO, Pubmed and ERIC databases (up to March 2008) using a combination of the following terms: behaviour analytic, behavioural, early, intervention, and autism and/or PDD. The first author read the headings and abstracts of all the papers from this initial search to decide whether the study warranted a more detailed coding. If it was possible that the study reported outcome data on the effects of behavioural treatment for children with ASD, the study was obtained for more detailed coding. In addition, the reference sections of obtained papers were browsed in an attempt to locate studies that might have been missed in the electronic search.

A coding scheme was used for coding all the selected studies (see appendix). First, it was coded whether or not the children had received behavioural intervention that generally adhered to common elements described by Green et al. (2002, p. 70). Second, a series of true/false scores were given for the following: (a) the participants were on average between two and seven years 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 behaviour such

as the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984) was conducted at intake and after intervention. Primarily administering a non-verbal intelligence measure such as the Leiter-R (Roid & Miller, 1997) or the Merrill-Palmer Scale of Mental Tests (Stutsman, 1948) led to the study's exclusion; (d) the duration of intervention was between 12 and 36 months; (e) the study was not a case study (or series of case studies) where participant selection may not be relevant, for instance only particularly positive effects may be published; (f) the results had been published in a peer-reviewed journal; and (g) the study included either a control or comparison group. The studies were classified as either a comparison or a control study (or both). If it was specified that the children in the study had received intervention(s) other than EIBI of similar duration and intensity in terms of 1:1 hours, it was classed as a comparison study. While it would probably be impossible to determine whether the children in the comparison groups had similar eclectic or specialist autism provision (even within a single study), classifying the studies in this way could still yield useful information. For example, it may facilitate the exploration of whether it is the number of 1:1 hours itself (i.e., "intensity") that makes a difference. Where no intervention (or a considerably less intensive one) was provided, or a poorly specified intervention was described, the study was classified as a control study.

The electronic and manual searches resulted in 2150 potential hits. Through the first screening process, we selected 34 papers for detailed coding. One of the database searches, resulting in 607 hits, was chosen for a reliability check. The screening results from the first author were compared to that of a second screener (another author) using the same decision criteria. Agreement was high overall in terms of whether the paper should be subject to further analysis (Cohen's Kappa = .85). Disagreements occurred only because the second screener included fewer studies than the first screener. Thus, there were no instances of the second screener including a study for further analysis that was not already included by the first screener. Thus, the sample included in this analysis was that identified by the first screener.

The 34 studies that remained after this initial screening were coded by the first author and two independent scorers (master students in behaviour analysis). 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 25 out of the 34 studies for one or more of following reasons: a) seven had inadequate intake and/or outcome data, such as primarily reporting performance IQ instead of full scale IQ (Bibby, Eikeseth, Martin, Mudford, & Reeves, 2002; Drew, et al., 2002; Fenske, et al., 1985; Luiselli, Cannon, Ellis, & Sisson, 2000; Magiati, Charman, & Howlin, 2007; Sheinkopf & Siegel, 1998; Solomon, Necheles, Ferch, & Bruckman, 2007); b) five had an intervention duration that was too short to meet inclusion criteria (Harris, Handleman, Gordon, Kristoff, & Fuentes, 1991; Ingersoll, Schreibman, & Stahmer, 2001; Reed, Osborne, & Corness, 2007b; Reed, Osborne, & Corness, 2007a; Stahmer & Ingersoll, 2004); c) two reported data from case studies only (Butter, Mulick, & Metz, 2006; Green et al., 2002); d) three reported data that were already included in other studies (Beglinger & Smith, 2005; Eikeseth, et al., 2007; McEachin, Smith, & Lovaas, 1993); e) upon closer inspection, one of the studies provided intervention that did not meet the definition of behavioural treatment (Gabriels, Hill, Pierce, Rogers, & Wehner, 2001); and finally, f) seven did not have a control or comparison group (Anderson, Avery, DiPietro, Edwards, & Christian, 1987; Ben-Itzhak & Zachor, 2007; Harris & Handleman, 2000; Hayward, et al., 2009; Sallows & Graupner, 2005; Smith, Buch, & Gamby, 2000; Weiss, 1999). It should be noted that several of the studies may have violated more than one of these criteria. For instance the Magiati et al., (2007), relied primarily on performance measures of intelligence and did not meet the definition of behavioural treatment since supervision was not provided frequently enough or was supervised by professionals that were not trained (or experienced) according to the Green et al. (2002) criteria.

Individual data needed to calculate effect sizes from the nine remaining studies were obtained by contacting the authors of each study. We asked them to provide the age, IQ, and adaptive behaviour scores at intake and after two years in intervention (or as close as possible to). Also, we asked if any of the children either in the EIBI or comparison/control groups was represented in other published studies. Thus, all computations in the present study were conducted by re-calculating pre and post-

group means and standard deviation on outcome measures rather than data reported in the original papers or extrapolated from these reports. Individual data from control group 2 (n=21) in the Lovaas (1987) study were not available, and four children in the comparison group from one study (Eldevik, et al., 2006) were excluded because they were already in the comparison group of an earlier study also included in the analysis (Eikeseth, et al., 2002). Figure 3.1 summarizes the study search and selection process.

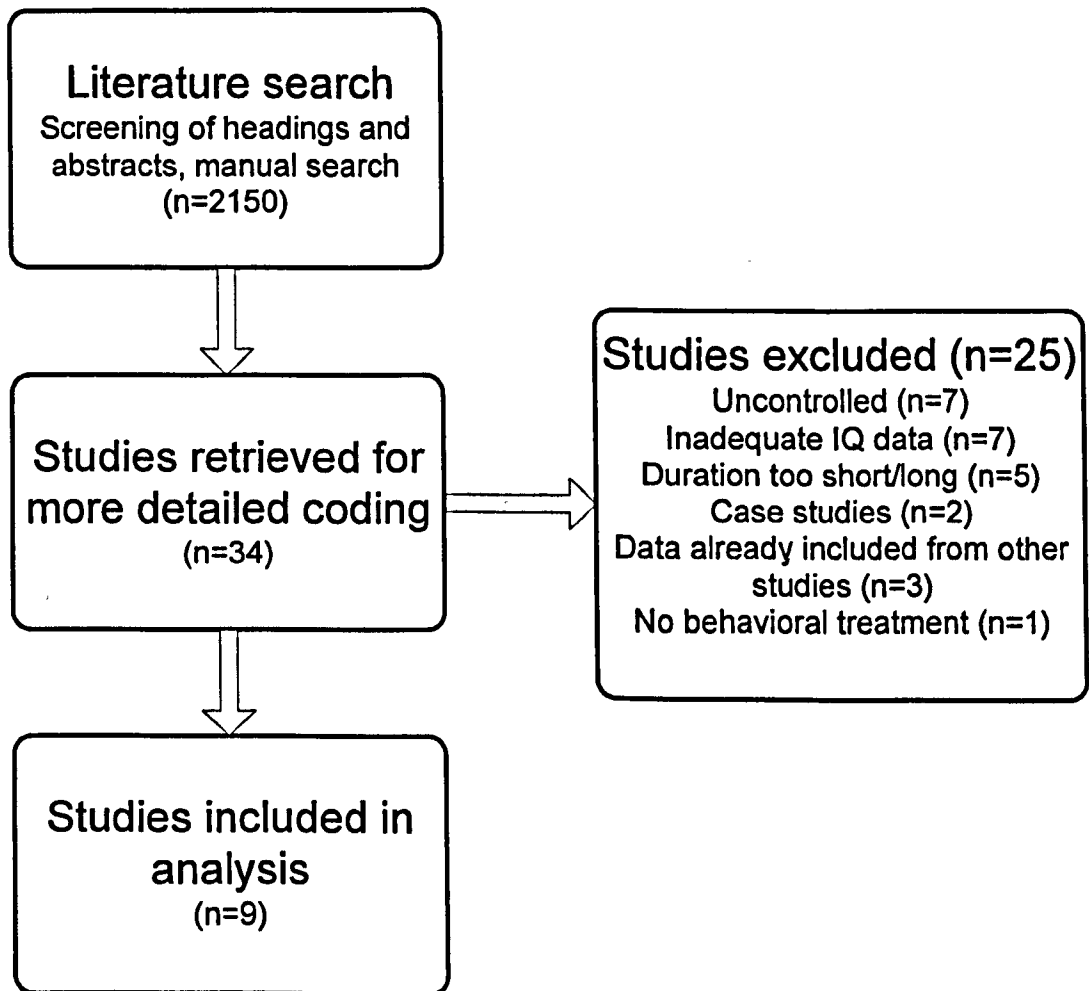


Figure 3.1. The search and selection procedure.

The total number of children in the nine intervention studies was 297; 153 in the EIBI groups, 105 in control groups and 39 in comparison groups. Table 3.1 summarizes the main characteristics of the children included in this analysis including: mean age at intake, IQ, and Vineland Adaptive Behavior Composite (ABC) scores at intake and post treatment. Not all authors reported both IQ and ABC data, or were able to give

the exact duration of intervention for each individual child. The average intensity in terms of weekly hours and duration is also provided in Table 3.1. The research design and assignment procedures employed are briefly described along with any inclusion criteria described in the original paper. If a study reported outcome data at more than one point in time, we chose the point that was closest to two year duration of treatment. All of the above calculations were conducted in SPSS (version 16.0) using raw-data provided from the authors. Hence, the pre and post group means and standard deviations may differ from those reported in the original published papers.

Child Measures

Intellectual functioning. The Bayley Scales of Infant Development (BSID), either the first or second edition (Bayley, 1969, 1993) were used for the youngest children or the children that scored below the basal on intelligence tests standardized for their chronological age. The BSID is a measure of mental development for children up to 42 months. It will yield a mental developmental index (MDI), which is considered broadly equivalent to an IQ score. 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). If the child scored below the norms on a test, researchers had generally computed a ratio IQ score by dividing the obtained mental age with chronological age and multiplying by 100. All of the tests have been used extensively and validated for children with pervasive developmental disorders and intellectual disabilities (Newsom & Hovanitz, 1997).

Adaptive behaviour. The Vineland Adaptive Behavior Scales (VABS), (Sparrow, et al., 1984) was the only measure for adaptive behaviour used in the studies included in the present analysis. The VABS yields standard scores on four domains; communication, daily living skills, socialization, and for children less than six years old, motor skills. Based on these scores it will also yield a standardized adaptive behaviour composite (ABC). In the present study we only used this composite score

Table 3 1:
Main characteristics of the studies included in the analysis of outcome. Studies included in other recent reviews are shown.

Study	Group	Age	Pre		Post		Intensity		n	Design/Assignment/Inclusion	Comments
			IQ	ABC	IQ	ABC	hrs	mins			
USA (Lovaas, 1987) 1, 2, 3, 4	EIBI Control	34.6 (8.9) 40.9 (10.3)	62.9 (13.7) 57.1 (14.5)		83.3 (28.6) 50.1 (22.4)		40 <10	24-36 24	19 19	QCT/Staff availability and archives included if CA < 40 months if mute or CA < 46 months if echolalic and prorated mental age of >11 months at CA 30 months	Five subjects deemed untestable at intake, 3 in experimental group and 2 in control group. 1. Intelligence scores based on mental age scores from the Vineland Social Maturity Scale (Doll, 1953) were used in these cases.
(Smith, Eikeseth, Klevstrand, & Lovaas, 1997) 1, 3, 4	EIBI Control	36.0 (6.9) 38.0 (5.4)	27.8 (4.9) 27.3 (5.4)	50.3 (9.1)	35.8 (14.3) 24.0 (8.2)	51.7 (17.9)	30 <10	24 24	11 10	QCT/Archival data included if CA ≤ 46 months and IQ < 35.	Post measures conducted 3-4 years after treatment in some cases. VABS data only available for 6 of the 11 children in the ABA group. Control group received minimal treatment.
(Howard, Sparkman, Cohen, Green, & Stanslaw, 2005) 2, 3	EIBI Comparison (AP) Control (GP)	30.9 (5.2) 37.4 (5.7) 34.6 (6.5)	58.5 (18.2) 53.7 (13.5) 59.9 (14.8)	70.5 (11.9) 69.8 (10.5) 71.6 (10.5)	89.9 (20.9) 62.1 (19.6) 68.8 (15.3)	81.3 (11.1) 69.1 (12.9) 68.3 (9.9)	25-40 25 15	14 13 15	29 16 16	QCT/Parental preference and IEP teams included if CA < 48 months.	EIBI: Multiple settings (home, school and community) 25-30 h per week under 3 years of age, 35-40 h per week over 3 years of age. AP (autism educational programming): Public classroom for children with autism. 1:1 or 1:2 staff: child ratio. 25-30 hrs 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. GP (generic educational programming): Local community special education classrooms. Average of 15 hrs per week intervention, 1:6 staff: child ratio. 13 children received speech and language therapy for 1-2 times per week.
(Smith, Groen, & Wynn, 2000) 1, 2, 3, 4	EIBI Companson	36.1 (6.0) 35.7 (5.4)	50.5 (11.2) 50.7 (13.9)	63.7 (9.6) 65.2 (9.0)	66.5 (24.1) 50.5 (20.4)	61.3 (28.7) 59.9 (16.7)	24.5 <10	24 -24	15 13	RCT/Matched-pair random. Included if CA < 42 months and ratio IQ between 35 and 75.	Clinic directed group: Number of hours for ABA group are for first year in treatment. Gradual reductions in year two. Treatment phased out after 18 months for children responding slowly. Average duration 33 months. Parent managed group. 5 hrs a week of parent training for first 3-9 months, parents asked to do 5 hours a week in between sessions: Total < 10 hours per week of ABA + 12.5 hours of special education classes per week. ABA treatment hours second year presumed to be gradually decreasing, school hours 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.
(Cohen, Amerine-Dickens, & Smith, 2006) 1, 2, 3, 4	EIBI Control	34.4 (5.4) 33.2 (3.7)	62.0 (16.4) 59.4 (14.7)	64.0 (8.4) 71.9 (11.5)	81.1 (21.8) 65.9 (16.5)	79.5 (13.4) 70.7 (13.3)	35-40 -	24 24	21 21	QCT/Parental preference. Included if CA < 48 months and ratio IQ > 35.	Community-non-university setting. Community services selected by family. In control group 1 child had a Early Start Autism Intervention Program 9 hrs a week, 2 children home-based development program 1-4 hrs a week, 17 SDC (special day class) eclectic, ratio 1:1 to 3:1, 3-5 days a week for up to 5 hrs. Speech, behavioral and occupational therapies 0-5 hrs per week. 3 where mainstreamed for up to 45 minutes a day.
Australia (Bimbrauer & Leach, 1993) 1, 2, 4	EIBI Control	38.1 (7.1) 33.2 (10.3)	45.3 (17.9) 45.0 (9.4)	47.5 51.5	57.6 (18.7) 43.2 (15.0)	41.0 (5.1) 42.5 (4.9)	19 -	22 22	9 5	QCT/Parent willingness and geographical included if CA between 24 and 48 months.	Untestable subjects set to IQ of 30. Ratio scores computed for the rest of subjects. Scores post treatment are deviation IQ used where available. Ratio VABS scores calculated both pre and post.
Norway (Eikeseth, Smith, Jahr, & Eldevik, 2002) 1, 2, 3, 4	EIBI Companson	66.3 (11.3) 64.8 (9.9)	61.9 (11.3) 65.2 (15.0)	55.8 (9.0) 60.0 (13.2)	79.1 (18.1) 68.9 (18.8)	67.0 (16.3) 60.2 (11.7)	28 29	12 14	13 12	QCT/Staff availability included if CA between 48 and 84 months and IQ ≥ 50.	Companson received eclectic treatment of similar intensity, but with less supervision.
(Eldevik, Eikeseth, Jahr, & Smith, 2006) 1, 2, 3, 4	EIBI Companson	53.1 (9.5) 45.1 (16.5)	41.0 (15.2) 42.8 (13.0)	52.5 (3.9) 50.1 (9.2)	49.2 (16.6) 38.5 (15.5)	52.4 (9.2) 44.6 (7.5)	13 12	20 23	13 11	QCT/Archival data included if CA < 72 months.	Companson received eclectic treatment of similar intensity. Four subjects from companson group taken out here, because included in Eikeseth et al. (2002).
United Kingdom (Remington et al., 2007) 2, 3	EIBI Control	35.7 (4.0) 38.4 (4.4)	61.4 (16.7) 62.3 (16.6)	60.2 (5.8) 57.0 (6.8)	73.5 (27.3) 60.1 (27.8)	61.5 (15.4) 54.6 (13.1)	26 16	24 24	23 21	QCT/Parent preference included if CA between 30 and 42 months.	Control group received TAU, special school, mainstream or mix, but little or no 1:1, speech therapy, TEACCH etc.

QCT=Quasi-experimental controlled clinical trial

RCT=Randomized controlled clinical trial

1, Included in Reichow & Wolery (2009), 2, Included in Eikeseth (2009), 3, Included in Howlin, Magiati & Charman (2009), 4, Included in Rogers & Vismara (2008)

as we did not have access to the domain scores for most of the children. The VABS is widely regarded as the best available instrument for assessing adaptive behaviour in children with autism (Newsom & Hovanitz, 1997).

Tests of Homogeneity and Publication Bias

Data were entered into the Comprehensive Meta-analysis Software (Borenstein, Hedges, Higgins, & Rothstein, 2005). To determine whether all studies were drawn from a population of studies with a common mean effect size, we performed a test of homogeneity using the Q -statistic and I^2 , utilizing these options in the Comprehensive Meta Analysis Software. These tests were conducted for the whole group of nine studies together. The I^2 gives the proportion of the variance that be explained by between-study variance. Using the software, we also assessed potential publication bias by a funnel plot of the standard error and effect size for each study (Egger, Smith, Schneider, & Minder, 1997) and the trim and fill method (Duval & Tweedie, 2000) for both IQ and ABC outcomes.

Effect Size Measures

The standardized mean difference effects size for EIBI were computed for IQ and ABC using the same software. Effect sizes were computed for each study separately but we also computed an overall effect size against the comparison and control groups. We used the Hedges' g effect size measure (Hedges & Olkin, 1985) to adjust for the relatively small sample sizes in the studies, typically <20 in each group. When computing an overall (meta-analytic) effect size the individual studies were weighted using the inverse of the variance, as is widely considered to be the best practice (Borenstein, Hedges, Higgins & Rothstein, 2009).

Results

Homogeneity

The Q -statistic was not statistically significant for either IQ ($Q(9) = 10.07, p = .345$), or ABC ($Q(7) = 8.50, p = .291$) scores across the nine identified studies. This

indicated that all of the studies could be combined for one common effect size. We also calculated the between-study variance for IQ ($I^2= 10.66$), and for ABC scores ($I^2= 17.65$), and these data supported the homogeneity conclusion in that relatively small proportions of variance were explained by between-study variance. Given these findings, we used a fixed effects model for computing all effect sizes.

Effect Size Measures

The standardized mean difference effect size was calculated for IQ and ABC. The Howard, et al., (2005) study contributed both a control and a comparison group. We decided to calculate separate effect sizes for these. Hence, the total number of effect sizes for IQ was ten from the nine studies included. Four studies had a comparison group and six studies had a control group. For ABC, four studies had a comparison group and four studies had a control group, the Howard, et al. 2005 study again contributing one to each group, making the total number of effect sizes eight from the seven studies included. A forest plot of the effect sizes for each study and an overall effect size for IQ and ABC are shown in Figure 3.2. The overall effect size for IQ change was 1.103 (95 % confidence interval (CI) .871, 1.335). The overall effect size for change in adaptive behaviour composite scores was .660 (95 % CI .41, .90).

Publication Bias

We found no statistical or visual evidence of publication bias. Funnel plots of the standard error against effect sizes for IQ and ABC changes are shown in 3.3. The Duval and Tweedie (2000) trim and fill method did not suggest the potential absence of any studies. However, the limitations of these techniques, particularly when there are few studies, mean we cannot exclude publication bias.

Discussion

Following EIBI treatment, our meta-analysis found an average large effect size for IQ change (based on 10 comparisons) and an average medium effect size for ABC change (based on eight comparisons) (Cohen, 1992). These estimates of effect size were also statistically significant from zero (the 95 % confidence intervals did not

contain zero). We also found that the sample of studies was reasonably homogenous and that there was no evidence of publication bias.

The only other published meta-analysis we have found (Reichow & Wolery, 2009) reported an effect size of .69 for IQ change. The effect size for IQ change is thus somewhat higher in the present analysis. There may be several explanations for the difference between our results and those of Reichow and Wolery that relate to the steps we took to expand and improve on their analysis. First, to be able to use the more methodologically rigorous standardized mean difference effect size, we only included studies that had a control or a comparison group. In contrast, Reichow and Wolery computed their mean effect size based on the change within the EIBI group only. Second, we applied a more precise yet inclusive definition of EIBI, introduced by Green et al, (2002), that seems to us to be more in keeping with how other EIBI professionals define their field (e.g., Eikeseth, 2009). Reichow and Wolery employed more restricted criteria for including studies in their review, in that it had to be a replication of Lovaas' UCLA/YAP model and/or based on their treatment manuals.

This may be the reason why the Howard et al. (2005) and the Remington et al. (2007) studies were not included in their analysis. However, in our opinion, the failure to include these two studies represents an inconsistency in the selection process. To us, the treatment provided in these studies is as much EIBI as the treatment provided in the studies that were included by Reichow and Wolery. Along the same lines, it seems inappropriate to us to include in the calculation of effect size a study that compared two models of EIBI service delivery (Sallows & Graupner, 2005). It is quite clear from the description of the provisions in this study, that while the groups differ in some respects, they are both examples of EIBI. A third difference in our analysis is that we only included studies that reported full-scale intelligence scores. Reichow and Wolery did not make a distinction between performance based and full-scale intelligence measures. As we have noted earlier, this may skew results in either direction, especially when tests are used interchangeably over time. Fourth, our analysis was based on individual raw data from each study rather than the data reported in the published papers. This meant that we had a slightly different sample of children, even from the studies in common to both analyses (see Table 3.1).

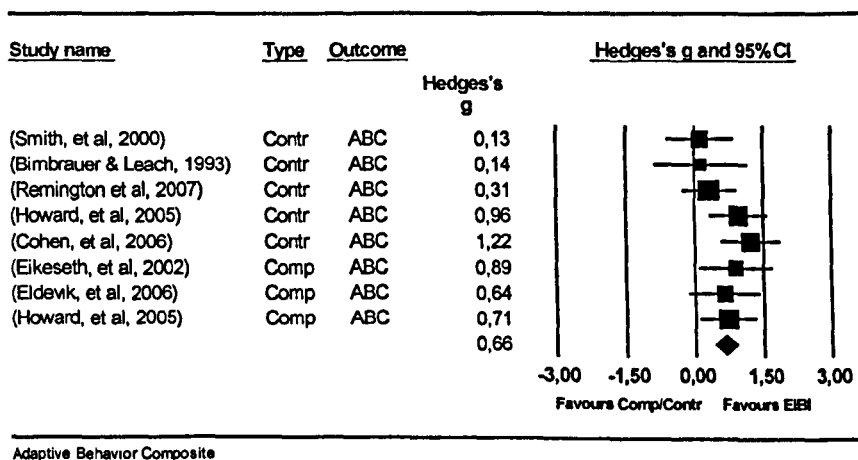
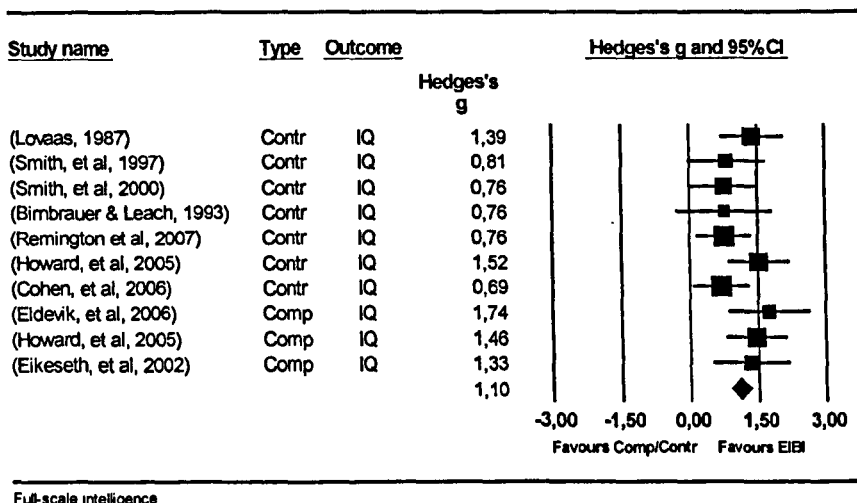


Figure 3.2. Forrest plots of standardized mean difference effect sizes (Hedges' g) and 95 % confidence intervals. In the first plot effect sizes for full scale IQ are shown and in the second ditto for adaptive behaviour composites. Studies are grouped as either comparison or control. The fixed model effect size is computed against both the comparison and control studies and also an overall effect size is computed.

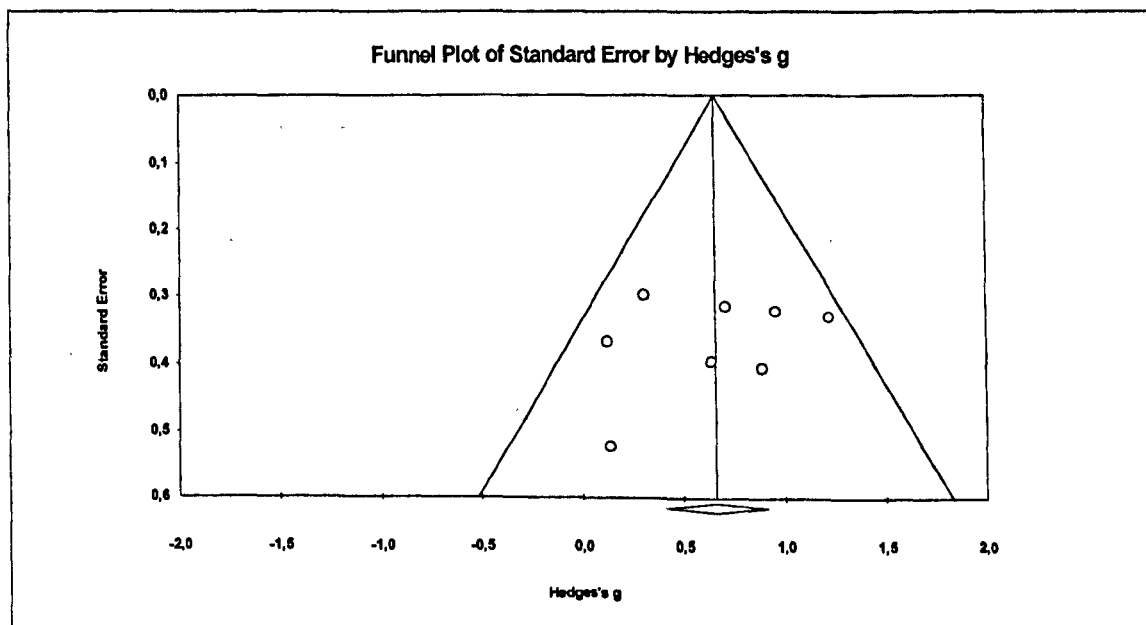
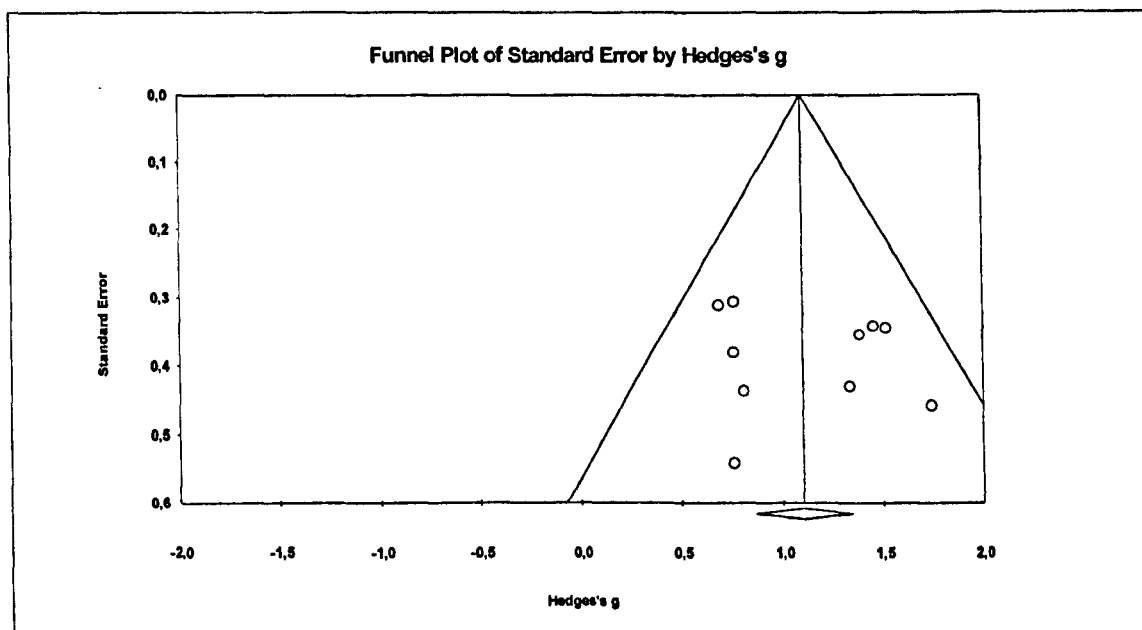


Figure 3.3. Funnel plots of effect sizes against the standard errors. In the first plot the full scale IQ data are shown, and in the second plot the adaptive behaviour composites are shown. The circles represent the studies included in the analysis, and the diamond represents the average effect size with a 95 % confidence interval.

As an extension to Reichow and Worley's meta-analysis, we were able to include an analysis of another important outcome measure, namely the adaptive behaviour composite. This measure adds substantial validity to the outcomes, because it tells us more about the children's skills in daily life. Interestingly, effects sizes were lower than for IQ. We also tested if intensity of treatment in itself may account for differences in outcome. This was possible by employing stricter criteria for what should constitute a comparison group. In the present study, the comparison groups had to be given a provision of similar intensity (measured as weekly hours of 1:1 provision) as the EIBI groups. In the studies included here, this meant an "eclectic" provision. While we agree with Reichow and Worley's (2009) point that it is still hard to determine whether this means a specific common provision, we think it is valuable to treat them as a group, especially since eclectic provision is probably similar to a treatment as usual for many children with autism. In eclectic programs, the particular composition of treatments is to be adjusted to the individual child's needs and may thus vary a great deal across children and across time for a given child. Attempts to measure this have been made (e.g., Eikeseth et al., 2002) but it proved difficult for teachers in the eclectic groups (and thus for the researchers) to say what specific treatment they were using because they tended to blend and apply them depending on the child's behaviour and needs through the day. Although difficult to specify, the eclectic approach seems to be the most common provision offered to children with ASD in service settings currently, even amongst those clinicians with behaviour analysis training (Schreck & Mazur, 2008). Our results add to the serious concerns raised by Rogers and Vismara (2008) about eclectic treatment models.

While we were able to refine Reichow and Worley's meta-analysis, there are some serious limitations that remain, such that any conclusions need to be drawn with caution and to be considered tentative. First, the number of studies included in our analysis may be considered small, although it is above the median for reviews listed in The Cochrane Database of Systematic Reviews. This database currently includes over 3000 reviews and the median number of studies in a review is six (Borenstein, Hedges, Higgins, & Rothstein, 2009). Second, a more serious limitation is the quality of the studies on effects of EIBI. Because of the lack of random assignment, only one study included in the present analysis met Type 1/highest level criteria of methodological rigor (Nathan & Gorman, 2002). Furthermore, the literature lacks

comparisons between EIBI and other approaches, perhaps other than the eclectic one. Third, while there is a clear difference in outcome between EIBI and the comparison intervention, it should be noted that this may be due to differences in the amount and frequency of supervision and training. We did not have enough data to control for this in the present study. However, based on the information in the studies included, it is clear that the EIBI group in general received more frequent and more total hours of supervision and training. This remains a threat for the validity of conclusions about the superiority of EIBI in relation to comparison intervention. Fourth, we decided to include two effect sizes from the Howard et al. (2005) study, one for EIBI against the comparison group and one for EIBI against the control group. This is problematic since they are not independent of each other as both involve contrasts with a single EIBI group. We did all calculation only including the effect size from the comparison group and this did not alter the overall results in any significant manner; Hedges' $g = 1.048$ (95 % CI .80, 1.30) for IQ, and Hedges' $g = .607$ (95 % CI .34 to .87) for ABC. Fifth, due to the limited number of studies and available variables we decided not to conduct an analysis of moderator variables that may explain variation in intervention outcome. However, this is certainly a priority for the future when more studies are published and more potential moderator variables can be analyzed.

Implications for Research, Policy, and Practice

With these limitations in mind, our general conclusions are very similar to those of other recent reviews: EIBI produces large to moderate effect sizes for changes in IQ and ABC scores for children with ASD when compared with no intervention controls and eclectic provision. These results support the clinical implication that EIBI at present should be an intervention of choice for children with ASD. However, randomized controlled trials comparing EIBI to other interventions are still needed. In particular, studies are needed where the comparison intervention is of similar intensity and where staff receives similar training and supervision.

Chapter 4: Using Individual Participant Data to Extend the Evidence Base for Intensive Behavioural Intervention for Children with Autism⁵

We gathered individual participant data from 16 group design studies of behavioural intervention for children with autism. 309 children received behavioural intervention, 39 received comparison interventions, and 105 were in a control group. More children receiving behavioural intervention achieved reliable change in IQ, 29.8 % vs. 2.6 % and 8.7 % for comparison and control groups, and adaptive behaviour, 20.6 % vs. 5.7 % and 5.1 %. This equated to a Number Needed to Treat, combining control and comparison interventions, of 5 for IQ and 7 for adaptive behaviour, and Absolute Risk Reduction of 23 % and 16 % respectively. Within the behavioural intervention sample, IQ and adaptive behaviour at intake predicted gains in adaptive behaviour. Intensity of intervention predicted gains in both IQ and adaptive behaviour.

⁵ This chapter is based on a paper that is accepted for publication: Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (in press). Using individual participant data to extend the evidence base for Intensive Behavioral Intervention for children with autism. *American Journal on Intellectual and Developmental Disabilities*.

Introduction

There is a growing body of evidence that Intensive Behavioural Intervention (IBI) can result in significant improvement in the intellectual, social, adaptive and language functioning of young children with autism spectrum disorders (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). 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, et al., 1993). Such data have led to positive conclusions about the evidence base for IBI in two recent narrative reviews (Eikeseth, 2009; Rogers & Vismara, 2008). According to Rogers and Vismara (2008), clinic based IBI (or what they call the Lovaas treatment approach) can be considered “well established” based on formal criteria (Chambless et al., 1996; Chambless et al., 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 different 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 post-tests scores in the experimental and control groups. Reichow and Wolery (2009) have recently conducted a synthesis of the research, including an aggregated data meta-analysis on the effects of IBI 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 following IBI of .69. Such an effect size would normally be considered clinically meaningful. In a second aggregated data meta-analysis of nine controlled studies of IBI, using a weighted mean difference effect size, Eldevik et al. (2009) found a large effect for IQ change in favour of IBI (Hedges' $g = 1.10$ [95 % CI .87, 1.34]) and a smaller although still

statistically significant effect for change in adaptive behaviour composite 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 then 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 within IBI studies varies considerably (Howlin, Magiati & Charman, 2009); an important step in the examination of the evidence base for IBI 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 (Jacobson & Truax, 1991); a construct borrowed from psychotherapy outcome research, to examine meaningful change in their IBI 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. Using Jacobson and Truax's formula, Remington et al. (2007) found that 6 out of 23 children (26 %) in their IBI group achieved positive reliable change on IQ after two years, whereas 3 out of 21 (14 %) in the treatment as usual group achieved this level of change and the IQ scores of three children in this group also reduced to a reliable extent. To date, no other published IBI 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 IBI 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 (NNT) and Absolute Risk Reduction (ARR) (Straus & Sackett, 2005). These statistics are particularly helpful as simple ways to communicate information about interventions to policy makers. The NNT 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, $NNT = 4$ 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 $NNT = 1$ 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 NNT, the less effective the treatment relative to the comparison (Kraemer, et al., 2003). When the confidence interval for NNT extends into negative numbers it means that the intervention may also be harmful. In such cases we compute the Number Needed to Harm (NNH), which indicates how many patients that needs to be exposed to the intervention to cause harm in one patient that would not otherwise have been harmed. The lower the number needed to harm, the greater the risk of being harmed by the intervention.

ARR is computed in a similar way as NNT but expressed as a measure of the difference in percentage response between two interventions (Pinson & Gray, 2003). When the ARR 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 treatment A: 50 % of patients do not respond to intervention, and in treatment B: 90 % do not respond to intervention, the ARR (also called risk difference) is 40 % in favour of intervention A.

A further advantage of establishing an objective criterion for meaningful outcome for individual children with autism receiving IBI is that the search for correlates or predictors of IBI outcome can become more consistent. For example, the children who achieved reliable change following IBI (n=6) in the Remington et al. (2007) study, were compared to the children in the IBI group whose IQ score decreased (n=6). The children that met reliable change criteria had higher IQ, mental age, Vineland composite scores (Sparrow, et al., 1984), along with higher Vineland communication and socialization scores at intake. In addition, these best outcome children at intake had lower Vineland motor scores, more behaviour problems on the Developmental Behavior Checklist (DBC) (Einfeld & Tonge, 1995) and more autistic symptoms on the DBC autism algorithm (Einfeld & Tonge, 2002), but also fewer treatment hours in their second year of IBI.

Apart from the Lovaas (1987) intensity comparison (40 vs. ≤ 10 hours), IBI 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 (Harris & Handleman, 2000), initial social skills (Ben-Itzhak & Zachor, 2007), toy play and socially avoidant behaviour at intake (Sherer & Schreibman, 2005), and autism subtype (Beglinger & Smith, 2005). Notably, despite its potential significance to the IBI 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 IBI there is currently no evidence base that can be used to identify children at intake that 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, from published IBI outcome studies identified via a systematic review. We then used all of these data to establish whether each individual child met reliable change criteria for changes in IQ or adaptive behaviour after approximately 2 years of intervention. These data were then used to address two aims. First, we carried out an individual participant data meta-analysis of IBI 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 IBI. This analysis was able to extend beyond the small n analyses from individual published studies and also to facilitate a more sophisticated analysis of outcome prediction in one important respect. We were able to explore both main effects and also 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.

Methods

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: behaviour analytic, behavioural, early, intervention, and autism and/or PDD. The first author read the headings and abstracts of all the papers from this initial search; studies that reported standardized outcome data on the effects of behavioural intervention for young children with autism were obtained for more detailed coding. The first author manually browsed the reference section of each in an attempt to locate studies that might have been missed during the electronic search.

Following this selection process we developed a coding scheme (see appendix) and coded the selected studies in two main ways. First, we coded whether or not the children had received behavioural 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 behaviour analytic procedures were used to build new repertoires and reduce interfering behaviour (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 ABA and experience with young children with autism directed intervention; (d) typical developmental sequences guided selection of intervention goals and short-term objectives; (e) parents served as active co-therapists for their children; (f) intervention was delivered in one-to-one fashion initially, with gradual transitions to small-group 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 hours 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 two 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 two and seven years 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 behaviour such as the Vineland Adaptive Behavior Scales (Sparrow, et al., 1984) was conducted at intake and after intervention—studies that had primarily administered a non-verbal intelligence measure such as the Leiter-R (Roid & Miller, 1997) or the Merrill-Palmer Scale of Mental Tests (Stutsman, 1948) were excluded 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) where only particularly positive effects may be published; 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 IBI 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". While 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 IBI 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 2150 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 that 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 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. Thus, the sample included in the present analysis was that identified by the first screener.

The remaining 33 studies were then coded by the first author and two independent scorers (master students in behaviour 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 following reasons: a) seven had inadequate

intake and/or outcome data, most often reporting primarily performance IQ instead of full scale IQ (Bibby, et al., 2002; Drew, et al., 2002; Fenske, et al., 1985; Luiselli, et al., 2000; Magiati, et al., 2007; Sheinkopf & Siegel, 1998; Solomon, et al., 2007); b) in five of the studies the duration of intervention was too short to meet inclusion criteria (Harris, et al., 1991; Ingersoll, et al., 2001; Reed, et al., 2007a; Reed, et al., 2007b; Stahmer & Ingersoll, 2004); c) two papers reported data from case studies only (Butter, et al., 2006; Green, Brennan, & Fein, 2002); d) three of the studies reported data that were already included in other studies (Beglinger & Smith, 2005; Eikeseth, et al., 2007; McEachin, et al., 1993); and finally, e) upon closer inspection one of the studies provided intervention that did not meet the definition of behavioural intervention (Gabriels, et al., 2001).

Of the 15 remaining studies, only four reported 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 four children in the comparison group of one study (Eldevik, et al., 2006) were extracted since they were also in the comparison group of another study included in the analysis (Eikeseth, et al., 2002). One of the authors who we contacted also volunteered an additional study that was in press (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 4.1 presents a flowchart of the search and selection procedure.

Table 4.1 summarizes the main characteristics of the studies included in this analysis, the mean age at intake, and the mean IQ and adaptive behaviour scores at intake and post intervention. Furthermore, the mean intensity 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 a study reported outcome data at more than one point in time, we chose the point that was closest to two year duration of intervention.

Participants

The total number of participants for whom individual data were provided was 453, including 309 who had received IBI, 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 behaviour data (2 studies) are missing (see below). A one-way analysis of variance showed that all three groups were similar on intelligence measures at intake. Children in the comparison intervention group were older than children in the other two groups at intake, and children in the control group had higher Vineland Adaptive Behavior Composite scores (see Table 4.2). However, the total sample, as well as the separate sub groups with the sample seem to be fairly representative of the autism population (Volkmar & Klin, 2005). Not all authors were able to provide data on the gender of each individual child (38.5 % missing), thus these data were not included in the present analysis. Within the IBI group, the number of weekly intervention hours for each individual child was only available for 75 out of 309 children or 24.3 %. To include intensity as a variable, a decision was made 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 individual child if these data were available. Second, we used the mean weekly hours of intervention for the IBI group that the child was in. Data on the group means for the IBI 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 hours 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 that scored below the basal on other intelligence tests. The BSID is a measure of mental developmental level for children up to 42 months. It will yield a

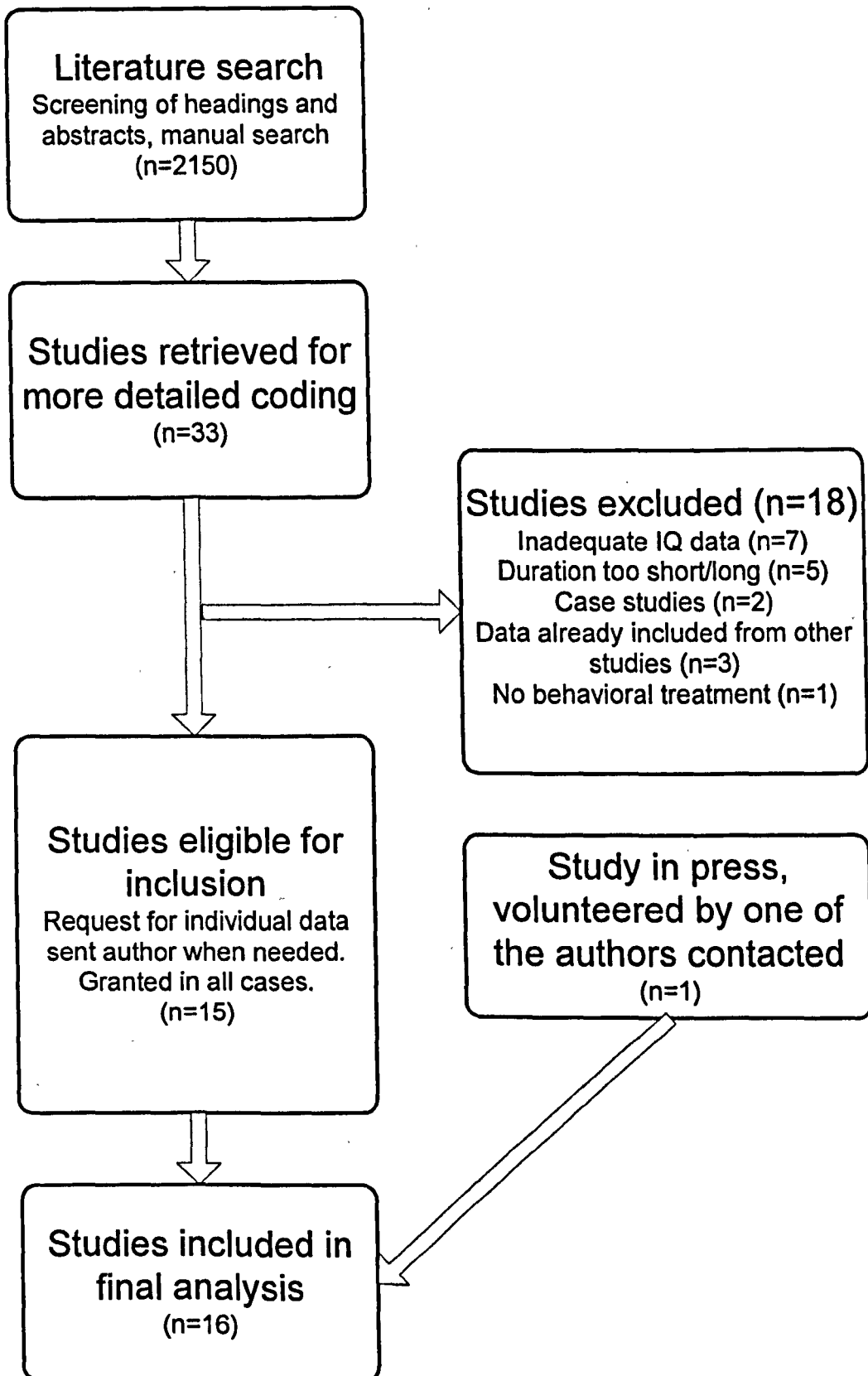


Figure 4.1. Flowchart on the procedure for selecting studies.

mental developmental index, which was considered broadly equivalent to an IQ score. For the older and higher functioning children the most frequently used measures of intelligence were the Stanford-Binet Intelligence Scale: Fourth Edition (Thorndike, et al., 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 score by dividing the obtained mental age with chronological age and multiplying by 100. Unfortunately, we do not have data about which tests were used for each child at what point, nor for whom a ratio IQ score was used. IQ outcome data were obtained from a total of 422 children (31 missing). These were divided as follows: 279 children in the IBI groups (30 missing), 104 children in the control groups (1 missing), and 39 children in the comparison groups (0 missing).

Adaptive behaviour. The Vineland Adaptive Behavior Scales (VABS) (Sparrow, et al., 1984), was the measure for adaptive skills in all studies included in this research. The VABS gives standard scores for communication, daily living skills, socialization, and for children under six years old, motor skills. It also yields a total adaptive behaviour composite (ABC). In the present study we only used the ABC scores as 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 behaviours were obtained from a total of 357 children (96 missing): 248 children in the IBI 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 behavioural intervention at the level of individual children we applied the statistical approach outlined by Jacobson & Truax (1991). The formula for computing reliable change requires that one can 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 two years by finding the correlation between pre and post scores 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 calculated the standard deviation for test scores from the whole sample of 453 children using intake data. Using the formula reported in 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).

Some IBI studies excluded children with intake IQ scores 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 IQ scores 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 purpose 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 Vineland mainly reflected a smaller standard deviation in the test scores at intake. None of the analyses reported here revealed a different pattern of results when the children with intake IQ scores below 35 were excluded, thus, no further results excluding those children are reported.

After classifying each individual child in terms of whether or not their intellectual functioning and adaptive levels changed to a reliable extent, we computed NNT and ARR (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 NNT and ARR calculations we used readily available free access online calculators (Straus, Newton, & Tomlinson, 2004).

Table 4.1:

Main characteristics of the studies included in the present analysis.

Study	Group	Age	Pre		Post		Intensity		n		Gender		Design/Assignment/Inclusion	Comments
			IQ	ABC	IQ	ABC	hrs	mths	Boys	Girls				
USA														
(Lovaas, 1987)	EIBI (clinic)	34.6 (8.9)	62.9 (13.7)		83.3 (28.6)		40	24-36	19	16	3		QCT/Staff availability and archives. Included if CA < 40 months if mute or CA < 46 months if echolalic and prorated mental age of >11months at CA 30 months.	Five subjects deemed untestable at intake, 3 in experimental group and 2 in control group 1. Intelligence scores based on mental age score from Vineland Social Maturity Scale (Doll, 1953) were used in these cases. INDA.
	Control	40.9 (10.3)	57.1 (14.5)		50.1 (22.4)		<10	24	19	11	8			
(Anderson, Avery, DiPietro, Edwards, & et al., 1987)	EIBI (clinic)	42.8 (11.8)	55.0 (18.9)	47.7 (8.2)	60.6 (25.1)	57.7 (15.3)	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 hours per week. 1 child was 18 month and 1 child was 23 months at intake. Duration was either 1 year or 2 years.	
(Smith, Eikeseth, Klevstrand, & Lovaas, 1997)	EIBI (clinic)	36.0 (6.9)	27.8 (4.9)	50.3 (9.1)	35.8 (14.3)	51.7 (17.9)	30	24	11	11	0	QCT/Archival data. Included if CA ≤ 46 months and IQ < 35.	Post measures conducted 3-4 years after treatment in some cases. VABS data only available for 6 of the 11 children in the ABA group. Control group received minimal treatment.	
	Control	38.0 (5.4)	27.3 (5.4)		24.0 (8.2)		<10	24	10	8	2			
(Weiss, 1999)	EIBI (parent)	42.0		49.9 (7.8)		83.6 (28.3)	40	24	20	19	1	UCT/Enrollment centre.	Workshops were done every 4-6 weeks. Mix of clinic and parent managed programs. Entire caseload of clinician from 80 children enrolled at centre.	
(Harris & Handleman, 2000)	EIBI (clinic)	49.0 (8.8)	59.3 (24.2)		77.6 (28.6)		40	12-36	27	23	4	UCT/Enrollment centre.	Number of hours per week between 35 and 45. Follow-up testing done 4-6 years after treatment. Duration of treatment 1, 2 or 3 years. INDA	
(Smith, Groen, & Wynn, 2000)	EIBI (clinic)	36.1 (6.0)	50.5 (11.2)	63.7 (9.6)	66.5 (24.1)	61.3 (28.7)	24.5	24	15	12	3	RCT/Matched-pair random. Included if CA < 42 months and ratio IQ between 35 and 75.	Clinic directed group: Number of hours for ABA group are for first year in treatment. Gradual reductions in year two. Treatment phased out after 18 months for children responding slowly. Average duration 33 months. Parent managed group: 5 hrs a week of parent training for first 3-9 months, parents asked to do 5 hours a week in between sessions: Total < 10 hours per week of ABA + 12.5 hours of special education classes per week. ABA treatment hours second year presumed to be gradually decreasing, school hours 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.	
	Control	35.7 (5.4)	50.7 (13.9)	65.2 (9.0)	50.5 (20.4)	59.9 (16.7)	<10	~24	13	11	2			

Table 4.1: continued

(Smith, Buch, & Gamby, 2000)	EIBI (parent)	35.8 (4.8)	43.2 (11.8)	54.8 (4.5)	58.3 (19.3)	60.2 (16.7)	26	24	6	6	0	UCT/Consecutive referrals. Included if CA < 48 months.	Two boys deemed untestable and IQ set to 30. Post treatment after 2-3 years. Children had to be under 48 months at intake. Average of 26.2 hrs a week the first 5 months, after that ca 30 hrs for 5 of the children (1 dropped out), supervision monthly.
(Sallows & Graupner, 2005)	EIBI (clinic)	33.7 (3.9)	48.8 (8.8)	59.8 (5.7)	70.8 (24.6)	63.4 (23.6)	38	24	13	11	2	RCT/Matched-pair random.	In order to keep the most variables as constant as possible year 2 outcome data was obtained from the authors. Intensity data from Sallows & Graupner (2005). INDA
	EIBI (parent)	30.2 (3.9)	44.4 (8.2)	54.4 (5.3)	63.8 (23.5)	58.9 (21.8)	31	24	11	8	3	Included if CA < 42 months and ratio IQ ≥ 35.	
(Howard, Sparkman, Cohen, Green, & Stansislaw, 2005)	EIBI (clinic)	30.9 (5.2)	58.5 (18.2)	70.5 (11.9)	89.9 (20.9)	81.3 (11.1)	25-40	14	29	25	4	QCT/Parental preference and IEP teams. Included if CA < 48 months.	EIBI: Multiple settings (home, school and community. 25-30 h per week under 3 years of age, 35-40 h per week over 3 years of age. AP (autism educational programming): Public classroom for children with autism. 1:1 or 1:2 staff: child ratio. 25-30 hrs 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. GP (generic educational programming): Local community special education classrooms. Average of 15 hrs per week intervention, 1:6 staff: child ratio. 13 children received speech and language therapy for 1-2 times per week. INDA.
	Comparison (AP)	37.4 (5.7)	53.7 (13.5)	69.8 (10.5)	62.1 (19.6)	69.1 (12.9)	25	13	16	13	3		
	Control (GP)	34.6 (6.5)	59.9 (14.8)	71.6 (10.5)	68.8 (15.3)	68.3 (9.9)	15	15	16	16	0		
(Cohen, Amerine-Dickens, & Smith, 2006)	EIBI (clinic)	34.4 (5.4)	62.0 (16.4)	64.0 (8.4)	81.1 (21.8)	79.5 (13.4)	35-40	24	21	18	3	QCT/Parental preference.	Community-non-university setting. Community services selected by family. In control group 1 child had a Early Start Autism Intervention Program 9 hrs a week, 2 children home-based development program 1-4 hrs a week, 17 SDC (special day class) eclectic, ratio 1:1 to 3:1, 3-5 days a week for up to 5 hrs. Speech, behavioral and occupational therapies 0-5 hrs per week. 3 where mainstreamed for up to 45 minutes a day. INDA.
	Control	33.2 (3.7)	59.4 (14.7)	71.9 (11.5)	65.9 (16.5)	70.7 (13.3)	-	24	21	17	4	Included if CA < 48 months and ratio IQ > 35.	
Australia													
(Bimbrauer & Leach, 1993)	EIBI (clinic)	38.1 (7.1)	45.3 (17.9)	47.5	57.6 (18.7)	41.0 (5.1)	19	22	9	5	4	QCT/Parent willingness and geographical. Included if CA between 24 and 48 months.	Untestable subjects set to IQ of 30. Ratio scores computed for the rest of subjects. Scores post treatment are deviation IQ used where available. Ratio VABS scores calculated both pre and post. INDA.
	Control	33.2 (10.3)	45.0 (9.4)	51.5	43.2 (15.0)	42.5 (4.9)	-	22	5	5	0		
Israel													
(Ben-Itzhak & Zachor, 2007)	EIBI (clinic)	25.9 (3.2)	71.4 (18.8)	65.9 (7.1)	82.9 (23.2)	90.7 (12.3)	35	12	21	20	1	UCT/Enrollment centre.	Only children that are available for year 2 follow-ups have been included here. 5 children 23 months and 2 children 21 months at intake. INDA.
Norway													
(Eikeseth, Smith, Jahr, & Eldevik, 2002)	EIBI (clinic)	66.3 (11.3)	61.9 (11.3)	55.8 (9.0)	79.1 (18.1)	67.0 (16.3)	28	12	13	8	5	QCT/Staff availability. Included if CA between 48 and 84 months and IQ ≥ 50.	Comparison received eclectic treatment of similar intensity. INDA.
	Comparison	64.8 (9.9)	65.2 (15.0)	60.0 (13.2)	68.9 (18.8)	60.2 (11.7)	29	14	12	11	1		
(Eldevik, Eikeseth, Jahr, & Smith, 2006)	EIBI (clinic)	53.1 (9.5)	41.0 (15.2)	52.5 (3.9)	49.2 (16.6)	52.4 (9.2)	13	20	13	10	3	QCT/Archival data. Included if CA < 72 months.	Comparison received eclectic treatment of similar intensity. Four subjects from comparison group taken out here, because included in Eikeseth et al. (2002). INDA.
	Comparison	45.1 (16.5)	42.8 (13.0)	50.1 (9.2)	38.5 (15.5)	44.6 (7.5)	12	23	11	10	1		
United Kingdom													
(Hayward, Eikeseth, Gale, & Morgan, in press)	EIBI (clinic)	35.7 (6.2)	53.5 (15.1)	62.3 (6.8)	72.0 (19.6)	68.3 (14.5)	37	13	23	19	4	QCT/Geographical. Included if CA < 42 months.	INDA
	EIBI (parent)	34.4 (5.7)	54.7 (15.3)	65.1 (10.4)	69.7 (22.9)	72.5 (17.3)	34	14	21	15	6		
(Remington et al., 2007)	EIBI (clinic)	35.7 (4.0)	61.4 (16.7)	60.2 (5.8)	73.5 (27.3)	61.5 (15.4)	26	24	23	18	5	QCT/Parent preference.	Control group received TAU, special school, mainstream or mix, but little or no 1:1, speech therapy, TEACCH etc. INDA.
	Control	38.4 (4.4)	62.3 (16.6)	57.0 (6.8)	60.1 (27.8)	54.6 (13.1)	16	24	21	18	3	Included if CA between 30 and 42 months.	

INDA=individual data obtained from author, UCT=Uncontrolled clinical trial, QCT=Quasi-experimental controlled clinical trial, RCT=Randomized controlled clinical trial

Table 4.2:

The available number of subjects and demographic characteristics of the entire sample and subgroups.

	Behavioral treatment (n=309)	Control (n=105)	Comparison (n=39)	Total (N=453)	F
Age (n)	278	95	39	412	
Mean and SD (months)	38.0 (11.4)	36.5 (7.1)	47.6 (15.9)	38.5 (11.5)	14.94**
Range	16-84	18-72	21-84	16-84	
Intelligence (n)	286	105	39	430	
Mean and SD	55.6 (18.2)	54.8 (17.1)	54.0 (16.2)	55.3 (17.7)	.183
Range	17-120	19-97	21-101	17-120	
Adaptive Behavior (n)	252	73	39	364	
Mean and SD	60.3 (10.9)	65.0 (11.6)	61.2 (13.7)	61.3 (11.5)	4.83*
Range	26-95	45-113	37-96	26-113	

**p<.001

*p<.05

To explore predictors of IBI outcomes, we conducted a multiple regression analysis for the behavioural 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 intake, and intensity of treatment (median split of intensity at 36 hours per week). To protect against some errors of statistical inference all variables were centred following the guidelines suggested by Kraemer and Blasey (2004). Thus, the binary independent variable (high or low intensity of treatment) was recoded as either +1/2 or -1/2, and all other independent variables (age, IQ, and ABC scores at intake) were centred 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 centred variables. For IQ change, we included interaction terms for: age and IQ at 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 IBI, control and comparison groups achieving reliable change on IQ and ABC is displayed in Figure 2. Each bar on the graph in 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 from 279 (29.8 %) in the IBI group achieved reliable change in IQ and 51 from 248 (20.6 %) achieved reliable change in ABC scores. In the control group, 9 from 104 (8.7 %) achieved reliable change in IQ and 4 from 70 (5.1 %) achieved reliable change in ABC scores. In the comparison interventions group, 1 from 39 (2.6 %) achieved reliable change in IQ and 2 from 39 (5.7 %) achieved reliable change in ABC scores. The proportions in the three groups were compared statistically using 3x2 chi-square tests. There was a significant difference in the proportions achieving reliable change for IQ (χ^2 (df=2, N=422) =29.11, $p<.001$) and for ABC (χ^2 (df=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 IBI group, and fewer children than expected achieving this change in the two other groups. Exploratory 2x2 chi-square comparisons between the control and comparison group for IQ and ABC change revealed no difference between these two groups.

Because there was little difference between the control and comparison groups, we combined them to carry out the individual participant data meta-analysis focusing on the NNT and ARR for IBI. The NNT 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 ARR of 23 % (95 % CI [16.0 %, 29.6 %]) and 16 % (95 % CI [10.2 %, 22.3 %]) respectively in favour of the IBI group. The NNT and ARR for IQ and ABC, along with the 95 % confidence intervals for the individual studies (i.e. the controlled studies where there is a comparison or control group against which to calculate an effect size) are shown in 4.s 3 and 4.4. At the level of individual studies, there is considerable variability in effect sizes and many of the individual studies focus on small samples and so are underpowered. The multiple regression analyses for prediction of IQ and ABC change are summarized in Tables 4.5 and 4.6. 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, $F(7, 208) = 3.87$, $p < .001$; $R^2 = .115$, adjusted $R^2 = .086$, and ABC change $F(7, 210) = 8.39$, $p < .001$; $R^2 = .219$, adjusted $R^2 = .193$. The results from the regression analyses showed that (high) intervention intensity was the only variable that independently 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 two 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, 2009), we were able to gather such data for every one of 16 evaluation studies of IBI identified via a systematic review. Only data from one of Lovaas' (1987) original control groups were unavailable. After combining data from control groups and comparison interventions, an individual participant data meta-analysis showed meaningful differences in outcomes for children with autism in favour of IBI. For IQ, the Number Needed to Treat was 5 (Absolute Risk Reduction = 23 %), and for the adaptive behaviour composite the NNT was 7 (ARR = 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 IBI outcome studies and also 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 favourably to psychological and medical treatments for common disorders such as major depression (NNT between 3 and 5), obsessive compulsive disorders (NNT between 4 and 5), and bulimia nervosa (NNT = 9) (Pinson & Gray, 2003). We have not been able to locate published NNT or ARR data for other interventions for autism. The decision to offer interventions cannot be made by looking at the NNT 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.

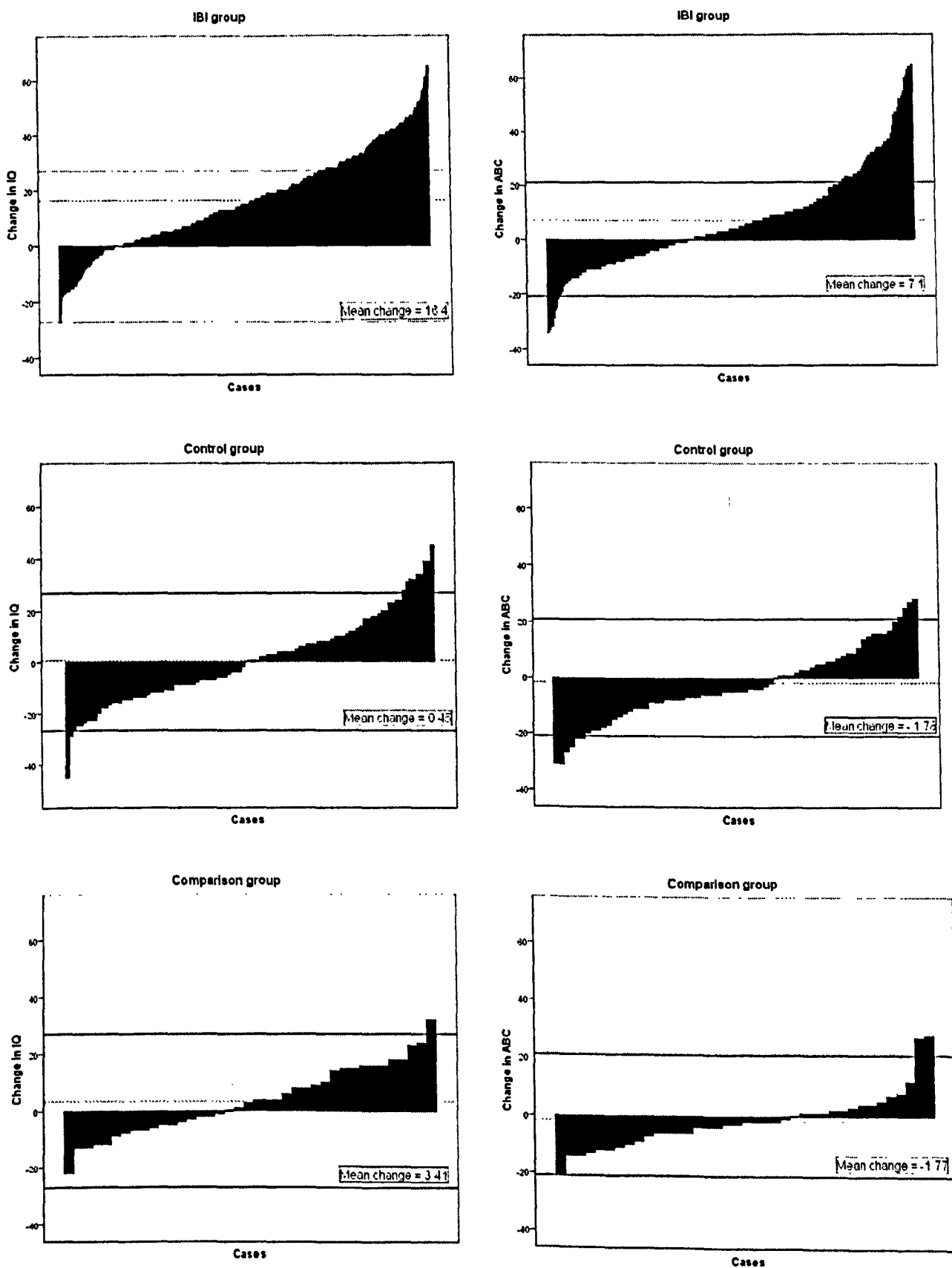


Figure 4.2. Bars indicate changes in IQ and ABC scores for children in the IBI, control and comparison groups. The lines at ± 27 IQ points and ± 21 ABC points show the criteria for reliable change. The dotted line shows the mean change for the group.

Table 4.3:

The number of children that met the reliable change criteria for intellectual functioning (27 IQ points), the number needed to treat (NNT) and the absolute risk reduction (ARR) with 95% confidence intervals for the controlled studies that were included in the present analysis.

Study	Groups	Outcome Intelligence					
		RCI+	RCI-	NNT	95% CI NNT/NNH	ARR	95% CI
Lovaas, 1987	IBI	9	10	3	1,5 - 5,7	42,0 %	17,5 - 66,7 %
	Control	1	18				
Eikeseth et al., 2002	IBI	3	10	5	2,2 - 575,3	23,1 %	0,2 - 46,0 %
	Comparison	0	12				
Bimbrauer & Leach, 1993	IBI	2	7	5	NNT 2,0 to ? to NNH 20,2*	22,2 %	-4,9 - 49,4 %
	Control	0	5				
Smith et al., 2000 (RCT)	IBI	6	9	3	1,5 - 6,6	40,0 %	15,2 - 64,8 %
	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 ? to NNH 12,7*	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 to ? to NNH 6,6*	7,5 %	-15,1 - 30,0 %
	Control	3	18				

*Because the 95% confidence interval for the absolute risk reduction extends from a negative number where treatment may harm (NNH), to a positive number where treatment may benefit (NNT), 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 ? to NNH x (adopted from Altman, 1998).

Table 4.4:

The number of children that met the reliable change criteria for adaptive behavior (21 ABC points), the number needed to treat (NNT) and the absolute risk reduction (ARR) with 95% confidence intervals for the controlled studies that were included in the present analysis.

Study	Groups	Outcome Adaptive Behaviors					
		RCI+	RCI-	NNT	95% CI NNT/NNH	ARR	95% CI
Eikeseth et al., 2002	IBI	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 to ? to NNH 1,8*	-20,0 %	-15,1 - 55,1 %
	Control	1	4				
Smith et al., 2000 (RCT)	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 to ? to NNH 14,7*	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 to ? to NNH 7,1*	9,2 %	-14,1 - 32,6 %
	Comparison	2	14				
Cohen et al., 2006	IBI	4	15	11	NNT 3,0 to ? to NNH 7,4*	9,9 %	-13,4 - 33,3 %
	Control	2	16				
Remington et al., 2007	IBI	2	21	26	NNT 5,4 to ? to NNH 9,3*	3,9 %	-10,8 - 18,6 %
	Control	1	20				

Table 4.5:
Regression analysis of predictors of IQ gain in the IBI group.

	IQ gain	
	beta	p
Main effects		
IQ at intake	-.132	.081
ABC at intake	.092	.263
Age at intake	-.197	.109
Intensity	.232	.001
Interactions		
Age at intake X IQ at intake	.069	.313
Age at intake X Intensity	-.209	.063
IQ at intake X Intensity	-.095	.184

Table 4.6:
Regression analysis of predictors of ABC gain in the IBI group.

	ABC gain	
	beta	p
Main effects		
IQ at intake	.411	.000
ABC at intake	-.366	.000
Age at intake	.014	.904
Intensity	.280	.000
Interactions		
Age at intake X ABC at intake	.073	.346
Age at intake X Intensity	-.076	.531
ABC at intake X Intensity	-.032	.657

Full data on these variables are not currently available in the field of autism. However, it is informative to note that there appears to be no additional negative psychological impact on family members associated with IBI (Hastings, 2003; Hastings & Johnson, 2001; Remington et al., 2007). Furthermore, autism-specific eclectic pre-school services may cost no less than home-based IBI (Magiati et al., 2007).

The present analysis confirms that IBI 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 IBI 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 IBI evidence base meets the criteria for 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.

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. (2009) 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 4.3 and 4.4 that there is considerable variability in the estimates of effect sizes (NNT and ARR) at the level of individual studies. These tables only include controlled studies that could be used to generate study level effect sizes (i.e., pre-post test 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 an intervention's evidence base.

In addition to the variability summarized in Tables 4.3 and 4.4, in applying the general common elements of IBI defined by Green et al. (2002), we may risk combining quite different interventions. For example, we have made no distinction between centre-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 IBI programs are likely to stress the use of techniques differently; some may be based heavily on discrete trial training, others incidental teaching, others pivotal response training, and still others, verbal behaviour 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 IBI group of 309 children. The results from these regression analyses showed 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 two 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 IBI. 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 IBI. 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 IBI 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 (Jacobson, Mulick, & Green, 1998; Jarbrink & Knapp, 2001). Under many circumstances, a change equivalent to one standard deviation 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 over-estimates 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 IBI used, quality of outcome measurement) but we did not exclude studies on the basis of research design (apart from excluding 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 later meta-analyses will need to incorporate research quality selection criteria when the body of randomized studies available for analysis is larger.

Chapter 5: Behavioural Intervention for Children with Autism in Local Mainstream Pre-school Settings⁶

In 2000 the county of Oslo, Norway opened an early behavioural intervention centre for children with autism. As is the current policy in Norway, children receiving services from the centre were enrolled in their local pre-schools. Children were referred from local education departments and the service model was funded using resources typically provided to support eclectic support for young children with autism. Thus, the behavioural intervention service was provided through a re-direction of existing financial resources. The services included training and supervision of local pre-school staff for the day to day running of the intervention program. Frequency of supervision and training were at the levels typically recommended for Intensive Behavioural Intervention (IBI) programs. Outcome, particularly on IQ, was satisfactory compared to recently suggested benchmarks for evaluating IBI programs. Correlation analysis revealed that the only predictors of outcome were weekly intervention hours and diagnosis. Strengths and weaknesses of the model are discussed along with some suggestions for improvements.

⁶ This research has received approval by the regional committee for medical research ethics in Norway.

Introduction

The benefits of Intensive Behavioural Intervention (IBI) for young children with autism are being increasingly documented. Recent reviews of the literature suggest that the effectiveness of IBI is well established according to commonly used criteria for evidence based practices (Eikeseth, 2009; Rogers & Vismara, 2008) and meta-analytic methods reveal large and moderate effect sizes for outcome assessed via standardized tests of intelligence and adaptive functioning respectively (Eikeseth, 2009; Eldevik, et al., 2009; Howlin, Magiati, & Charman, 2009; Reichow & Wolery, 2009; Rogers & Vismara, 2008). To date, the majority of published studies of IBI included in systematic reviews have focused on its delivery by specialist teams and often with implementation staff engaged as part of a research evaluation project. The question of whether IBI might be delivered successfully in more typical service settings, or on a reasonably large scale, has been relatively neglected.

There are various models described in the literature on how IBI can be implemented (Handleman & Harris, 2001). These range from a full time place in a centre (or an institution), home-based programs, directed either through a clinic or the parents themselves, to full time placement in mainstream pre-school, with IBI being implemented at home before or after pre-school and at the weekends. Most of the outcome data published so far has come from home-based programs in some form. Home-based IBI has been organized through university clinics (Lovaas, 1987; Remington et al., 2007), and community based agencies or clinics (Cohen, et al., 2006; Howard, et al., 2005; Sallows & Graupner, 2005). There are a few examples of evaluations of intervention models delivered outside of the home setting. In Norway, Eikeseth, et al., (2002) evaluated IBI for young children with autism in pre-school settings, and in Israel Ben-Itzhak and Zachor (2007) evaluated IBI provided at an intervention clinic.

In terms of outcomes for IBI when delivered on a large scale, Perry et al. (2008) recently reported on IBI delivered throughout a Canadian provincial area. In this study, IBI was provided in a variety of settings (including centre-based and integrated child care settings) and the evaluation was focused on the effectiveness of IBI as it

might be typically delivered clinically rather than in a controlled research study. Notably, staff training had to be conducted on a large scale in a limited time and no children were excluded based on co-morbid diagnosis, low cognitive ability, or age. Also, children were referred from a large and diverse socioeconomic group. Ostensibly, any one of these factors could lead to less favourable outcome. The study reported outcome for 332 children with autism between 2 and 7 years. Like in other outcome studies there was considerable variation in outcome but overall the children made statistically and clinically significant improvements in intellectual and adaptive functioning and autism severity.

Initial data on delivery of IBI in various settings and even on a large scale and in typical clinical practice are encouraging. However, there is still a need to investigate the effectiveness of models of service delivery in real world settings. Such settings will vary considerably from country to country (and within countries), and thus a variety of models will need to be evaluated. In the county of Oslo, Norway in 2000, a project was started that focused on the provision of an IBI intervention model for pre-school children with autism. As is the current policy in Norway, all children receiving services from this project were enrolled in their local mainstream pre-schools. Referrals were taken from local education departments (of which there were seven) each covering a designated geographical area of the city, and the services involved no extra financial costs for the family, the pre-school, or the community. The project was to provide specialist intervention services directly to the pre-schools, thus supporting the local education departments with some of their most difficult cases.

No financial/staffing resources in addition to those typically given to a pre-school enrolling a child with autism were given. A pre-school enrolling a child with autism will typically receive funding for one additional full-time staff member and supervision and training from a special education teacher and/or speech and language therapist employed by the education departments for 2–5 hours a week. The aim of the current project was instead to provide a behavioural intervention program using the resources associated with this support model. The intervention model was implemented using the same staff resources but instead of the pre-school receiving supervision and training from local education department professionals, this was provided through the behavioural intervention centre. The pre-school staff were

responsible for the day-to-day running of the behavioural intervention program while being supervised and trained through the centre.

The purpose of the present paper is to describe the key features of the Oslo mainstream pre-school delivery model and to provide a summary of outcome data for all children who have received services through the centre from its inception to July 2009.

Methods

Participants

All children who received intervention through the centre from its inception to July 2009 were included in the present analysis. In terms of evaluation data we included children who: (a) had been independently diagnosed with autism or atypical autism based on the ADI-R (Lord, et al., 1994); (b) were between 2 and 6 years of age at intake; (c) had a full-scale intelligence test and a measure of adaptive behaviour at intake and after 1 and/or 2 years of intervention, and; (d) had received at least 5 hours per week of behavioural intervention. In terms of inclusion criterion (c), the final group of children had been tested on average 14 months from intake and then again at 27 months from intake. For ease of reference, these outcome points are referred to in the remainder of this paper as Year 1 and Year 2.

Twenty five children (21 boys, 4 girls) met all the inclusion criteria and their data are presented below, 23 of these had 1 year outcome, and 18 had 2 year outcome. Two children with 2 year outcome data did not have 1 year outcome data available. A more detailed description of the children is provided in Table 5.1. Overall, the sample is fairly representative of the autism population (Volkmar & Klin, 2005). Six children referred to the centre were excluded from the present study either due to inadequate assessments (4 children) or because their programs provided fewer than 5 intervention hours per week (2 children).

Table 5.1:

The mean (SD, range) age at intake, weekly hours and time between testing, along with diagnosis and gender.

Descriptor	Intervention Group (N=25) Mean, SD and range
Age at intake (months)	41.9 (7.8, 26-56)
Diagnosis	
Autistic disorder	19 (2 girls, 17 boys)
PDD-NOS	6 (2 girls, 4 boys)
Weekly intervention hours	
Year 1 (n=23)	13.6 (6.2, 5-28)
Year 2 (n=18)	14.3 (5.5, 5-28)
Time of testing (months)	
Year 1 (n=23)	14.2 (2.6, 6-19)
Year 2 (n=18)	27.1 (3.8, 15-33)

Pre-School Setting

The county provided funding for an intervention service employing one psychologist and four supervisors. The psychologist (first author) was a Board Certified Behavior Analyst with approximately 15 years of experience implementing IBI programs. The supervisors had bachelor degrees (in caring for and habilitating individuals with intellectual disability) covering the basics of applied behaviour analysis and between two and ten years of training and experience with IBI programs.

All children attended their local mainstream pre-school. As required by Norwegian regulations, mainstream pre-schools units were staffed on 1:3 staff to child ratio for children under the age of 3 years, and a 1:6 staff to child ratio for children between three and six years of age (children started school proper at six years of age). In their last year of pre-school, special "clubs" are arranged to prepare the children for

school. Typically a unit would either consist of nine children below the age of 3 with three staff, or 18 children between the age of three and six with three staff. When the unit enrolls a child with autism it will typically receive resources for one more full-time staff member; thus making it possible to cover this child 1:1 without taking resources away from the other children in the unit. This extra resource is granted independent of any intervention (and of the current study).

Rather than having just this extra staff member cover the child with autism, we recommended that a rota was made so that all members of staff at the unit would work with the child with autism during a week, and that 2–3 staff members should form a team that would be responsible for the day-to-day implementation of the intervention. This was done so that the child would get used to interacting with a number of different adults and so as not to make the program dependent on just one person. One of these staff would be given responsibilities for scheduling and monitoring intervention hours for the staff involved, preparing the weekly-team meetings, updating the programme folder, and finding the instructional materials needed for the various programs. All units had a separate room available where 1:1 intervention could be done without disturbance from the rest of the unit. Pre-schools were typically open weekdays from 7:00 am to 5:00 pm, and children in the present study typically stayed for at least 20 hours every week (typically, a minimum of four hours per day).

Staff Training and Supervision

The IBI centre was responsible for training and supervision of all staff involved in the intervention. The model used for staff training and supervision was similar to that described as clinic-supervised intervention by Smith, Donahoe, & Davis (2001) and Eikeseth, et al., (2002). Staff training started with a 3 day workshop and continued throughout the duration of the intervention with weekly (or eventually in some cases bi-weekly) consultations lasting 1 to 4 hours. In addition, weekly two-hour team meetings were held for each child. The child, primary caregiver(s), and staff attended both the workshops and the team meetings, and all were trained using an apprenticeship model. The supervisor first explained and demonstrated how to do a program, and then the staff took turns doing the program with the child, while being

coached by the supervisor and the other team members. At team meetings, the child's program and/or intervention procedures were reviewed and modified based on the child's progress during the preceding week. The program was comprehensive and balanced covering all important areas of the child's life. Each week the child would normally be actively engaged with 10–20 teaching programs. Parental participation was central to ensure generalization and maintenance of skills to the home and other community settings.

Depending on the needs of the individual child, parts of the intervention were provided outside of the separate teaching room, targeting specific weekly goals (e.g., conducting incidental teaching for expressing wants and needs, providing instruction on self-help skills such as putting on shoes, teaching peer interaction skills, or implementing behaviour management plans). However, some of the 1:1 hours in the unit merely focused on practical help for getting dressed or undressed, eating, toileting, and going outside on the playground, without the use of systematic behavioural teaching methods. For the purposes of this study, only the hours spent implementing behavioural teaching methods towards the specified weekly targets, (whether inside or outside of the teaching room), were counted as behavioural intervention hours, while the hours of other practical assistance were not counted. The supervisors had a caseload of 4–8 children. They met weekly (or more often if required) with the psychologist to discuss programming or any particular problems arising with the individual child. The psychologist would also oversee individual programs by attending team-meetings at least once a semester.

Behavioural Intervention

The behavioural intervention was based on several widely used IBI manuals (Lovaas, 1981; 2003; Maurice, et al., 1996). In short, the intervention began with establishing basic tasks, such as expressing wants and needs, responding to simple requests made by an adult, imitation of gross motor behaviours, matching of objects or pictures, and teaching of simple toy play such as completion of puzzles or putting shapes in a shape sorter. When these tasks were mastered, the intervention moved on to more complex skills such as imitation of fine motor and oral motor behaviours, imitation of sounds and words, and recognizing objects and actions upon request.

After the child had acquired vocal imitation of words and basic receptive language, the child was taught to use the words functionally, for example by naming objects and actions. Next, more abstract concepts such as colour, size, adjectives, and prepositions were targeted. Subsequent intervention goals included discriminating wh-questions, conversing, and making friends with peers. From the start, the intervention also targeted other play and social skills, progressing from functional toy play and parallel play to symbolic play and cooperative play.

All procedures were based on documented operant conditioning techniques such as differential reinforcement, shaping, chaining, task analysis, and prompt and prompt fading. In the early stages of intervention, most teaching took place in a 1:1 discrete trial format. Later, the focus gradually shifted to include small group settings with typical children to help generalize skills and adjust to the routines of the unit. The ultimate goal of the intervention was to improve the ability of the children to learn in natural settings as might be expected of their typically developing peers.

Outcome Measures

We employed measures of full scale intellectual functioning and adaptive behaviour that are widely used and recommended for assessing children with autism (Klin, Saulnier, et al., 2005).

Intellectual functioning. The Bayley Scales of Infant Development (BSID-II) (Bayley, 1993) were used for the youngest children or children that scored below the basal on intelligence tests standardized for their chronological age. The BSID-II is a measure of mental development for children up to 42 months. It will yield a mental developmental index (MDI), which was considered broadly equivalent to an IQ score. If the child scored below the norms on this test or was too old for the norms, we computed a ratio IQ score by dividing the obtained mental age with chronological age and multiplying by 100. For the older children we used the Stanford-Binet Intelligence Scale: Fourth or Fifth Edition (Thorndike, et al., 1986; Roid, 2003), or the Wechsler Preschool and Primary Scale Intelligence-Revised (Wechsler, 1989).

Adaptive behaviour. The Vineland Adaptive Behavior Scales (VABS I or II); (Sparrow, et al., 1984; Sparrow, et al., 2005) were used for measuring adaptive behaviour. The VABS yields standard scores on four domains; communication, daily living, socialization, and for children under six years old, motor skills. Based on these scores the VABS also yields a standardized adaptive behaviour composite (ABC).

Approximately 65 % of the assessments were carried out by professionals blind to the purposes of the present study (76 % at intake, 44 % after one year, and 78 % after 2 years). The remainder of the assessments was conducted by the first author. Six of these administrations were performed both by an independent professional and by the first author within 3 months of each other. In these instances a conservative measure of improvement was obtained by using the higher score at intake and the lower score in the later assessments. Agreement on the total scores was within ± 5 points in all of these cases.

Results

We analyzed the outcome data in three ways. First, we examined pre-post mean change after 1 and 2 years of intervention respectively. Because the number of participants available after 1 and 2 years differed, we conducted two separate paired samples t-tests in order to maximize the sample evaluated at each time point. This analysis was carried out for intelligence scores and the adaptive behaviour composite scores, as well as for scores for the communication, daily living and socialization domains on the VABS. Effect sizes for pre-post change on all measures were also calculated. For pre-post change in IQ, the results were benchmarked against Reichow and Worley's (2009) meta-analytic pre-post change effect size from 12 studies (Hedges' $g = .69$).

The second level of analysis was to examine meaningful change at the level of the individual children, following Remington et al. (2007) who used a reliable change analysis (Jacobson & Truax, 1991) for the children in their outcome research. An analysis of reliable change establishes with 95 % certainty that observed changes at an individual level are meaningful and not accounted for by measurement error and sample variance. The amount of change required for IQ and ABC scores to be

considered as reliable was established in chapter 3 from an analysis of almost 300 individual children who received IBI across 16 separate evaluation studies. Here I established, using formulae from Jacobson and Truax (1991) that change in IQ over approximately two years would need to be 27+ points to be considered reliable and 21+ points for ABC scores. I found that 29.8 % and 20.6 % of children met these criteria for IQ and ABC change respectively following IBI. Thus, these results were used to benchmark the results from the present study at the two year outcome point only.

The final exploratory analysis focused on correlates of change. Pearson correlations (2-tailed, and point biserial correlations where a correlate was dichotomous) were computed between IQ and ABC change after 1 and after 2 years with the following variables: age at intake, IQ at intake, ABC scores at intake, child gender, diagnosis (PDD-NOS or autism), and intensity of intervention in year 1 and year 2.

Results of the pre-post test change analysis show that after one year of intervention only change in IQ scores emerged as statistically significant. However, after two years of intervention change in all outcome variables except the daily living domain on the VABS were statistically significant. After two years, the mean change effect sizes (Cohen's *d*) were 2.26 [95 % CI -7.21, 8.88] for IQ scores and 1.43 [95 % CI -.338, 4.56] for ABC scores (see Table 5.2 for mean scores on all measures after 1 and 2 years). The analysis of change after two years included the two children who did not have outcome data available after one year. To check if the inclusion of these two children at the two year evaluation point explained the differences between one and two year outcomes, we carried out the same analyses only including children who had outcome data after both one year and two years ($n=16$). All outcomes except the daily living domain were also significant in this reduced sample. Data from the individual children after one and two years of intervention are displayed in Figure 5.1. Only two variables were significantly correlated with outcome. Intensity of intervention (weekly hours) in year 2 was positively correlated with the total change in IQ after 2 years, of intervention and a diagnosis of PDD-NOS was related to more change in ABC scores also after 2 years of intervention (see Table 5.3).

Table 5.2:

Mean scores at intake and after one and two years of treatment along with the results of the paired samples t-test.

Measures	Intake (n=25)			After 1 year (n=23)				df	t	Effect Size (95% CI)	M	SD (range)	n
	M	SD (range)	n missing s's	M	SD (range)	n	missing s's						
Intellectual functioning	50.0	16.9 (24-94)	25	61.0	22.5 (28-103)	23	1, 2	22	4.865**	1.59 (-7.61-8.21)	69.6	20.5 (29-96)	18
Adaptive Behavior Composite	60.7	8.0 (46-77)	25	61.9	10.1 (46-80)	23		22	.721	0.20 (-3.92-3.34)	70.0	10.4 (47-97)	18
Communication	60.8	9.1 (48-87)	24	63.3	13.3 (42-91)	23		21	1.198	0,40 (-5.04-4.04)	73.7	15.2 (45-114)	18
Daily Living	67.4	10.3 (48-88)	24	67.3	10.7 (47-87)	23	2, 10	21	.122	-0.01 (-4.39-4.11)	72.4	9.5 (52-87)	18
Socialization	62.2	10.4 (49-97)	24	63.8	9.9 (52-85)	23		21	.703	0.26 (-3.79-4.42)	70.2	10.7 (54-90)	18

*p < .01. **p < .001.

Table 5.3:

Correlations between scores at intake, weekly intervention hours, gender and outcome in terms of gain in IQ and ABC scores.

Outcome measures	Intake variables					Intervention variables	
	Age	IQ	ABC	Gender	Diagnosis	Weekly hours year 1	Weekly hours year 2
Change in IQ year 1 (n=23)	.033	.178	.147	-.158	.048	.240	.472
Change in IQ year 2 (n=18)	-.238	.091	.093	-.293	.149	-.047	.502*
Change in ABC year 1 (n=23)	.278	.113	-.280	.308	.341	-.009	.074
Change in ABC year 2 (n=18)	.407	.163	-.269	.314	.516*	.124	.406

*p < .05 (2-tailed)

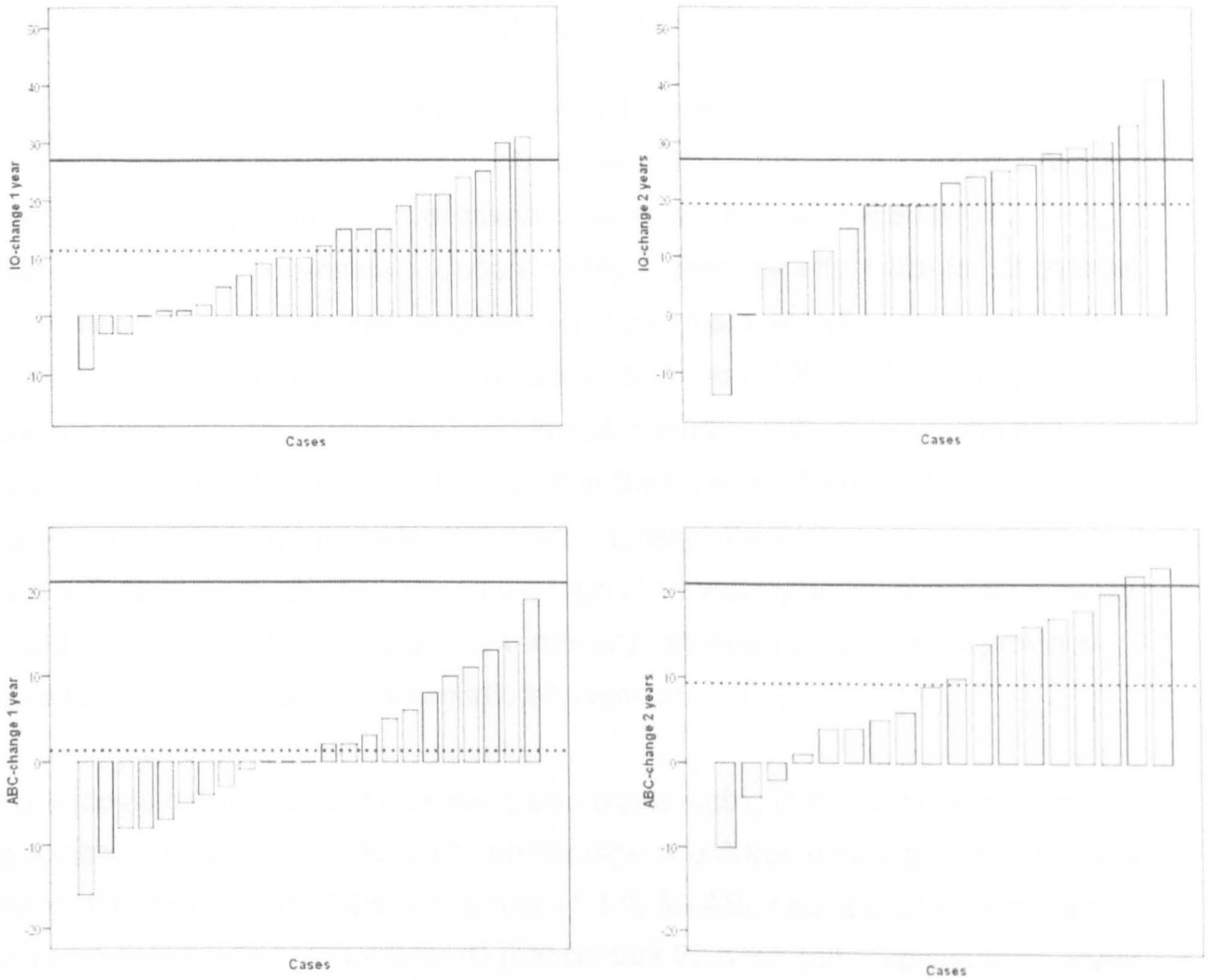


Figure 5.1. Bars indicate changes in IQ and ABC scores for individual children following 1 and 2 years of intervention. The left column shows the change after one year and the right column the change after two years. Results are sorted from highest negative to highest positive. The solid lines represent the suggested Reliable Change Index (RCI) from chapter 4, 27 for IQ and 21 for ABC, the dotted line the mean change in each group. For IQ five out of 18 children (27.8 %) met criteria for reliable change (27+ IQ points) after two years of intervention. For ABC, two out of 18 children (11.1 %) met the criterion for reliable change (21+ standard points) after two years of intervention. Binomial tests revealed that neither of these percentages differed significantly from the benchmarks suggested in chapter 4.

Discussion

Children receiving intervention under the current model made statistically significant gains in IQ after one and two years of intervention. The mean change effect size for both IQ and ABC after two years of intervention should be considered large according to current conventions (Cohen, 1988). Indeed the effect size for IQ change was somewhat higher than that reported in a recent meta-analytic review employing the same type of effect size measure (Reichow & Wolery, 2009). The average gain from this meta-analytic review, however, included studies with children who had substantially lower IQ scores at intake (27.8 in Smith, et al., 1997; and 41.0 in Eldevik, et al., 2006, for example) and lower-intensity interventions (12.5 hours a week in Eldevik, et al., 2006). With an average of 14 weekly hours, the present study should probably also be considered low-intensity, although children were provided with almost 3 hours a day of systematic intervention.

The outcome data presented here were also encouraging in terms of meaningful outcomes for individual children. The percentage of children meeting reliable change criteria after two years (27.8 % for IQ and 11.1 % for ABC) compared favourably with the benchmarks from chapter 3 for IQ (benchmark 29.8 %), but is considerably lower for ABC (benchmark 20.6 %). The present outcomes did not differ statistically from the ABC benchmark, but this is very likely to be related to a lack of power to detect this difference in the small sample. Relatively low gains in ABC scores have also been reported in other low-intensity intervention studies (Eldevik, et al, 2006).

We found that weekly hours of intervention correlated significantly with gains in IQ after 2 years. This result is in line with what we found in chapter 3 and others (e.g., Lovaas, 1987) and underpins the point that intensity of intervention may be one determinant of outcome. However, there is no agreed standard for measuring intensity of behavioural intervention and the validity of our measurement needs to be questioned. In particular, it may be that we have counted intervention hours in a more conservative way than other authors. For example, since the staff had considerable training in the IBI techniques it is possible that they provided some sort of intervention (e.g., incidental teaching, systematic fading of prompt) in addition to the intervention hours actually counted. Since the study lasted for nine years there is also a chance of

procedural drift, both planned and unplanned. It is conceivable that the children being enrolled in the program at a later date would get a better program since the field is under constant development.

There was also tentative evidence of an association between autism diagnosis (with those with PDD-NOS performing better) and ABC outcome. This pattern has been found in another study (Smith, et al., 2000). However, the number of PDD-NOS cases in the present evaluation is small (as it was in Smith et al.) and thus this result may not be robust. Examination of differential outcomes for a variety of Pervasive Developmental Disorder remains an issue for future research.

There are also limitations in the research design of the present study. Although our main goal was to evaluate the intervention model against benchmark data, it would have added strength to our findings if we had a comparison group of children receiving another form of intervention with similar intensity. This could rule out that the possibility that gains reported were due to maturity, attention, or simply the fact that children were placed in local mainstream pre-schools. In addition, only IQ and adaptive behaviour outcome data were available. Behavioural intervention has also been associated with positive effects on language and deficits associated with autism (e.g., joint attention) in previous research (e.g., Remington et al., 2007). Thus, future evaluations of this pre-school model and other service delivery models should endeavour to examine a wider range of outcomes.

In addition to the formal outcome data, it is important to review the strengths and weaknesses of this service model for the delivery of behavioural intervention to children with autism. Strengths include that trained staff are with the children the entire day, and inclusion in mainstream provide opportunities for interaction with peers, who may also serve as role models. The weaknesses of this model are also notable: it was in most cases difficult to reach the recommended weekly intervention hours due to competing contingencies on the staff in the mainstream pre-school. Also, there was often a 3–9 months period before the program was up and properly running. The staff and pre-school management were in most cases unfamiliar with (and in some cases opposed to) IBI. Behavioural intervention is different in many ways to the generic education provided in mainstream pre-schools. The close

supervision and monitoring of staff performance and the child's learning, the intensity of intervention, and the structure of teaching (in particular, the discrete trials format) may be at odds with the educational approaches pre-school staff were used to. In most cases, such scepticism was overcome, but in two cases these problems led to the children receiving so few weekly intervention hours that the programs were discontinued. Another limitation is that it was difficult to achieve parental involvement. Since the children were referred through local education departments, it could be that parents had no knowledge of IBI and the homework that was required at the start of the child's program. Indeed, some parents had little or no knowledge of autism. A further improvement of this study would be to keep records of parental involvement (e.g. if they were present at the team meetings, if they did the assigned homework, and so forth). These data could then be correlated with outcome measures. It should also be noted that the intervention did not entail other common elements of EIBI.

The outcome data, and the clinical experience of involvement with this mainstream pre-school model for 9 years, suggest it is a viable option for the delivery of behavioural intervention to young children with autism. The model could perhaps be improved on, if the extra staff that was provided to the pre-schools when they enrol a child with autism were employed directly through the intervention centre instead of being appointed through the local education departments. If they were employed through the intervention centre, it could provide for better continuity and these staff would have the opportunity to build up more experience with IBI which would in turn make it possible to get the programs up and running more quickly. Furthermore, the intervention centre would not have to use as many resources in training new staff.

Chapter 6: Cognitive and Adaptive Behaviour Outcomes of Behavioural Intervention for Young Children with Intellectual Disability⁷

Data from Norway were analyzed to evaluate early behavioural intervention for children with intellectual disabilities. The intervention group (n=11) received approximately 10 hours per week of behavioural intervention; the eclectic comparison group (n=14) received treatment as usual. After 1 year, changes in intelligence and adaptive behaviour scores were statistically significant in favour of the behavioural intervention group (effect sizes of 1.13 for IQ change and .95 for change in adaptive behaviour composite). Approximately 64 % of the children in the behavioural intervention group met objective criteria for reliable change in IQ, whereas 14 % in the eclectic comparison group did so. These results suggest that children with intellectual disability may profit from behavioural intervention typically provided for children with autism.

⁷ This chapter is based on a paper that is accepted for publication: Eldevik, S., Jahr, E., Eikeseth, S., Hastings, R. P., & Hughes, J. C. (2010). Cognitive and adaptive behavior outcomes of behavioral intervention for young children with intellectual disability. *Behavior Modification*, 34, 16-34. This research has received approval by the regional committee for medical research ethics in Norway

Introduction

There are an increasing number of studies using global outcome measures to evaluate effects of Early Intensive Behavioural Intervention (EIBI) as a comprehensive intervention for children with autism spectrum disorders. Recent narrative and meta-analytic reviews suggest that EIBI may meet criteria as a “well-established” intervention, and that effect sizes for IQ and adaptive behaviour outcomes are in the medium to large range (Howlin, et al., 2009; Eikeseth, 2009; Eldevik, et al., 2009; Reichow & Wolery, 2009; Rogers & Vismara, 2008). For example, using a weighted mean difference effect size across 12 controlled studies of EIBI, Eldevik et al. (2009) found a large effect for IQ change (Hedge's $g = 1.10$ [95 % CI .87, 1.34]) and a smaller although still statistically significant effect for change in adaptive behaviour composite scores (Hedge's $g = .66$ [95 % CI .41, .90]).

While there are no obvious reasons why a comprehensive intervention based on behavioural principles (Lovaas, 1987) should not be effective for diagnostic groups other than children with autism, we found only two reports of the use of EIBI with other non-autistic populations. Smith, Eikeseth, Klevstrand, & Lovaas (1997) provided EIBI for children with autistic features and severe intellectual disability. A group of 11 children receiving EIBI was compared to a group of 10 children receiving minimal treatment. Groups were similar on all measures at intake, with a mean age at intake of 3.08 years and mean intake IQ of about 28. Mean IQ gain for the EIBI group after a minimum of 24 months of intervention was 8 points while the comparison group on average lost 3 IQ points over the same period. Furthermore, children in the EIBI group acquired more expressive speech than children in the comparison group. The authors concluded that EIBI resulted in clinically meaningful gains, although the children remained significantly developmentally delayed. Smith, Klevstrand, & Lovaas (1995) reported data on results of EIBI for three girls between the ages of 31 and 37 months at intake who were, during the course of intervention, discovered to have Rett syndrome rather than autism. Results from this study failed to show positive outcomes in terms of potential intervention gains

Results of EIBI for children with autism have considerable promise for young children with intellectual disability for at least two reasons. First, the main evidence for EIBI outcomes relates to changes in IQ and adaptive skills and these are the defining features (i.e., limitations) in intellectual disability. Second, behavioural methods have been shown to be successful intervention approaches for a wide range of difficulties in children with intellectual disability including reducing challenging behaviour (Gardner & Griffiths, 2004), teaching communication skills (Buckhalt, Baird, & Reilly, 2004), toileting (Lancioni, O'Reilly, & Basili, 2004), and social skills (Carter & Hughes, 2007). Despite such potential, we could find no studies reporting the effects of intensive behavioural intervention methods (cf. Lovaas, 1987) with children with intellectual disability who do not have autism.

Other interventions for children with intellectual disability have resulted in positive effects on early language and communication skills. Prelinguistic milieu teaching (PMT) has been shown to have positive effects on both receptive and expressive language (Fey et al., 2006; Yoder & Warren, 2001) as have other naturalistic language interventions (Kaiser & Trent, 2007). These types of interventions are less intensive than EIBI: typically 3 to 4 weekly sessions of 20 minutes are provided. One will often recognize techniques based directly on applied behaviour analysis in these interventions (such as prompting and reinforcement). Researchers do not typically use global outcome measures to test these naturalistic language interventions; rather more direct measures of the targeted skills are used such as various assessments of language and communication.

Some centres in Norway that were originally established to serve children with autism have also enrolled children with intellectual disability and provided intervention for them in line with Norwegian intensive behavioural intervention service models for children with autism (Eikeseth, et al., 2002; Eikeseth, et al., 2007). However, although the service model used with non-autistic children with intellectual disabilities was similar, due to fewer financial resources allocated for working with this population the intervention programmes have not been as intensive in terms of the number of weekly intervention hours as those typically provided for children with autism. The present research offered a unique opportunity to evaluate behavioural intervention methods with young children with intellectual disability in comparison to

an eclectic "Treatment as Usual" group of children. The intervention model tested was that associated with an evidence base in the field of autism (Lovaas, 1987) but delivered in a less intensive format.

Method

Participants

We examined records of children admitted to three treatment centres in Norway between 2000 and 2008. Centres in Akershus, Oslo, and Vestfold counties (total population of approximately 1.3 million) are part of the national health and education system, and their services are available to all residents determined to have a need for them. All children who met the following criteria were included in the study: (a) intellectual functioning within the intellectual disability range (mental retardation) according to ICD-10 criteria (World Health Organization, 1993); (b) between 2 and 6 years of age at the start of intervention; (c) no medical conditions that could interfere with intervention such as sustained uncontrollable epilepsy, and major motor or sensory impairments; (d) no record of a diagnosis of autism, and (e) had received assessments of intellectual functioning and adaptive behaviour conducted at intake and again after at least 4 months and no longer than 24 months of intervention. These inclusion criteria resulted in a group of 25 children.

We conducted a thorough examination of each child's clinical record to establish intervention group status: a behavioural intervention group and an eclectic intervention comparison group. Children who had received intervention primarily based on applied behaviour analysis (ABA) for at least 8 hours a week consistently, constituted the behavioural intervention group; the remaining children, who had all received a mix of interventions types, comprised the eclectic intervention group. A majority of the children in the eclectic intervention group did receive some ABA intervention. However, the focus of this work was on specific skills and not as a comprehensive curriculum (cf. Lovaas, 1987). Thus, the ABA input in this group was for a maximum of 2–5 hours per week and only for a short period typically 3–4 months. In the Centres that were a part of this study, the decision about the type of intervention the child was to receive was made through a process involving the local

educational authorities, specialist services such as the Child Habilitation Services and/or the Child- and Adolescent Psychiatric Services, and the child's parents. Thus, assignment of intervention type was made independent of the present research.

The local educational authorities are required to make a written individual intervention plan for any child receiving special education services. The individual plan has to be evaluated and revised at least every 6 months. All children in the present study were subject to this routine. The behavioural intervention group consisted of 11 children (9 boys) (mean age 4 years and 6 months, range 34–67 months) and the eclectic comparison group of 14 children (10 boys) (mean age 3 years and 10 months, range 24–66 months). Sixteen children were recruited from Akershus (4 behavioural intervention), eight children from Oslo (6 behavioural intervention), and one child from Vestfold (who received behavioural intervention). Only three of the children had a known medical history that could be related to their intellectual disability, one child in the eclectic comparison group was diagnosed with Fragile-X syndrome, and two children in the behavioural intervention group were born prematurely.

The mean intensity of intervention in the EBI group was 10.3 (SD 1.8, range 8–15 hours) hours per week spent on weekly goals, and the mean duration was 12.1 months (SD 4.1, range 5–22 months). Although we do not have specific measures of intervention intensity in the eclectic intervention group, individual records indicate that resources allotted to the children were about the same across the two groups. The mean duration of treatment in this group was 14.8 months (SD 3.5, range 12–21 months).

Child Measures

Intelligence. The Bayley Scales of Infant Development (BSID-II) (Bayley, 1993) were used for the youngest children or the children that scored below the basal on intelligence tests standardized for their chronological age. The BSID-II is a measure of mental development for children up to 42 months. It yields a mental developmental index (MDI), which for the purposes of this study was considered equivalent to an IQ score. If the child scored below the norms, a ratio score was computed by dividing the obtained mental age with chronological age and multiplying by 100. For the older

and higher functioning children the Stanford-Binet Intelligence Scale: Fourth Edition (Thorndike, et al., 1986) or the Wechsler Preschool and Primary Scale Intelligence-Revised (Wechsler, 1989) were used.

Adaptive behaviour. The survey form of the Vineland Adaptive Behavior Scales [either the VABS or the revised VABS-II; (Sparrow, et al., 1984; Sparrow, et al., 2005)] were the only measure used for adaptive behaviours. This interview yields standardized scores for communication, daily living skills, socialization, and, for children less than six years old, motor skills. It also gives a total adaptive behaviour composite score (ABC). Standardized scores are given in much the same way as intelligence tests, with a mean of 100 and a standard deviation of 15.

Setting

All participants attended regular mainstream pre-schools for typical children in the local community. As required by Norwegian regulations, pre-schools are divided into units with a maximum of 18 children between 3 and 6 years of age, staffed with 3 adults (6 to 1 staff to child ratio). If the unit enrolls a child with special needs, it receives additional staff depending on the child's needs. The units involved in the present study received between 5 and 20 hours a week of extra staff time. The pre-schools are typically open weekdays from 7:30 am to 5:00 pm. All pre-schools had a separate treatment room or area that was made available where staff could work with the children without being disturbed. When not working in this room, the child was mainstreamed in the unit. Depending on the needs of the particular child, assistance was given for daily practical skills, such as eating, getting dressed to go outside, and toileting. If required, the child was helped to partake in play and general activities such as art projects and assemblies. However, unless staff was specifically instructed to do so, little or no systematic training towards educational goals was provided while the child was mainstreamed in the unit. There was one instance where two participants in this study were enrolled in the same pre-school unit (two boys in the behavioural intervention group).

Interventions

Behavioural intervention. The intervention model was similar to that described as clinic based behavioural intervention (Smith, et al., 2001). The model has been adopted for Norway community implementation for children with autism and has been described in detail elsewhere (Eikeseth, et al., 2002). A team of one to three therapists was recruited from the pre-school staff to provide the daily intervention for the child. One of the three treatment centres offered all training and supervision for these staff. Staff training started with a three-day training workshop and continued thereafter with weekly consultations lasting two to four hours. Included in this was a one to two hour team meeting conducted in the pre-school for the individual child. The child, primary caregiver(s), and therapists attended both the workshops and the team meetings. All staff and parents were trained in an apprenticeship manner; the supervisor first explained and demonstrated how to do a program, then the program was role played and/or staff took turns practicing the program with the child while being coached by the supervisors and the other team members. Based on the child's progress during the preceding week, that child's program goals and treatment procedures were updated and expanded. One person on the team was assigned responsibility for scheduling and monitoring intervention hours, updating program sheets, and finding materials needed for the various programs. Parental participation was also an important part of the intervention. Typically, parents' homework was to give the child opportunity to practice mastered skills in the home and other community settings.

Most of the work on new and difficult skills was carried out one-to-one in the treatment room. During some teaching sessions, particularly those targeting social skills, typically developing children from the unit worked with the child in the treatment room. During other sessions, programs for enhancing group interaction could take place in the unit, and programs focusing on self-help skills such as dressing were implemented in the reception area or bathroom. For the purposes of this study, we counted only the hours spent toward the specified weekly goals, whether inside or outside of the treatment room, as intervention time; time spent on practical assistance in daily life was not considered part of the intervention.

The supervisors had a minimum of one year's experience with implementation and supervision of behavioural intervention programs with children with autism, and possessed qualifications similar to those specified by Smith, et al., (2001). They met weekly with a consultant (one of the first three authors of this study), all of whom were licensed psychologists with at least 10 years experience of implementing intensive behavioural intervention. The first and second authors were collaborating researchers in the Multi-Site Young Autism Project, and the third author was a centre director and had completed a nine-month internship at the UCLA Clinic for the Behavioural Treatment of Children.

The intervention described here was based on widely used and validated treatment manuals (Leaf & McEachin, 1999; Lovaas, 1981, 2003; Maurice, Green, & Foxx, 2001; Maurice, et al., 1996). Typically the programmes progressed as follows: first work was conducted to establish basic skills, such as, responding to simple requests, imitation of gross motor behaviours, matching of objects, teaching of construction play (such as completion of puzzles or putting shapes in a shape sorter), making basic needs and wants known, and using sounds or words. Once these skills were mastered, we moved onto more advanced skills, such as, imitation of fine motor and oral motor behaviours, imitation of sounds and words, and labelling objects and actions. Following this we taught the children more abstract concepts, such as, colour, size, adjectives, and prepositions. Subsequent treatment goals included academic skills such as numeracy and literacy, discriminating wh-questions, and conversing and making friends with peers. From the beginning we targeted play and social skills, progressing from functional toy play and parallel play to symbolic play and cooperative play.

All procedures were based on operant conditioning techniques, including but not limited to, differential reinforcement, shaping, chaining, and prompt fading. In the early stages of intervention we taught most skills using the discrete trial format, but as the children progressed to more complex skills we included more naturalistic delivery such as incidental teaching and teaching in small groups of peers. The ultimate goal of the intervention was to help the child establish skills that would enable them to benefit from learning in various natural settings in a similar manner to the typically developing child. The major difference between the intervention provided

in the present study and that typically provided to children with autism is that the comprehensiveness and number of weekly treatment goals had to be adjusted to fit with the fewer available hours.

Eclectic intervention. For children in this group, elements from various types of interventions were combined in an attempt to best meet the child's educational needs. The intervention typically included a mix of the following intervention types: alternative communication, applied behaviour analysis, total communication, sensory motor therapies, programs based on the principles from TEACCH, as well as other methods that were incorporated based on the personal experience of the particular special education teacher and staff. As in the behavioural intervention group, the same therapists who provided teaching also served as assistants in the mainstream unit. Also, similar to the arrangements in the behavioural intervention group, the majority of the time in the mainstream unit was not used to implement standardized instructional programs or on systematic work towards educational goals. The organization of supervision and staff training for the eclectic programs was in some ways similar to that of the behavioural group. A special education teacher from the local educational authorities (12 children) or a supervisor from one of the three centres that provided intensive behavioural intervention (2 children) would do one or two weekly consultations totalling about 2 to 5 hours a week. The agency to be responsible for supervision and training was determined in the interdisciplinary educational planning process for each child. Within each agency a particular person was assigned on the basis of capacity and/or geographical location. As in the behavioural group, 1 to 3 therapists were recruited from the pre-school staff to do the daily work with the child.

The intervention components typically found in the eclectic group can be summarized as follows. ABA would typically include working on a small number of selected programs from ABA treatment manuals such as matching, imitation, or toilet training. Alternative communication would typically include working towards a symbol or sign based communication system. Symbols were typically line drawings, Bliss symbols or photographs taken of objects, persons, or activities from the child daily life. Signs would typically be hand signs for expressing needs such as "Food," "Water," or "Toilet". Sensory integration would typically involve daily activity sessions of 15 to 20

minutes of going on a swing, rocking and stretching while listening to music, or getting a massage. Total Communication elements would include strategies for the complementary use of signs, symbols and speech to enhance verbal comprehension, improve expressive language and develop a form of literacy. It would focus on broadening the medium of communication to include signs, symbols, pictures, photographs and objects, as well as speech. It might also involve the use of drama, mime, or other forms of visual communication. Elements taken from TEACCH would typically involve making length and content of sessions predictable by using baskets to separate the tasks and assigning areas for specific activities. The intervention elements that were reported to be based on the teacher's clinical experience would typically involve the use of worksheets, learning through educational software on a computer, and training social skills through listening to stories and looking at picture sequences.

Unfortunately, we were not successful in measuring accurately the total time spent on intervention in the eclectic group, or measuring the proportion of time spent on the various intervention approaches that were implemented. Typically, sessions were conducted throughout the day taking advantage of opportunities that arose in daily life when the child was motivated. Also, intervention approaches were often combined in the same session, so that when applying principles derived from TEACCH, such as structuring daily activities with baskets, the staff simultaneously used techniques derived from ABA, such as reinforcement and prompting, and principles from total communication, such as combining visual and verbal modalities to promote spoken language.

Data Collection Procedure

The first and second authors (both licensed clinical psychologists) carried out 82 % of the test administrations and were not blind to intervention group status. Independent professionals, who were blind to the group status of the children, carried out the remaining 18 % of the administrations. The independent professionals were either licensed clinical psychologists or examiners with a master's degree in special education and a license to administer psychological tests. Two of the administrations were performed both by an independent professional and by one of the authors

within 3 months of each other. When two test results were available, a conservative measure of improvement was obtained by using the higher score at intake and the lower score in post-treatment tests. Agreement on the total scores was within ± 5 points in both of these cases.

Results

Data Analysis

To evaluate the effectiveness of behavioural intervention, we used ANCOVA models. Because the children were not randomly assigned to groups or actively matched, the intake score for the specific outcome measure was entered as a covariate in each analysis. ANCOVAs were run for IQ and adaptive behaviour, including all sub domains (except for motor skills) and the adaptive composite scores. We also computed an index of reliable change (RCI) for the dependent variables where there was a statistically significant group difference in the initial ANCOVA models. The RCI is computed using formulae that yield a value representing the amount of change required to be 95 % confident that results are not accounted for by the variability in scores in the sample and/or by measurement error (Jacobson & Truax, 1991). The formula for computing reliable change requires that one can determine the stability and distribution of the test scores (in this case IQ and adaptive behaviour scores). To ensure that this test was conservative, we used our relatively small sample to generate this information (following Remington, et al., 2007). We estimated stability of test scores from intake to post intervention by computing the correlation between these scores in the eclectic comparison group where no specific intervention model had been applied, and thus where stability might be better estimated than from groups receiving a specific active intervention. We calculated the standard deviation for test scores from the whole sample of 25 children using intake data.

Applying the formula reported in Jacobson and Truax (1991, p.14), we established the absolute change in scores required to achieve a reliable change index score of 1.96 (i.e., statistically significant at the 5 % level). The index for reliable change was computed to be 17.96 for IQ, 11.38 for the adaptive behaviour composite score, and 13.36 for the Communication sub domain on the VABS. Reasons for the considerably

lower indexes on the VABS were mainly explained by a lower standard deviation on intake scores.

A further advantage of establishing a dichotomous outcome variable for change, such as whether or not each child meets reliable change criteria, 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 (NNT) and Absolute Risk Reduction (ARR). The NNT represents number of children who would need to be treated with a specified intervention to obtain one additional successful outcome over the success rate in a comparison intervention. For example, $NNT = 4$ means that for every four children who are treated with intervention X, one additional child will respond to the intervention, who would not have responded to a comparison intervention. A result of $NNT = 1$ means that all children receiving an intervention succeed when they would not have done so following a comparison intervention. In other words, the lower the NNT, the more effective the treatment relative to the comparison (Kraemer, et al., 2003). ARR is computed in a similar way as NNT, but expressed as a measure of the difference in percentage response between two interventions (Pinson & Gray, 2003). When the ARR is used as a measure of intervention effectiveness, the results are usually given in negative outcome. This means that intervention effectiveness is measured in how much it reduces the risk of having a bad outcome. For example, if in treatment A: 30 % of patients do not respond to intervention, and in treatment B: 80 % do not respond to intervention, the ARR (also called risk difference) is 50 % in favour of intervention A. These statistics are particularly helpful as a simpler way to communicate information about the effects of interventions to policy makers.

After classifying children in early behavioural intervention and comparison groups, in terms of whether or not their intellectual functioning and adaptive skills changed to a reliable extent, we computed NNT and ARR. To do these calculations for the present sample, we used readily available free access online calculators (Straus, et al., 2004).

Main Findings

Means and standard deviations for both groups at pre- and post-intervention are displayed in Table 6.1. The ANCOVA model we used to analyze outcome after, on average, approximately one year of intervention showed that the behavioural intervention group made significantly larger gains on the main two outcome variables. For intelligence, the group difference was statistically significant, $F(1, 23) = 7.97, p = .010$, and also for the adaptive behaviour composite scores, $F(1, 22) = 8.25, p = .009$. For the sub domains on the VABS, Communication scores showed a statistically significant group effect $F(1, 22) = 6.50, p = .019$ but no other VABS domain led to a significant group effect. On average, the behavioural intervention group gained 16.6 IQ points, and 2.9 adaptive behaviour composite points. The eclectic comparison group gained 3.9 IQ points and lost 2.8 adaptive behaviour composite points. On the communication sub domain of the VABS the intervention group gained 2.7 standard points whereas the eclectic comparison group lost 3.2 points. These differences equate to effect size estimates (Cohen's d) of 1.13 for IQ, .95 for the adaptive behaviour composite, and .83 for the Communication sub domain (using the formula differences in mean change/pooled standard deviation). The proportions of children in the behavioural intervention group and the eclectic comparison group, achieving reliable change on IQ, adaptive behaviour composite, and the Communication sub domain are displayed in Figure 6.1. Each line on the graph in Figure 6.1 represents an individual child's change in test score. These have been sorted left to right from highest negative to highest positive change. A reference line on the y-axis shows the criterion for reliable change.

Overall, 7 out of 11 children (64 %) in the behavioural intervention group achieved reliable change in IQ, 2 out of 11 children (18 %) achieved reliable change in adaptive behaviour composite scores, and 1 out 11 (9 %) did so on the Communication sub domain. In the eclectic comparison group, 2 out of 14 children (14 %) achieved reliable change in IQ and none achieved reliable change on either adaptive behaviour score. The NNT was computed to be 2.0 (95 % CI [1.2, 6.4]) for achieving a reliable change in IQ. The NNT for reliable change in adaptive behaviour was 5.5 and for communication 11.0 For adaptive behaviour and communication the 95 % confidence interval for the absolute risk reduction extends from a negative

number, where treatment may harm (NNH), to a positive number, where treatment may benefit (NNT). As such it was hard to compute a 95 % CI. 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 these instances is that we can be 95 % certain that one of these statements is true. For adaptive behaviour, the experimental treatment is harmful (compared to control), and the NNH is greater than 21.7. The experimental treatment is helpful (compared to control), and the NNT is greater than 2.4 (expressed as NNT 2.4 to ∞ to NNH 21.7; adopted from Altman, 1998). For change in communication the NNH is greater than 12.7, and the NNT is greater than 3.8 (expressed as NNT 3.8 to ∞ to NNH 12.7). These numbers translate to absolute risk reductions of 49.4 % (95 % CI [15.5 to 83.2 %]), 19 % (95 % CI [-4.6 to 41.0 %]) and 9.1 % (95 % CI [-7.9 to 26.1 %]) respectively, all in favour of the behavioural intervention group. Reflecting the results for the NNT calculations, the confidence intervals of the ARR statistics for adaptive behaviour outcomes include zero and so these proportions meeting reliable change criteria on these measures do not represent a statistically significant difference in this small sample.

Discussion

Results of this study showed that children with intellectual disability who received behavioural intervention made statistically significant and clinically meaningful gains when compared to a group of children receiving eclectic "Treatment as usual" after on average approximately one year of intervention. These gains were apparent for IQ, adaptive behaviour composite scores, and communication standard scores. All of the group effect sizes were, by convention, considered large (Cohen, 1988). Objective measures of reliable change also showed significant advantages in favour of the behavioural intervention group for IQ, although adaptive behaviour differences on these criteria were not statistically different across the groups. Using the dichotomous outcomes of reliable change at the individual child level, especially for IQ outcomes, measures of potency focused on Numbers Needed to Treat and Absolute Risk Reduction compared favourably to those for medical and psychological interventions for many common disorders (Pinson & Gray, 2003).

Table 6.1:

Unadjusted means and SDs of scores at intake and after ca one year of intervention by group.

Measures	Behavioral group						Comparison group					
	Intake		After 1 year		Change		Intake		After 1 year		Change	
	M	SD (range)	M	SD (range)	M	SD	M	SD (range)	M	SD (range)	M	SD
Intellectual functioning*	56.4	12.7 (36-70)	73.0	12.2 (56-90)	16.6	10.2	49.5	10.0 (35-64)	53.4	16.5 (21-81)	3.9	12.1
Vineland Adaptive Behavior Scales												
Adaptive Behavior Composite**	62.7	7.8 (51-76)	65.8	7.3 (51-76)	2.9	6.2	59.3	6.9 (48-69)	56.2	7.7 (42-71)	-2.8	5.9
Communication*	68.3	10.6 (51-84)	71.4	9.2 (55-83)	2.7	6.9	63.8	10.8 (48-80)	60.1	11.5 (38-79)	-3.2	7.3
Daily Living	66.0	10.9 (55-85)	67.1	7.4 (57-79)	1.6	6.4	63.2	7.1 (48-76)	61.0	9.8 (47-77)	-1.6	9.3
Socialization	66.5	8.4 (57-82)	71.5	10.7 (53-86)	4.2	9.5	64.8	8.4 (54-82)	61.6	11.0 (42-83)	-2.7	10.6

*p < .05. **p < .01 on main effects. Behavioral group n=11 and comparison n=14, except on VABS scores n=13.

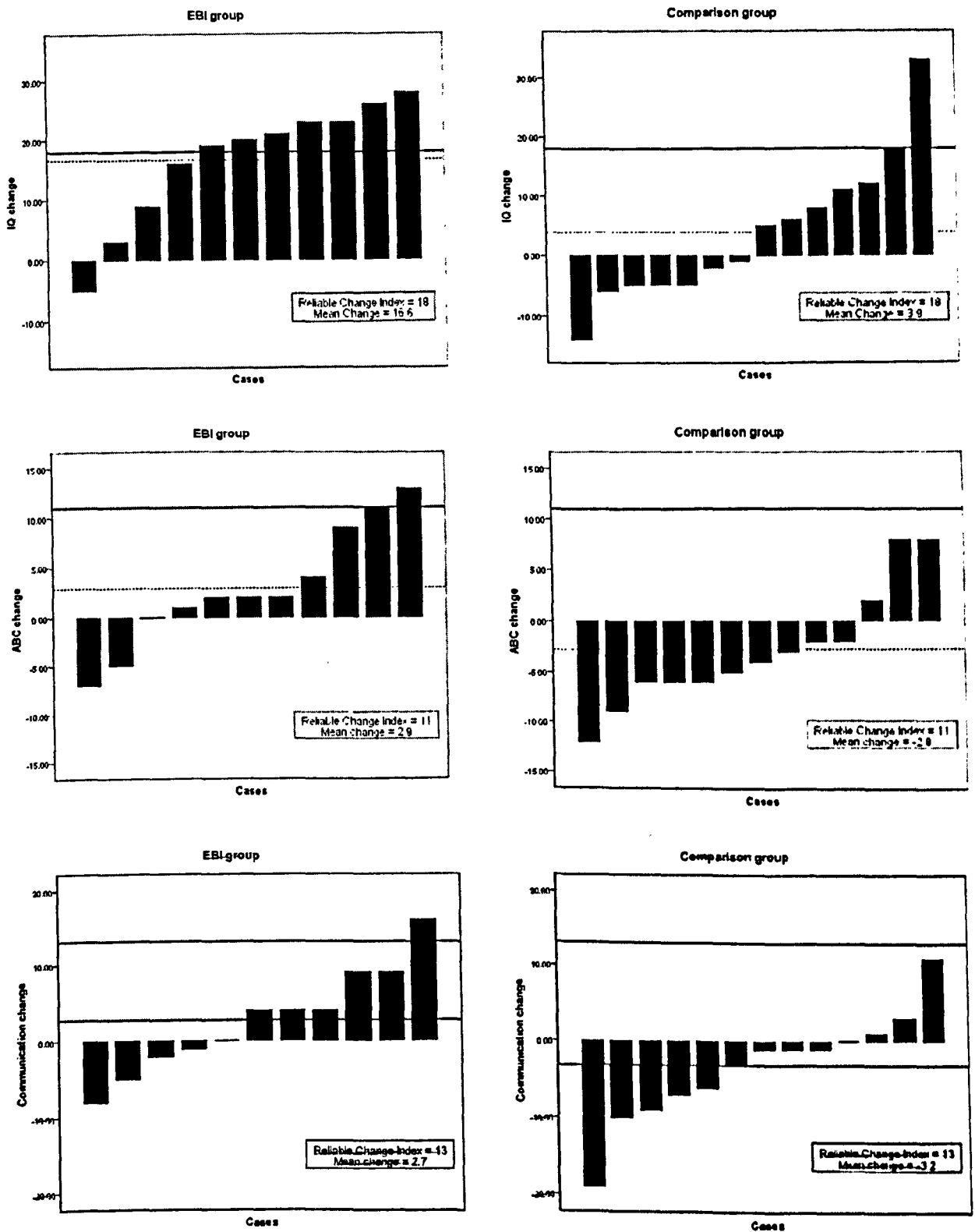


Figure 6.1. Bars indicate changes in IQ, ABC and Communication sub domain scores for individual children in the EBI and comparison groups. The solid lines represent the Reliable Change Index and the dotted line the average change in each group.

There is increasing empirical support for behavioural intervention in the field of autism. These developments, combined with the promising results from the present study, support further exploration of a behavioural intervention model for children with intellectual disability. The present data also suggest, in contrast to the evidence in the field of autism, that a relatively low intensity intervention may be beneficial to children with intellectual disability. In a study where children with autism were provided with a comparable number of weekly hours (mean 12.5 hours) to that in the present study (mean 10.3 hours), changes in IQ (8.2 points) and ABC (-0.2 points) were smaller (Eldevik, et al., 2006). In future research, intensity of intervention needs to be directly manipulated within the context of a randomized controlled trial of behavioural intervention for children with intellectual disability.

Drawing on the experience of working with children with intellectual disability using behavioural intervention methods, it is important to consider the reasons why children with intellectual disability may profit more from behavioural intervention than children with autism. We found that it is often much easier to find effective reinforcers and that many children will, unlike those with autism, respond well to natural social reinforcement. As a teacher, it can be easier to get and sustain the attention of children with intellectual disabilities because one is less likely to have to compete with stereotyped behaviours. Furthermore, children with intellectual disability are more likely to respond to more aspects of the teaching material, as has also been demonstrated in the laboratory (Lovaas, Schreibman, Koegel, & Rehm, 1971). Another facet that appeared to be easier when working with children with intellectual disabilities, and indeed often emerged without special programming, was generalization of skills. This was true both within programs (e.g., imitation) and across settings and persons. This may be explained by the other features mentioned above such as the affinity to more natural reinforcers and lower frequencies and severity of stereotyped behaviours and stimulus over selectivity.

The main contribution of the present research has been to extend the evidence for behavioural intervention typically used with children with autism to children with intellectual disability. Our findings clearly need to be replicated and extended before early (intensive or less intensive) behavioural intervention might become established as an intervention of choice for young children with intellectual disability. It is also

important to consider some limitations of the present research. First, this was a natural experiment and thus there was no random allocation and there were thus threats to internal validity. These included the fact that testers were not generally blind to intervention group status and the intervention groups were defined pragmatically to reflect practice rather than experimentally imposed. Second, we included only limited outcome measures. Third, it will be important in future research to also follow up children over time to establish whether gains from early intervention can be maintained. Finally, the majority of children's intake IQ scores in the present study would have placed them in the moderate to mild intellectual disability range. The potential of behavioural intervention of the kind used here to benefit children with more severe intellectual disability is worthy of further study. Nevertheless, an intervention that can be shown to have considerable impact for children with moderate to mild intellectual disability has the potential to benefit a large number of children.

Chapter 7: General Discussion

Aggregated Data Meta-analysis

In the preceding four chapters I have made an attempt to answer some of the questions relating to behavioural interventions for children with developmental disabilities or autism. In chapter 3, I conducted a refined and updated meta-analysis addressing some of the shortcomings in earlier reviews. Nine controlled studies were included following a systematic literature search. The results confirmed that IBI yields significantly better effects in terms of gains in intelligence (Hedges' g effects size = 1.10) and adaptive behaviours (Hedges' g effects size = .66) than other interventions and control groups. According to convention these effect sizes would be considered large and moderate respectively. Furthermore, I found that the present meta-analysis appears to be without significant publication bias, and included studies with homogeneous results. In the introduction I mentioned some of the most common criticism of meta-analysis. Some of this criticism is relevant to reviews in general (whether they are quantitative or narrative) and some may have particular relevance for the review presented in chapter 3. I will therefore go through these in more detail.

Can One Number Summarize Intervention Effects?

The first criticism often raised is that one number cannot accurately summarize the research in a field, since it only focuses on a combined effect and ignores that effects may vary from study to study. In the meta-analysis reported in chapter 3, tests of homogeneity were not significant (the Q -statistic and I^2). Hence, effect sizes appear to be comparable across studies. It should be noted however, that although the outcomes across studies are similar, there may be (and indeed there is) considerable variability in outcome for the individual participants within each study. This fact has been noted in earlier reviews (e.g., Howlin, et al., 2009) and in most of the published studies on behavioural intervention, including Lovaas' (1987) seminal report. This was also part of the reason why I wanted to pursue the individual participant data analysis in chapter 4.

Are we Compiling Different Interventions?

Second, when doing a meta-analysis one risks mixing apples and oranges (i.e., different interventions) and report these as the result of one intervention. In my reviews (both the AD and IPD meta-analysis) I applied the Green et al (2002) definition of behavioural intervention. I believe this definition better captures the common elements of what is actually provided, regardless of whether the authors happen to mention Lovaas, the UCLA model, EIBI, ABA, and so forth in their paper. After all, the intervention is based on the principles of applied behaviour analysis (and not Lovaas' principles or any other person), a point Lovaas has also been careful to make on several occasions (e.g., Lovaas, 2003). I might add that I think other reviewers have applied their definition of behavioural intervention somewhat inconsistently. One example of this is how the Howard et al (2005) study is treated, sometimes as supporting the UCLA/Lovaas model and sometimes not (Rogers & Vismara, 2008). Furthermore, the four studies that are the foundation for Rogers and Vismara classifying the UCLA/Lovaas as "well established" include partial replications, where the intervention is delivered in settings other than a university clinic (Eikeseth, et al 2002; Cohen, et al 2006; Howard et al., 2005), and with fewer hours (Smith et al, 2000). The intervention provided fell within the definition of IBI, but these reports indicate that the UCLA model of service delivery is not the only effective approach. This is further corroborated with the evaluation of the IBI model I describe in chapter 5. The utility and real world effectiveness of IBI may be more dependent on the actual behavioural analytical procedures being implemented, than exactly how staffing and other logistics are dealt with.

The Spreckley and Boyd (2009) study provides an example of what may happen when different interventions are mixed. They confused the parent managed behavioural intervention group in the Sallows and Graupner (2005) study with standard care. Since the outcomes reported were similar between parent managed and clinic directed behavioural intervention, Spreckley and Boyd went on to conclude that behavioural intervention was no better than standard care. Thus, how intervention is defined and how studies are categorized can have a large impact on the conclusions drawn even from the same set of reviewed studies.

As has been pointed out by other reviewers the various behavioural intervention teams are likely to weight different elements of the behavioural intervention programs differently. In addition, the way intervention has been carried out may have varied over the years within the same clinics as more research has become available documenting more efficient procedures and it may have varied because of individual children's needs and skill profile. Furthermore, variation may be due to the research teams emphasizing particular procedures, for instance some may rely heavily on discrete trials teaching whereas others have relied more extensively on incidental teaching; some may introduce augmentative communication such as PECS (Bondy & Frost, 2003) earlier than others for children with no spoken language, some may instead introduce sign-language, some may employ procedures based on Skinner's terminology of verbal behaviour (Sundberg & Partington, 1998), emphasizing mand-training heavily in the early stages of the intervention and still others may incorporate procedures from pivotal response training (Koegel, Koegel, Harrower, & Carter, 1999; Koegel, Koegel, Shoshan, & McNeerney, 1999). Unfortunately, I did not find the information in the studies included in the present chapters detailed enough to code them on these variables. Only a very crude coding would perhaps be feasible, leading to problems with validity and perhaps also misleading results. I have, however, communicated with most of the researchers in the process of obtaining the individual participant data. There seems to be a general consensus on the definition of IBI I applied in the present analyses, and also that the intervention provided fell well within the limits of the definition.

A related issue may be how and why I decided to divide the remaining children into a control group and a comparison group. The difference between comparison and control groups may seem trivial and confusing. Because these groups were combined for some of the analyses, why are they not combined throughout the reviews? And, why did I choose also to include children receiving low-intensity behavioural intervention? I think that the distinction between comparison and control groups is worthwhile, because it may be possible then to analyse more directly the effects of intensity per se. One of the criteria for including a child in the comparison group rather than the control group was documentation in the original paper that the child had received an intervention of similar intensity as the IBI group. Thus permitting a more valid analysis of whether or not intensity in itself is the active

ingredient or if it is the actual intervention methods that make the difference. The reason for combining the comparison and control groups in the aggregated data meta-analysis effect size calculations is that the initial analyses showed that the outcome in these two groups was similar. This was also very helpful in that I could provide a more pragmatic summary of the effect sizes using the NNT and ARR methods. In order to get variability in possible predictors of outcome, I chose to include all IBI groups with acceptable documentation of the weekly intervention hours, even the ones that had provided low-intensity intervention. Most likely, the results (NNT and ARR) would be more favourable for behavioural intervention if I had made a cut-off for inclusion of children at 30 hours a week, or 35 hours a week. Hence the current analyses yielded conservative results.

Is the Quality of the Studies Good Enough?

The third point of concern when doing a meta-analysis is that low-quality studies may be included in the analysis. As has been pointed out in earlier reviews (e.g., Howlin, et al., 2009; Eikeseth, 2009), there are a host of issues relating to the general quality of the studies in the IBI literature.

Crude measures of intervention intensity. There is an issue with accurately measuring intervention intensity (Rogers & Vismara, 2008). The decision to include weekly intervention hours as a predictor in my regression analysis may be questionable, given the large amount of data that are missing, available only in the form of a retrospective estimate or as averages for all children in a particular study. I have acknowledged that weekly hours is a crude estimate of intervention intensity, but a good and widely applied quantification of intensity does not exist at present. Thus, I dichotomized intensity using the median split so I could include more subjects than the approximately 25 % I had accurate individual data on. This would hopefully yield a more valid analysis on how this variable is related to outcome. This procedure, I believe, provided a reasonable balance between the crudeness of the data and what I could draw out from them with some level of certainty. Even if we got significant differences between the high intensity group and the low intensity group using this procedure, we may also get significant results if we set the cut off point

based on recommendations in the literature at lets say 30 hours a week (Green, 1996). There are other issues however pertaining to intensity of intervention. It is conceivable that the children who do very well will be provided fewer hours of intervention (McEachin, 2006), simply because it is not necessary to have a high intensity program. Along the same lines, it is unlikely that all children will need 40 hours a week of intervention. This is supported by studies providing fewer hours which do indeed have some children doing very well on all outcome measures (Eikeseth, et al., 2002; Smith, et al., 2000), although the average outcomes for all children in these studies are not as good as the higher intensity studies (Howard, et al., 2005; Sallows & Graupner, 2005). We are still far from making specific prescriptions on intensity for individual children, but at group level it appears that intensity is the most important variable related to outcome. As important as this variable may be, there is very little literature on this topic, and with one possible exception (Lovaas, 1987) there is no study in which intensity is treated as the independent variable with all other intervention variables kept constant (Warren, et al., 2007). This can be contrasted with drug research where research is almost always conducted on its effects at various doses (Piantadosi, 1997). This is done to determine side-effects but also to estimate effects of different doses of the drug. But before intensity could be manipulated in any trial on early intervention a consistent approach to defining and measuring it is needed.

Crude measures of duration. Also relevant here is the often inaccurate reporting and large variability in the time between intake and post intervention measures. Although, in our review we included studies that had an intervention period between one and three years, the post intervention assessment was in some cases done later. For instance, in the Smith, et al., (2000) randomized control study, follow-up testing was conducted when the children were 7–8 years, with an average time between intake assessments and follow-up of 54 months. Similarly, post measures were taken 3–4 years after intervention had stopped for some of the children in the Smith, et al., (1997) study. The impact of this is unknown, and clearly it would be an advantage to assess children following a more standardized time protocol.

Choice of outcome measures. Another issue in the IBI literature is the problem of outcome measurement (Howlin, et al., 2009). The most commonly reported outcome

measures are intelligence and adaptive behaviours. But how valid are these when evaluating intervention for autism spectrum disorders? Why not use direct measures of autistic behaviours and severity? Indeed almost all of the published studies have some measure of this included in their evaluation. The problem these raise is that different and often non-standardized measures have been employed (e.g. custom made checklist measuring pathology, as in Lovaas, 1987). However, this has improved with the new wave of publications over the last 5 years; with the more recent studies often including more direct measures of autistic behaviour such as the ADI-R (Sallows & Graupner, 2005), the ADOS (Ben-Itzhak & Zachor, 2009), the Autism Screening Questionnaire (Remington, et al., 2007), and measures of joint attention skills (Remington, et al., 2007). Also, longer term outcome has been measured using standard personality tests (Eikeseth, et al., 2007; McEachin, et al., 1993). Although, the IBI groups tend to do better than comparisons or controls on these measures, the results are much less robust. Problems still remain with respect to validity of these tests, because they were not developed to gauge changes and improvement. Instead they were developed primarily as diagnostic instruments. Another measure of outcome that some studies report is educational placement (e.g., Eikeseth, et al., 2007; Sallows & Graupner, 2005). However, this is not necessarily indicative of child progress as it may be as much a result of local policies and parental preference. Recent studies have also included measures of other variables that may influence outcome such as socioeconomic status, and measures of parental well-being (stress, anxiety, depression etc - see for example Remington, et al., (2007).

Weak research designs. In my analysis I only had one study with random assignment (Smith, et al., 2000). The outcome of this study does not seem to differ in any way from the other studies included (if anything outcome seems to be above average). However, lack of random assignment remains a problem, and one that is not likely to be resolved easily. Bearing in mind the evidence base supporting IBI it may now be regarded as ethically problematic to conduct random assignment studies. At the same time random assignment cannot be a goal in itself, as this procedure also has shortcomings (Borenstein, et al., 2009), in particular when applied to small samples like those typically found in the IBI literature. Hence, it is argued by many that other assignment procedures may be as appropriate and

convincing (Baer, 1993). For example, there is no inherent bias in assigning children based on therapist availability, and indeed most of the assignment procedures used in the IBI literature have resulted in groups being similar at intake on most of the variables included (Cohen, et al., 2006; Eikeseth et al., 2002; Eldevik, et al., 2006; Howard, et al., 2005; Remington, et al., 2007). However, the lack of random assignment is perhaps the most commonly raised criticism of IBI studies and the main reason for why IBI is not achieving the highest scores on levels of evidence. Future intervention studies must make attempts to work around this, for instance by having a waiting list control group that receive intervention after 9-12 months.

Are All Relevant Studies Included?

Finally, the last common criticism of meta-analysis that needs to be mentioned here is the concern that important studies may not be included or are ignored. In my analysis, I have outlined the selection procedure with flowcharts as well as detailing it in the text. As a precautionary measure I also added a reliability check on a subset of the papers retrieved in the initial search. Still, one could argue that I should have included more studies. I excluded seven studies for instance because their outcome data were not according to my criteria in that they relied heavily or solely on performance measures (e.g., Magiati, et al., 2007; Sheinkopf & Siegel, 1998) and I excluded 5 studies because the duration of the intervention was less than a year (e.g., Reed, et al., 2007b; Stahmer & Ingersoll, 2004). Some of these studies report much more moderate effects and one study suggested only very small or potentially negative changes in IQ following behavioural intervention (Magiati, et al., 2007). Other studies indeed show that outcome is on average more modest when duration is shorter (Reed, et al., 2007b) and supervision is given on a less frequent basis (e.g., Bibby, et al., 2002). Hence, outcome of behavioural intervention may not be perceived as positively if other inclusion criteria are applied.

Individual Participant Meta-analysis

In chapter 4, individual participant data from published outcome studies on IBI were analysed. After a systematic literature search and obtaining data directly from

researchers, 16 studies with a total of 453 children were included in this analysis (including all studies from chapter 3). This type of analysis may be considered a next step for synthesising research data and a scientifically stronger approach than the traditional meta-analysis, which is based on aggregated data. Outcomes were analysed applying state of the art methodology developed within the framework of evidence based practices. First, the individual children were divided into three groups: those that had received behavioural intervention, those that had received another intervention of similar intensity or to a control group where no specific intervention was provided. Second, indices for reliable change were computed for intelligence and adaptive behaviours. These were used as cut-offs (≥ 27 points gain in IQ scores and ≥ 21 points gain in adaptive behaviour scores) for dichotomizing outcome—either meeting these criteria or not meeting them. Based on the number of children meeting the criteria in each of the groups I was able to compute the number needed to treat (NNT) and absolute risk reduction (ARR) following IBI treatment. The outcome for the behavioural intervention group were significantly better than those for the control and comparison groups and were even demonstrated to compare well with numbers reported in mainstream healthcare (e.g., interventions for common medical and psychological disorders). Approximately 30 % in the behavioural intervention group achieved reliable change in IQ, vs. 3 % and 9 % in the comparison and control groups. For adaptive behaviour, approximately 21 % met the criterion in the behavioural intervention group vs. 6 % in the comparison group and 5 % in the control group. These proportions equated to a NNT of 5 for IQ and 7 for adaptive behaviour, and an ARR of 23 % and 16 % respectively. I then used multiple regression analyses to see if any of the intake variables often suggested to be related to outcome (or interactions among these) predicted outcome in the behavioural intervention group. I found that intensity of intervention was related to both IQ and ABC gains, and that IQ and ABC scores at intake were related to gains in ABC scores. However, none of the interaction terms were significant. Finally, I propose that these outcome data may serve as benchmarks for evaluating outcomes for behavioural interventions for children with autism in a variety of contexts.

Intervention in Other Settings and with Other Diagnostic Groups

In chapter 5, another question was addressed. Could behavioural intervention for children with autism be implemented in a mainstream community setting? In this chapter I described a model where behavioural intervention was implemented in mainstream pre-schools, utilizing the existing community resources. All 25 children that had received services through this model were included in the evaluation. Outcomes on intelligence and adaptive behaviour after 1 and 2 years of behavioural intervention were compared against those reported in the literature, and the benchmarks proposed in chapter 4. Children were provided an average of only about 14 hours a week of intervention. The average IQ gain was about 11 points after one year and about 20 points after two years. For adaptive behaviours the gain was about 1 point after one year and about 9 points after two years. Paired samples t-tests conducted at each time point show that only change in IQ was statistically significant after the first year, whereas all outcome variables had significantly improved after the second year of intervention. The mean change Hedges' *g* effect size was 1.05 for IQ and 0.99 for adaptive behaviours after two years of intervention. Twenty eight per cent of children met the reliable change criterion proposed in chapter 4 for IQ and about 11 % met the criterion for adaptive behaviour change. The overall results appear better than what would be expected given the relatively low intensity of intervention. The Hedges' *g* effect size for IQ is indeed higher than that reported in a recent meta-analytic review (0.69) also employing the mean change effect size (Reichow & Wolery, 2009). NNT and ARR could not be computed since the study did not have a control or comparison group. Although there were problems associated with implementation in this setting, in particular with achieving the recommended intensity of intervention, outcome appears to be acceptable at least in terms of gains in intelligence scores

In chapter 6, I address the issue of whether or not behavioural intervention could be suitable for populations other than children with autism spectrum disorders. A group of 11 children with intellectual disability was provided with behavioural intervention and compared to a group of 14 children that was provided with standard special education. The behavioural intervention was similar to that typically provided for children with autism, except that it was less intensive, only about 10 hours a week,

due to fewer resources generally available for this population. Outcome appears promising, with children in the behavioural intervention group doing significantly better than the comparison group on IQ and ABC measures after one year of intervention. Average gain in IQ scores for the behavioural intervention group was about 17 points, whereas average gain in adaptive behaviour scores was more modest at about 3 points. Still, the average effect sizes were large at 1.13 for IQ and .95 for adaptive behaviour composite scores.

Further, I employed a similar procedure to that in chapter 4 for evaluating clinical significance. First a reliable change index was computed (in this population resulting in a less stringent criterion at ≥ 18 point gain in IQ scores and ≥ 11 point gain in adaptive behaviour scores), then the proportion of children meeting these criteria in each group was calculated. About 64 % in the behavioural intervention group met this criterion for change in IQ whereas only about 14 % in the comparison group did so. The numbers for change in adaptive behaviour were 18 % and 0 % respectively. This translates to a NNT of about 2 for intelligence and about 6 for adaptive behaviours, and an absolute risk reduction of about 50 % and 20 %, respectively.

General Implications

Based on the findings from both the reviews in chapter 3 and 4 and the applications described in chapter 5 and 6 I will discuss some possible implications and areas of future research.

How Can Intervention Methods be Further Improved?

Although the results show that effects of IBI are clinically meaningful, there are still a number of issues with respect to the intervention itself that need further research. Why the large variability in outcome? Is this due to short-comings in the intervention itself? Is it due to variability in how the intervention is implemented or is it due to the biological and genetic make up of the individual children? Based on what research from developmental psychology tells us (e.g. Novak & Pelaez, 2004) a combination of these factors is the most likely answer. How much each of these factors contributes is an impossible question to answer at present. However, the only

variable we can do something about at present is the actual intervention methods. Could we be better at adjusting these to the individual child? Could we develop new and improved methods to better help more children? Is it a realistic goal to achieve clinically meaningful change on intelligence, adaptive behaviour and other measures (such as measures of autistic behaviours) for all children, even if we could provide intervention of the optimal quality and intensity?

I think the answer to all of these questions is yes. If behaviour analysis continues to accumulate research findings and improve on and invent new methods, we should also see a corresponding improvement in general outcome. We may also have to expand our research areas to achieve this. For instance, one area that appears to have some promise is intervention with even younger children, below two years of age. Is it possible to intervene earlier and put them on better developmental trajectories through conditioning social reinforcers such as touch, proximity and smiling? At least in theory affinity to such reinforcers could improve interaction with care-givers and thus have disproportionate positive effects on development.

Search for Moderators and Mediators of Intervention Outcome. This is an important but difficult undertaking. A mediator is a variable that is influenced by the intervention, and in turn influences outcome (Smith, et al., 2007). It can thus be thought of as the mechanism by which the intervention works. In behavioural intervention social initiations may mediate outcome. For instance, in one study, children with low-rates of social initiation were taught through pivotal response training to initiate social interaction. When these children were assessed at age 8–10 years they had high-rates of social initiation and near-normal functioning in all other areas assessed (Koegel, Koegel, Shoshan, et al., 1999). The authors considered social initiation most important for achieving gains in other areas such as communication, friendship and participation in various activities outside school. A moderator variable is a factor that can change the impact of the intervention. It may specify for whom and under what circumstances an intervention is likely to yield the best outcome.

The search for mediators and moderators of intervention outcome requires a very large sample and sophisticated statistical techniques (Kraemer, Wilson, Fairburn, &

Agras, 2002), and random assignment to groups is essential. Hence, I did not attempt to conduct such an analysis in the present thesis (instead I analysed correlates of outcome applying multiple regression, because all the assumptions required for such an analysis seem to hold). No study has been adequately powered or designed to properly address the issue of moderating variables. Applying less rigorous methodology (usually correlations), intensity of intervention (Granpeesheh, et al., 2009; Lovaas, 1987), intake age and IQ (Ben-Itzhak & Zachor, 2007; Harris & Handleman, 2000) biological variables (in particular dysmorphology) (Stoelb, et al., 2004) and frequency of supervision (Eikeseth, Hayward, Gail, Gitlesen, & Eldevik, 2009) have all been found to relate to outcome. Identification of mediator and moderator variables may help prescribe intervention type to individual children and thus help save precious resources and reduce family despair. This remains an important area for future research

How Can We Improve The Implementation of EIBI?

Another important area of future research is how we can improve the quality of the implementation. Although the intervention is to a large extent manualised very few studies measure the extent to which the manuals are actually followed or employ other direct measures of overall quality of intervention. Research on many of these questions is underway. For instance, large scale implementation of IBI in Canada prompted the researchers to develop protocols for staff training and performance, as well as other measures of implementation (Perry, Penn Flanagan, Prichard, 2008). Also, some studies have specific staff-training protocols and levels of training (Cohen, et al. 2006; Eikeseth, et al., 2002; Smith, et al., 2000), but in most cases the information on fidelity is anecdotal. Thus, we still need to develop better and more standardized ways of measuring variables such as intensity, order of specific curriculum programs, procedural fidelity and weighting of different approaches (incidental teaching, discrete trials, etc).

Related to this are questions like how long the intervention should go on for? Should intervention continue into school age or even into high-school age? What sort of criteria should we have for terminating or continuing the intervention? And also we need to know more about how much room that can be allowed for local adoptions? In

some cases it may be difficult to get parents involved in the intervention. For instance children from families that have emigrated from third-world countries may be referred for intervention. In such cases the parents may not have heard about autism and developmental problems may be associated with shame or a belief that it will go away as the child grows older. These challenges were seen in particular in the study described in chapter 5. We need to know more about what we should do in such cases; both in terms of how to involve the parents and also how much the parents need to be involved. One of the lessons Lovaas learned (Lovaas et al., 1973) is that outcome is improved if parents are involved in the intervention. But this does not necessarily mean that intervention needs to take place at home. If intervention is provided in school parents could still get the same training and still come in to work with their child, and also still get “homework” to do every week.

In terms of local adaptations numerous questions arise such as; what sort of background the intervention team need to have, how often the intervention needs to be supervised and whether it can be supervised using recent technological advancements such as videoconferencing? Ideally, I think we need to develop the general professionalism of the services. This entails establishing more schools and centres that can provide IBI. There also needs to be a formal education and licensing of therapists, and a clearer route for how to make a career. This is underway with international board certification of behaviour analysts (Shook, 2005) and a move to expand this to specialities for behavioural intervention for autism.

Closing the Practice Gap. In most fields there is a considerable gap between knowledge and practice. This gap is well documented and is by no means limited to early intervention for children with autism spectrum disorders. It is persistent throughout psychology (Frazier, Formoso, Birman, & Atkins, 2008), education (Nuthall, 2004) and medicine (Gautam & Gautam, 2008). The reasons for this are many and beyond the scope of this thesis to explore in detail. The question is however if it is harder for behavioural interventions to get a foothold than other evidence based interventions? And if so, why might this be the case? Is the evidence not convincing enough? Are there ideological or philosophical reasons? Are there financial reasons? Is applied behaviour analysis being misrepresented or misunderstood? The building block of behaviour analysis, namely that behaviour can be a study matter in itself (Skinner, 1953) is still controversial, and attempts (not only

by behaviour analysts) to evaluate interventions applying various criteria for evidence based practices have met considerable resistance within the field of psychology and education (e.g., Poplin, 1988). Within the field of intervention for autism there is an increasing array of evidence based methods (National Autism Centre, 2009). It appears however that it is hard to compete with more politically correct interventions that often focus on such entities as child empowerment and self-esteem. There are a number of vocal opponents claiming that evidence based methods are inhumane, perfunctory, ill conceived, and dangerous to the psychological health and educational well-being of children (Maurice & Taylor, 2005; Morris, 2009). Also, it is claimed that it would be impossible and reductionist to quantitatively measure the outcome of intervention let alone to specify techniques and procedures used.

When evidence-based methods are compared against the multitude of interventions that claim to empower the child and increase self-esteem (whilst addressing the whole child and so forth), many parents and professionals are left confused. The situation is often made more difficult by popular media reporting miracle stories and policy makers and various associations for autism not taking a clear stance. For instance, the Autism Society of America embraces a philosophy that primarily seeks to empower the child with autism and their family, claiming that intervention should be based on the child and not any particular program or method, and that what is best for the child should be determined by parents and other people directly involved (Autism Society of America, 2009). Hence, intervention may be sidetracked into a debate on parental choice. What can we do to close this gap? It does not seem that good results and a scientific approach will be enough. Most likely this is a process that will take decades, and depend in particular on the education system changing and one aspect of this is how these issues are included in training curriculum for new teachers, psychologists, social workers etc.

Measuring Outcome

Recovery. Following the publication of the Lovaas study in 1987 there was a heated debate on whether or not the children in the best outcome group should be considered recovered from autism. Lovaas himself questioned this notion maintaining that certain residual deficits may remain in this group that may only be identified on

closer psychological assessments as the children grow older (Lovaas, 1987). In the 1993 follow-up study when the children were on average 13 years of age the best outcome children were put through a more lengthy series of tests, including assessment of areas likely to be difficult for children with autism of average intelligence (compulsive or ritualistic behaviour, empathy for and interest in others, a sense of humour etc). They found that 8 out of 9 of the children in the best outcome group continued to be successful in regular education classes and were indistinguishable from their peers on all of the assessments conducted (McEachin, et al., 1993). Although there are many case stories of recovery from autism following IBI (e.g., Cipani, 2008; Maurice, 1993; Perry, Cohen, & DeCarlo, 1995), this is still the only long term follow up study published with some scientific merit. There appears then to be some hope of recovery, but we do not know at what rate this could be expected or what variables that may mediate or moderate such outcome.

Clinical significance. Even though the reliable change index may be appropriate for dichotomizing outcome success, it does not tell us if the child has moved outside of the dysfunctional range. Jacobson and Truax (1991) propose a number of other criteria that may be used in conjunction with the reliable change index. The one that seems most applicable for the current studies is the criterion that the child should be closer to the mean of the functional population than the mean of the dysfunctional population after intervention (criterion c; Jacobson & Truax, p. 13). One can identify a cut off score above which the child is more similar to the typical population than that of the sample from which it was drawn. To find this cut off score you need to find the halfway point between the mean of the typical population (100 for both IQ and ABC) and that of the groups at intake on the same measures. However, Jacobson and Truax (1991) noted that these criteria might be too strict for people with serious, life-long conditions such as autism. For the analyses in chapter 3 and 4 I was also primarily interested in reliable change rather than children losing the diagnosis. Such clinical change is based on the assumption that you can normalize those who are the target of intervention, and may be more appropriate when assessing small changes (for instance in anxiety) in already high functioning individuals.

Research on Child Development and Other Populations

Can behavioural intervention research teach us anything about child development, such as how to better raise and educate children in general? Can the techniques be applied with children regardless of psychiatric diagnosis or learning difficulties? Can the teaching techniques even be applied to better teach complex skills to adults, like how to program a computer or how to perform surgery? But also we need to further investigate if this type of research with children with autism can tell us anything about how typical development happens? Perhaps we can learn more about the complexities of development and the coalescence between environment, ontogenic and phylogenic histories? Certainly behaviour analysis has solid traditions for inductive research; many of the findings from laboratory experiments with animals hold well with adult human beings.

To my knowledge chapter 5 presents the first study evaluating comprehensive behavioural intervention following the IBI autism model for children with intellectual disability. In the early intervention literature I have found only two other studies that report IQ gains following some form of early intervention. These studies were with children at risk or from disadvantaged homes and it is not known if they would fulfil current diagnostic criteria for developmental disabilities. In one of these studies beginning in the 1930s, 12 children that continued to live in an orphanage were compared to 12 children that were adopted (Skeels, 1966). After two years the children that were adopted scored on average 30 IQ points higher than the children who remained in the orphanage, and they also continued to fare much better in adulthood. Whether or not this should be classed as an intervention study by today standards is questionable. However, these data suggests that a rich a stimulating environment is important. It also underscores the importance of having a comparison or control group when conducting studies with this population, as in contrast to what seems to be the case with children with autism, intervention (regardless of type) may yield significant benefits.

In another study 20 children from disadvantaged homes were given about 30 hours per week of intervention for 2 years and compared to children who received no intervention (Garber, 1988). Intervention consisted of infant stimulation, including 1:1

interaction with a trained caregiver followed by high-quality preschool, parent training, and vocational skills training. At age 10, the children had IQ's like those of low-risk children and substantially higher (on average about 30 IQ points) than those of a randomly selected control group of at-risk children. Although these results were promising at the time, failure to replicate the results, coupled with the fact that the researchers did not answer questions that were raised pertaining to the specific intervention procedures and how the assessments were done, tempered how these data were received by the research community (e.g., Spitz 1986).

There is also a substantial amount of studies on the effects of various large scale childhood programs for children at risk, such as the Head Start program in the United States. However, the children enrolled are not generally intellectually disabled, and IQ and ABC scores are not used systematically to measure outcome. Even though some positive effects have been seen for four and five year olds, the outcome of these types of programs has been disappointing in the long term (Muhlhausen & Lips, 2010). Other studies have monitored the effects of full-inclusion in pre-schools for children with mild developmental delays (not intellectually disabled), and found that this may be related to some improvement in intellectual and adaptive functioning after three years (Guralnick, Neville, Hammond, & Connor 2008). Given the promising results of behavioural interventions for children with more pervasive developmental problems future research should look at how the principles of early behavioural intervention could help children at risk.

Clearly, the findings presented in the present thesis with respect to children with intellectual disability need to be replicated, preferably on a larger scale and with better experimental control (i.e., random assignment to groups). As opposed to the situation that has arisen for behavioural intervention with autism, it should still be feasible (ethically) to conduct such a study given that at present it is uncertain whether behavioural intervention will yield more positive results than other interventions for this population. Furthermore, the research may be easier and more straightforward than research with children with autism. For instance, measures of intellectual functioning and adaptive behaviour may be perfectly appropriate and have excellent face-validity for evaluating intervention effects, as they directly measure the defining properties of the condition. Also, it would perhaps be

appropriate to apply some of the other criteria for clinical significance (in addition to reliable change) suggested by Jacobson and Truax (1991). In particular their criterion b, where it is proposed that the level of functioning subsequent to intervention should fall within the range of the functional/normal population may be relevant. The normal range is defined as within two standard deviations of the mean of the population. This would be similar to the criteria for losing the diagnosis of intellectual disability. The possibility of losing the diagnosis would be in line with the American Association of Intellectual and Developmental Disabilities (AAIDD) moving away from the term mental retardation and instead using the term intellectual disability, as this emphasizes the sense that this condition is no longer considered an absolute, invariable trait of a person (Schalock, et al., 2010).

Scaling and Economical Implications

A pressing issue is whether or not it is possible to provide this intervention on a larger scale? What would happen if IBI became standard provision? What would this require from society at large in terms of funding and logistics? What would it take in terms of educating professionals at various levels? Could it be done? Again, we can consider the larger scale implementation in Canada that suggests that this may be possible, although the costs and benefit of this project are not yet fully known. We also need to consider the potential financial strain on society the growing number of children with autism is posing. It is estimated that on average each person with autism will cost 3.2 million dollars across the lifespan (Ganz, 2007), but that these costs could be cut by about 65 % following effective intervention (Jarbrink & Knapp, 2001). In one study costs associated with generic special education were compared to the costs of three years of IBI. Assuming that approximately 50 % of children achieve near-normal functioning, 40 % have moderate gains and 10 % have no significant gains, it was estimated that an average of 208,500 dollars could be saved for each child before they reach 18 years of age (Chasson, Harris, & Neely, 2007).

Improving the Evidence Base Standard

In summary the evidence base for IBI meets the criteria level 1b for evidence based practices as outlined by the Oxford Centre for Evidence Based Medicine (2009). The

main reason for not meeting the highest criteria (level 1a) is the lack of true random assignment to groups in almost all of the studies published so far. However, the IBI evidence base does meet the highest level of quality indicators for research in special education (Gersten, et al., 2005).

Also, the National Autism Centre has recently published a comprehensive report on the level of scientific evidence that exists to support behavioural and educational interventions for autism spectrum disorders (National Autism Centre, 2009). Based on the literature published until 2007 (22 studies met their inclusion criteria) they concluded that IBI should be considered an established intervention for autism and atypical autism. Although the evidence may seem convincing these conclusions are not universally accepted. For instance some countries (e.g. Scotland and Northern Ireland) have published reports where the conclusions are that a variety of provisions will meet these children's needs and that no single approach meets the needs of children with autism. In the future we clearly need to address these controversies, by publishing outcome research of high quality.

Summary and Conclusions

While the evidence base for IBI is the best documented intervention, it may not be the best intervention. This still remains an empirical question. Furthermore, IBI is likely to continue to change and evolve. Numerous reviews have been published over the last 5 years. However, I have extended and improved on these by applying criteria developed within the framework of evidence based practices. These include measures such as the NNT and ARR which may be easier to communicate to policy makers and funding agencies. Also, the proportions of children meeting reliable change criteria may serve as benchmarks for evaluating future intervention projects, both at the group and individual level. We may ask for instance, how many percent of the children in a particular study met the reliable change criteria proposed here (as I did with the outcome data reported in chapter 4). Similarly, on an individual level we may ask, how much gain did the child make? If the gains are above the criteria for reliable change we would probably have a much stronger argument for continuing behavioural intervention towards potential funding agencies. Finally, these reviews

confirm earlier findings that behavioural interventions should at present be the intervention of choice for autism spectrum disorders, yielding large effect sizes for change in IQ and moderate effect sizes for change in adaptive behaviour.

I have also shown that IBI may be implemented rather successfully on a larger scale in mainstream pre-school settings, without requiring extra resources. This can be thought of as effectiveness research relevant both to policy and general clinical application. Finally, I have evaluated the effects of IBI on children with intellectual disability. Here I demonstrated that this population might benefit from IBI as much as children with autism spectrum disorders. Though this study seems to be first of its kind, it has serious methodological limitations, and clearly a large scale experimental replication is needed. Results are promising and suggest that the behavioural intervention technology may have benefits far beyond children with autism, perhaps also for typical children (and adults) in general education settings.

References

- Altman, D. G. (1998). Confidence Intervals for the number needed to treat. *British Medical Journal*, *317*, 1309–1312.
- 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*, 352–366.
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders : DSM-IV* (4th ed.). Washington, DC: American Psychiatric Association.
- Autism Society of America (2009). Treatment options. Retrieved 30.9, 2009, from http://www.autism-society.org/site/PageServer?pagename=life_treat
- Baer, D. M. (1993). Quasi-random assignment can be as convincing as random assignment. *American Journal on Mental Retardation*, *97*, 373–375.
- Baer, D. M., Wolfe, M., & Risley, T. (1968). Some Current Dimensions of Applied Behavior Analysis. *Journal of Applied Behavior Analysis*, *1*, 91-97.
- Bailey, A., Le Couteur, A., Gottesman, I., & Bolton, P. (1995). Autism as a strongly genetic disorder: Evidence from a British twin study. *Psychological Medicine*, *25*, 63–77.
- Baird, G., Charman, T., Baron-Cohen, S., Cox, A., Swettenham, J., Wheelwright, S., et al. (2000). A screening instrument for autism at 18 months of age: A 6-year follow-up study. *Journal of the American Academy of Child Psychiatry*, *39*, 694–702.
- Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., et al. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *The Lancet*, *368*, 210–215.
- Baltaxe, C. A. (1981). Acoustic characteristics of prosody in autism. In P. Mittler (Ed.), *Frontiers of knowledge in mental retardation*. Baltimore: University Park Press.
- Baltaxe, C. A., & Simmons, J. Q. (1975). Language in childhood psychosis: A review. *Acta Symbolica*, *49*, 439–458.

- Baltaxe, C. A., & Simmons, J. Q. (1981). Disorders of language in childhood psychosis: Current concepts and approaches. In J. Darby (Ed.), *Speech evaluation in psychiatry*. New York: Grune & Stratton.
- Baron-Cohen, S., Allen, J., & Gillberg, C. (1993). Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *Annual Progress in Child Psychiatry & Child Development*, 95–103.
- Bayley, N. (1969). *Bayley Scales of Infant Development*. New York: Psychological Corporation.
- Bayley, N. (1993). *Bayley scales of infant development (2nd ed.)*. San Antonio: The Psychological Cooperation.
- Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development, Third Edition*. San Antonio, TX: Harcourt Assessment.
- Beglinger, L. J., & Smith, T. (2005). Concurrent validity of social subtype and IQ after early intensive behavioural intervention in children with autism: a preliminary investigation. *Journal of Autism & Developmental Disorders*, 35, 295–303.
- Ben-Itzhak, E., & Zachor, D. A. (2007). The effects of intellectual functioning and autism severity on outcome of early behavioural intervention for children with autism. *Research in Developmental Disabilities*, 28, 287–303.
- Ben-Itzhak, E., & Zachor, D. A. (2009). Change in Autism classification with early intervention: Predictors of outcome. *Research in Autism Spectrum Disorders*, 3, 967-976.
- 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": Erratum. *Research in Developmental Disabilities*, 23, 79–104.
- Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data From Berkeley. *Science*, 187, 398–404.
- Bolton, P., Macdonald, H., Pickles, A., Rios, P., & et al. (1994). A case-control family history study of autism. *Journal of Child Psychology and Psychiatry*, 35, 877–900.
- Bondy, A., & Frost, L. (2003). Communication Strategies for Visual Learners. In O. I. Lovaas (Ed.), *Teaching Individuals with Developmental Delays; Basic Intervention Techniques* (pp. 291–304). Austin, Texas: Pro-ed.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. NJ: Wiley.

- Bryson, S. E., Clark, B. S., & Smith, I. M. (1988). First report of a Canadian epidemiological study of autistic syndromes. *Journal of Child Psychology and Psychiatry, 29*, 433–445.
- Buitelaar, J. K., & Willemsen-Swinkels, S. (2000). Medication treatment in subjects with autistic spectrum disorders. *European Child and Adolescence Psychiatry, 9*, 185–197.
- Butter, E. M., Mulick, J. A., & Metz, B. (2006). Eight case reports of learning recovery in children with pervasive developmental disorders after early intervention. *Behavioural 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. *The 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. *The Clinical Psychologist, 49*, 5–18.
- Chasson, G. S., Harris, G. E., & Neely, W. J. (2007). Cost Comparison of Early Intensive Behavioural Intervention and Special Education for Children with Autism. *Journal of Child and Family Studies, 16*, 401–403.
- Chawarska, K., & Bearss, K. (2008). *Assessment of cognitive and adaptive skills*. New York, NY: Guilford Press.
- Chawarska, K., Paul, R., Klin, A., Hannigen, S., Dichtel, L. E., & Volkmar, F. (2007). Parental recognition of developmental problems in toddlers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*, 62–72.
- Cipani, E. (2008). *Triumphs in early autism treatment*. New York, NY: Springer Publishing Co.
- Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early Intensive Behavioural Treatment: Replication of the UCLA Model in a community Setting. *Developmental and Behavioural Pediatrics, 27*, 145–155.
- Coonrod, E. E., & Stone, W. L. (2005). *Screening for Autism in Young Children*. Hoboken, NJ: John Wiley & Sons Inc.

- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied Behavior Analysis* (2nd ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.
- 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.
- Cunningham, A. B., & Schreibman, L. (2008). Stereotypy in autism: The importance of function. *Research in Autism Spectrum Disorders, 2*, 469–479.
- 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 & Adolescent Psychiatry, 11*, 266–272.
- 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.
- Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal, 315*, 629–634.
- Eikeseth, S. (2001). Recent critiques of the UCLA young autism project. *Behavioural Interventions, 16*, 249–264.
- Eikeseth, S. (2009). Outcome of Comprehensive Psycho-Educational Interventions for Young Children with Autism. *Research in Developmental Disabilities, 30*, 158–178.
- Eikeseth, S., Hayward, D. W., Gail, C., Gitlesen, J.-P., & Eldevik, S. (2009). Intensity of Supervision and Outcome for Preschool Aged Children Receiving Early and Intensive Behavioural Interventions: A Preliminary Study. *Research in Autism Spectrum Disorders, 3*, 67–73.
- Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2002). Intensive behavioural 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 behavioural 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 behavioural and

- emotional disturbance in children with mental retardation. *Journal of Autism & Developmental Disorders*, 25, 81–104.
- Einfeld, S. L., & Tonge, B. J. (2002). *Manual for the Developmental Behavior Checklist (2nd ed.)*. Melbourne, Centre for Developmental Psychiatry, Monash University, Clayton, Victoria,: University of New South Wales, School of Psychiatry.
- Eldevik, S., Eikeseth, S., Jahr, E., & Smith, T. (2006). Effects of low-intensity behavioural treatment for children with autism and mental retardation. *Journal of Autism & 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 Behavioural Intervention for Children with Autism. *Journal of Clinical Child and Adolescent Psychology*, 38, 439–450.
- Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (in press). Using individual participant data to extend the evidence base for Intensive Behavioural Intervention for children with autism. *American Journal on Intellectual and Developmental Disabilities*.
- Fay, W. H., & Schuler, A. L. (1980). *Emerging language in autistic children*. Baltimore: University Park Press.
- Fenske, E. C., Zalski, S., Krantz, P. J., & McClannahan, L. E. (1985). Age at intervention and treatment outcome for autistic children in a comprehensive intervention program. *Analysis & Intervention in Developmental Disabilities: Special Issue: Early intervention*, 5, 49–58.
- Ferster, C., & Demyer, M. K. (1962). A method for the experimental analysis of the behavior of autistic children. *American Journal of Orthopsychiatry*, 32, 89–98.
- Filipek, P. A., Accardo, P. J., Baranek, G. T., Cook, E. H., Jr., Dawson, G., Gordon, B., et al. (1999). The screening and diagnosis of autistic spectrum disorders. *Journal of Autism & Developmental Disorders*, 29, 439–484.
- Fombonne, E. (1999). The epidemiology of autism: A review. *Psychological Medicine*, 29, 769–786.
- Fombonne, E. (2003). Epidemiological Surveys of Autism and Other Pervasive Developmental Disorders: An Update. *Journal of Autism and Developmental Disorders*, 33, 365–382.
- Fombonne, E. (2005). The changing epidemiology of autism. *Journal of Applied Research in Intellectual Disabilities*, 18, 281–294.

- Fox, N. A., Calkins, S. D., & Bell, M. A. (1994). Neural plasticity and development in the first two years of life: Evidence from cognitive and socioemotional domains of research. *Development and Psychopathology*, 6, 677–696.
- Foxx, R. M. (1993). Sapid effects awaiting independent replication. *American Journal on Mental Retardation*, 97, 375–376.
- Frazier, S. L., Formoso, D., II, Birman, D., & Atkins, M. S. (2008). Closing the research to practice gap: Redefining feasibility. *Clinical Psychology: Science and Practice*, 15, 125–129.
- 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.
- Ganz, M. L. (2007). The Lifetime Distribution of the Incremental Societal Costs of Autism. *Archives of Pediatrics & Adolescent Medicine*, 161, 343.
- Garber, H. L. (1988). *The Milwaukee Project: Preventing mental retardation in children at risk*. Washington, DC: American Association on Mental Retardation.
- Gautam, K., & Gautam, K. (2008). Addressing the research-practice gap in healthcare management. *Journal of Public Health Management & Practice*, 14, 155–159.
- 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.
- Ghaziuddin, M., & Mountain-Kimchi, K. (2004). Defining the intellectual profile of Asperger syndrome: Comparison with high-functioning autism. *Journal of Autism and Developmental Disorders*, 34, 279–284.
- Gillberg, C. (2007). The autism spectrum. In J. W. Jacobson, J. A. Mulick & J. Rojahn (Eds.), *Handbook of intellectual and developmental disabilities* (pp. 41–59). New York, NY: Springer Publishing Co.
- Granpeesheh, D., Dixon, D. R., Tarbox, J., Kaplan, A. N., & Wilke, A. E. (2009). The effects of age and treatment intensity on behavioural intervention outcomes for children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 3, 1014–1022.

- Green, G. (1996). Early behavioural intervention for autism: What does research tell us? *Behavioural intervention for young children with autism: A manual for parents and professionals* (pp. 29–44). Austin, TX: Pro-Ed, Inc.
- Green, G., Brennan, L. C., & Fein, D. (2002). Intensive behavioural treatment for a toddler at high risk for autism. *Behavior Modification, 26*, 69–102.
- Gresham, F. M., & MacMillan, D. L. (1997). Autistic recovery? An analysis and critique of the empirical evidence on the Early Intervention Project. *Behavioural Disorders, 22*, 185–201.
- Guralnick, M. J. (1998). Effectiveness of early intervention for vulnerable children: A developmental perspective. *American Journal on Mental Retardation, 102*, 319–345.
- Guralnick, M. J., Neville, B., Hammond, M. A., & Connor, R. T. (2008). Continuity and Change From Full-Inclusion Early Childhood Programs Through the Early Elementary Period. *Journal of Early Intervention, 30*, 237–250.
- Handleman, J. S., & Harris, S. L. (2001). *Preschool Education Programs for Children with Autism* (Second ed.). Austin, Texas: Pro-Ed.
- 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 & 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 & Developmental Disorders, 21*, 281–290.
- 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.
- Hewitt, F. J. (1965). Teaching speech to an autistic child through operant conditioning. *American Journal of Orthopsychiatry, 35*, 927–936.
- 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 behavioural interventions (EIBI) 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 behavioural intervention for young children with autism-general model and single state case. *Behavioural 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 & Clinical Psychology*, 59, 12–19.
- Jahr, E., Eldevik, S., & Eikeseth, S. (2000). Teaching children with autism to initiate and sustain cooperative play. *Research in Developmental Disabilities*, 21, 151–169.
- Jarbrink, K., & Knapp, M. (2001). The economic impact of autism in Britain. *Autism*, 5, 7–22.
- Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, 2, 217–250.
- Kazdin, A. E. (1993). Replication and extension of behavioural treatment of autistic disorder. *American Journal on Mental Retardation*, 97, 377–379.
- Klin, A., McPartland, J., & Volkmar, F. R. (2005). *Asperger Syndrome*. Hoboken, NJ: John Wiley & Sons Inc.
- Klin, A., Saulnier, C., Tsatsanis, K., & Volkmar, F. R. (2005). *Clinical Evaluation in Autism Spectrum Disorders: Psychological Assessment within a Transdisciplinary Framework*. Hoboken, NJ: John Wiley & Sons Inc.
- Koegel, L. K., Koegel, R. L., Harrower, J. K., & Carter, C. M. (1999). Pivotal response intervention I: Overview of approach. *Journal of the Association for Persons with Severe Handicaps*, 24, 174–185.
- Koegel, L. K., Koegel, R. L., Shoshan, Y., & McNerney, E. (1999). Pivotal response intervention II: Preliminary long-term outcomes data. *Journal of the Association for Persons with Severe Handicaps*, 24, 186–198.
- 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.[see comment]. *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 & Adolescent Psychiatry*, *42*, 1524–1529.
- Kraemer, H. C., Wilson, G., Fairburn, C. G., & Agras, W. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*, 877–884.
- Langthorne, P., & McGill, P. (2008). Functional analysis of the early development of self-injurious behavior: Incorporating gene-environment interactions. *American Journal on Mental Retardation*, *113*, 403–417.
- 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.
- Leaf, R., & McEachin, J. (1999). *A work in progress*. New York: DRL Books.
- LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials.[see comment]. *New England Journal of Medicine*, *337*, 536–542.
- Leyfer, O. T., Folstein, S. E., Bacalman, S., Davis, N. O., Dinh, E., Morgan, J., et al. (2006). Comorbid Psychiatric Disorders in Children with Autism: Interview Development and Rates of Disorders. *Journal of Autism and Developmental Disorders*, *36*, 849–861.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The Autism Diagnostic Observation Schedule - Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*, 205–223.
- Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism & Developmental Disorders*, *24*, 659–685.
- Lovaas, O., Koegel, R. L., & Schreibman, L. (1979). Stimulus overselectivity in autism: A review of research. *Psychological Bulletin*, *86*, 1236–1254.

- Lovaas, O., Smith, T., & McEachin, J. J. (1989). Clarifying comments on the young autism study: Reply to Schopler, Short, and Mesibov. *Journal of Consulting and Clinical Psychology, 57*, 165–167.
- Lovaas, O. I. (1977). *The autistic child: Language development through behavior modification*. Oxford, England: Irvington.
- Lovaas, O. I. (1981). *Teaching developmentally disabled children : the me book / by O. Ivar Lovaas, with Andrea Ackerman ... [et al.] and with contributions by Edward G. Carr and Crighton Newsom*. Baltimore: University Park Press.
- Lovaas, O. I. (1987). Behavioural treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting & Clinical Psychology, 55*, 3–9.
- Lovaas, O. I. (2003). *Teaching individuals with developmental delays: Basic intervention techniques*. Austin, TX: Pro-Ed.
- Lovaas, O. I., Berberich, J. P., Perloff, B. F., & Schaeffer, B. (1966). Acquisition of imitative speech by schizophrenic children. *Science, 151*, 705–707.
- Lovaas, O., Koegel, R., Simmons, J. Q., & Long, J. S. (1973). Some generalization and follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis, 6*, 131–166.
- Lovaas, O. I., Schaffer, B., & Simmons, J. Q. (1965). Building social behaviors in autistic children by use of electric shock. *Journal of Experimental Research in Personality, 1*, 99–109.
- Lovaas, O. I., & Smith, T. (1988a). *Intensive behavioural treatment for young autistic children*. New York, NY: Plenum Press.
- Lovaas, O. I., & Smith, T. (1988b). Intensive behavioural treatment for young autistic children *Advances in clinical child psychology, Vol. 11* (pp. 285–324). New York, NY: Plenum Press.
- Lovaas, O. I., & Smith, T. (1989). A comprehensive behavioural theory of autistic children: Paradigm for research and treatment. *Journal of Behavior Therapy & Experimental Psychiatry, 20*, 17–29.
- Luiselli, J. K., Cannon, B. O. M., Ellis, J. T., & Sisson, R. W. (2000). Home-based Behavioural 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: The International Journal of Research and Practice, 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.
- Matson, J. L., & LoVullo, S. V. (2008). A review of behavioural treatments for self-injurious behaviors of persons with autism spectrum disorders *Behavior Modification*, 32, 61–76.
- Matson, J. L., Matson, M. L., & Rivet, T. T. (2007). Social-skills treatments for children with autism spectrum disorders: An overview. *Behavior Modification*, 32, 61–76.
- Matson, J. L., & Smith, K. R. (2008). Current status of intensive behavioural interventions for young children with autism and PDD-NOS. *Research in Autism Spectrum Disorders*, 2, 60–74.
- Maurice, C. (1993). *Let me hear your voice : a family's triumph over autism / Catherine Maurice* (1st ed.). New York: Knopf.
- Maurice, C., Green, G., & Foxx, R. M. (2001). *Making a difference: Behavioural intervention for autism*. Austin, TX: Pro-Ed.
- Maurice, C., Green, G., & Luce, S. C. (Eds.). (1996). *Behavioural intervention for young children with autism: A manual for parents and professionals*. Austin, TX: Pro-Ed, Inc.
- Maurice, C., & Taylor, B. A. (2005). Early Intensive Behavioural Intervention for Autism. In W. L. Heward, T. E. Heron, N. A. Neef, S. M. Peterson, D. M. Sainato, G. Cartledge, R. Gardner III, L. D. Peterson, S. B. Hersh & J. C. Dardig (Eds.), *Focus on Behavior Analysis in Education; Achievements, Challenges and Opportunities*. Upper Saddle River: Pearson.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376-390.
- McEachin, J. J. (2006). Comment made to a presentation of a preliminary version of this analysis. Atlanta.
- McEachin, J. J., Smith, T., & Lovaas, O. (1993). Long-term outcome for children with autism who received early intensive behavioural treatment. *American Journal on Mental Retardation*, 97, 359–372.

- Morris, E. K. (2009). A case study in the misrepresentation of applied behavior analysis in autism: The Gernsbacher lectures. *The Behavior Analyst, 32*, 205–240.
- Muhlhausen, D. B., & Lips, D. (2010). Head Start Earns and F: No lasting Impact for Children by First Grade. *Background, 2363*, 1-13.
- Nathan, P., & Gorman, J. M. (2002). *A guide to treatments that work*. New York: Oxford University Press.
- National Autism Center (2009). National Standards Report. Randolph, MA.
- Newsom, C., & Hovanitz, C. A. (1997). Autistic disorder *Assessment of childhood disorders (3rd ed.)* (pp. 408–452). New York, NY: The Guilford Press.
- Novak, G., & Pelaez, M. (2004). *Child and Adolescent Development*. Thousand Oaks: Sage.
- Nuthall, G. (2004). Relating Classroom Teaching to Student Learning: A Critical Analysis of Why Research Has Failed to Bridge the Theory-Practice Gap. *Harvard Educational Review, 74*, 273–306.
- Oxford Centre for Evidence-based Medicine (2009). Levels of Evidence. Retrieved 1.7, 2009, from <http://www.cebm.net/index.aspx?o=1025>
- Perry, R., Cohen, I., & DeCarlo, R. (1995). Case study: Deterioration, autism, and recovery in two siblings. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*, 232–237.
- Perry, A., Penn Flanagan, H. E., & Prichard, E. A. (2008). York Measure of Quality of Intensive Behavioral Intervention. York University.
- Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*. New York, NY: John Wiley & Sons, Inc.
- Pinson, L., & Gray, G. E. (2003). Number needed to treat: An underused measure of treatment effect. *Psychiatric Services, 54*, 145–146.
- Piven, J., Gayle, J., Chase, G. A., Fink, B., & et al. (1990). A family history study of neuropsychiatric disorders in the adult siblings of autistic individuals. *Journal of the American Academy of Child & Adolescent Psychiatry, 29*, 177–183.
- Poplin, M. S. (1988). Holistic/constructivist principles of the teaching/learning process: Implications for the field of learning disabilities. *Journal of Learning Disabilities, 21*, 401–416.

- Reed, P., Osborne, L. A., & Corness, M. (2007a). Brief report: Relative effectiveness of different home-based behavioural 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 Behavioural 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 behavioural intervention: Outcomes for children with autism and their parents after two years. *American Journal on Mental Retardation*, *112*, 418–438.
- Risley, T., & Wolf, M. M. (1967). Establishing functional speech in echolalic children. *Behavior Research & Therapy*, *5*, 73–88.
- Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, *15*, 351–357.
- 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. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition, technical manual*. Itasca, IL: Riverside Publishing.
- Roid, G. M., & Miller, L. M. (1997). *Leiter International Performance Scale-Revised: Examiners Manual*. Wood dale, IL.: Stoelting Co.
- Sallows, G. O., & Graupner, T. D. (2005). Intensive Behavioural Treatment for Children With Autism: Four-Year Outcome and Predictors. *American Journal on Mental Retardation*, *110*, 417–438.
- Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V. J., Buntinx, W. H. E., Coulter, D. L., Craig, E. M. P., et al. (2010). *Intellectual Disability: Definition, Classification, and Systems of Supports (Eleventh edition)* Annapolis: AAIDD Publications.

- Schopler, E., Short, A., & Mesibov, G. (1989). Relation of behavioural treatment to "normal functioning": Comment on Lovaas. *Journal of Consulting & Clinical Psychology, 57*, 162–164.
- Schreck, K. A., & Mazur, A. (2008). Behavior analyst use of and beliefs in treatments for people with autism. *Behavioural Interventions, 23*, 201-212.
- Schreibman, L., Kohlenberg, B. S., & Britten, K. R. (1986). Differential responding to content and intonation components of a complex auditory stimulus by nonverbal and echolalic autistic children. *6*, 109–125.
- Sheinkopf, S. J., & Siegel, B. (1998). Home based behavioural treatment of young children with autism. *Journal of Autism & Developmental Disorders, 28*, 15–23.
- Shook, G. L. (2005). An Examination of the Integrity and Future of the Behavior Analyst Certification Board Reg. Credentials. *Behavior Modification, 29*, 562-574.
- Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., Baird, G., et al. (2008). Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*, 921–929.
- Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Table. *Journal of the Royal Statistical Society, 238–241*.
- Skeels, H. (1966). Adult status of children with contrasting early life experiences. *Monographs of the Society for Research in Child Development, 31*, 1–65.
- Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270-277.
- Skinner, B. F. (1953). *Science and human behavior*. New York: The Macmillan Company.
- 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., Donahoe, P. A., & Davis, B. J. (2001). The UCLA Young Autism Project. In J. S. Handleman & S. L. Harris (Eds.), *Preschool Education Programs for Children with Autism* (Second ed.). Austin, Texas: Pro-Ed.
- Smith, T., Eikeseth, S., Klevstrand, M., & Lovaas, O. I. (1997). Intensive Behavioural 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.
- Smith, T., Klevstrand, M., & Lovaas, O. (1995). Behavioural treatment of Rett's disorder: Ineffectiveness in three cases. *American Journal on Mental Retardation*, 100, 317–322.
- Smith, T., & Lovaas, O. (1997). The UCLA Young Autism Project: A reply to Gresham and MacMillan. *Behavioural Disorders*, 22, 202–218.
- Smith, T., McEachin, J. J., & Lovaas, O. (1993). Comments on replication and evaluation of outcome. *American Journal on Mental Retardation*, 97, 385–391.
- Smith, T., Mruzek, D. W., & Peyton, R. T. (2008). *A study in perseverance: The emergence of early intensive behavioural intervention*. New York, NY: Springer Publishing Co.
- Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., et al. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders*, 37, 354–366.
- 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*. Minnesota: American Guidance Service.
- Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales: Second Edition*. Circle Pines Minnesota: AGS Publishing.
- Spitz, H. H. (1986). Preventing and curing mental retardation by behavioural intervention: An evaluation of some claims. *Intelligence*, 10, 197–207.
- Spreckley, M., & Boyd, R. (2009). Efficacy of Applied Behavioural Intervention in Preschool Children with Autism for Improving Cognitive, Language, and Adaptive Behavior: A Systematic Review and Meta-analysis. *Journal of Pediatrics*, 154, 338–344.
- 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.

- Stein, D. G., Brailowsky, S., & Will, B. (1995). *Brain repair*. New York: Oxford University Press.
- Stoelb, M., Yarnal, R., Miles, J., Takahashi, T., Farmer, J. E., & McCathren, R. B. (2004). Predicting Responsiveness to Treatment of Children with Autism: A Retrospective Study of the Importance of Physical Dysmorphology. *Focus on Autism and Other Developmental Disabilities, 19*, 66–77.
- Straus, S. E., Newton, D., & Tomlinson, G. (2004). Centre For Evidence-Based Medicine Retrieved April 1st, 2008, 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.
- Sundberg, M. L., & Partington, J. W. (1998). *Teaching language to children with autism or other developmental disabilities*. Danville, CA: Behavior Analysts, Inc.
- 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: John Wiley & Sons Inc.
- Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research; A missing link to creating optimally effective communication interventions. *Mental Retardation & Developmental Disabilities Research Reviews, 13*, 70–77.
- Wechsler, D. (1974). *The Wechsler Intelligence Test for Children-Revised*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1989). *Wechsler Preschool and Primary Scale of Intelligence-Revised*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1993). *Wechsler Intelligence Scale for Children (3rd ed.)*. San Antonio, TX: the Psychological Corporation.

- Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence-III*. San Antonio, TX: Psychological Corporation.
- Weinberg, R. A. (1989). Intelligence and IQ: Landmark issues and great debates. *American Psychologist: Special Issue: Children and their development: Knowledge base, research agenda, and social policy application*, 44, 1998–1104.
- Weiss, M. J. (1999). Differential rates of skill acquisition and outcomes of early intensive behavioural intervention for autism. *Behavioural Interventions*, 14, 3–22.
- World Health Organization (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.
- World Health Organization (1993). *ICD-10, the ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*. Geneva: World Health Organization.

Appendix: Code sheet for selecting studies.

First author:
Three first words in title of paper:
Year published:

Type of treatment

		Yes	No
1	<i>Clinic directed behavioural treatment: Treatment managed through a University clinic, centre or community service. Treatment may be carried out in the child's home, or at a centre or at another place. Supervision on a regular basis (i.e. at least every 6 weeks). Treatment intensity (number of</i>		
2	<i>Parent managed behavioural treatment: Treatment managed by the parents and carried out in their home. Supervision on a regular basis (i.e. at least every 6 weeks). Treatment intensity (number of hours per week) is specified.</i>		
3	<i>Workshop based behavioural treatment: Treatment is generally carried out in the child's home. Consultations are typically given every 3 months, but may not be given more frequent than every 6 weeks (if so it would be considered parent managed). Treatment intensity (number of hours per week)</i>		
4	<i>Comparison groups: Called comparison groups because some other identifiable treatment is provided for about the same number of hours per week as behavioural treatment. This may also include giving a mixture of other treatments. Treatment intensity (number of hours per week) is</i>		
5	<i>Control groups: Called control groups because an unknown or poorly specified provision is given. Sometimes this may be referred to only as TAU (treatment as usual). However, the treatment may be known but given at a substantially lesser intensity than behavioural treatment.</i>		
5a	<i>A minimum of an AB design (Not a case study).</i>		

Other variables

6	<i>Children in the study are between 2 and 7 years of age at intake.</i>		
7	<i>All children in the study are diagnosed with autism or PDD-NOS, based on standard diagnostic manuals such as the ICD or DSM.</i>		
8	<i>Diagnosis of autism/PDD-NOS given by clinician and/or using standardized measures such as the ADI-R.</i>		
9	<i>Duration of treatment is at least 12 months</i>		
10	<i>Testing done at intake (no more than 3 months after/before treatment started).</i>		
11	<i>Follow up or post testing done.</i>		
12	<i>Global IQ measure. The IQ measure must include performance and verbal subtests. Only administering a performance IQ-test such as the Leiter or the Merrill-Palmer will not be adequate.</i>		
13	<i>A standardized measure behaviour (such as the Vineland Adaptive Behavior Scales) is done at intake and after 12-36 months.</i>		
14	<i>Include in analysis? Inclusions criteria: To be included there must be: (i) a YES on criterion 1, 2 or 3; and 5a (ii) a YES on 6, 7, 8, 9, 10, and 11 and a YES on 12 and/or 13.</i>		
Comments/questions:			

Scorer initials and date: