Review Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s40489-023-00355-9

REVIEW PAPER

Measures Used to Assess Treatment Outcomes in Children with Autism Receiving Early and Intensive Behavioral Interventions: A Review

Samantha Ridout[1,2] · **Sigmund Eldevik**[3]

Received: 6 July 2021 / Accepted: 3 February 2023

Abstract

This review is aimed at identifying assessment instruments used to measure treatment outcomes in children with autism spectrum disorder who received early and intensive behavioral interventions. Forty three articles were included and appraised using the Council for Exceptional Children's Standards for Evidence Based Practice quality index rater. Ninety-two outcome measures were discovered. Measures of adaptive functioning (91%), intellectual functioning (86%), and core symptoms (67%) of autism were represented with the highest frequencies. Measures of challenging behavior and parent or caregiver wellbeing were reported at 30% and 14% respectively. Reliability and validity of each measure were determined by recently published psychometric data. The utility of outcome measures in clinical practice is discussed.

Keywords Autism · Early intensive behavioral intervention · Outcome measures · Treatment outcomes

Autism spectrum disorder is classified as a neurodevelopmental disorder with marked impairments in social interactions, communication, and the presence of restricted and repetitive behaviors. Heterogeneity within the disorder is large, with presentation or symptom severity ranging from mild to severe. It is estimated that 1 in 54 children in the United States will receive an autism spectrum diagnosis (Centers for Disease Control, 2020). The diagnostic process includes a clinical evaluation alongside caregiver reports, with most children receiving their diagnosis between the ages of 2 and 6 years (Fletcher-Watson & McConachie, 2017).

Given the substantial empirical support, Early Intensive Behavioral Interventions (herein referenced as EIBI) are well-established and effective treatments for children with autism based on the principles of applied behavior analysis which are typically employed to very young children at intensities of 20–40 h per week (Reichow et al., 2018). EIBI models such as the UCLA or Lovaas model employ one-to-one, systematic teaching procedures known as Discrete Trial Teaching and Incidental Teaching (Lovaas, 1987). Models such as the Early Start Denver model embed learning opportunities into the contexts of the child's naturally occurring routine (Rogers & Dawson, 2010) whereas other naturalistic models such as Pivotal Response Training may focus on pivotal areas of the child's development such as the child's motivation and self-management (Koegel & Koegel, 2006). These structured, individualized teaching programs that are designed to address a wide range of developmental areas (Vismara & Rogers, 2010) focus on acquiring new skill repertoires and/or decreasing challenging behavior, are typically carried out in the child's home or clinical center, and are usually funded through public health, education budgets, or insurance. Desired outcomes of these interventions include a reduction in the severity of autism core symptoms such as increased social communication and language, increased adaptive behaviors, and a reduction in the frequency and severity of restricted and repetitive behaviors and maladaptive behaviors.

Several systematic reviews and meta-analyses have discussed positive outcomes in intellectual functioning and adaptive behavior regarding treatment outcomes for children who participated in EIBI programs (Eldevik et al., 2009; Peters-Scheffer et al., 2011; Reichow et al., 2018), with some evidence that these gains are maintained overtime (see Smith et al., 2019a, b). Emerging evidence for similar behavioral based interventions has shown results in developmental changes in infants and toddlers such as normalized brain activity (Dawson et al., 2012) and improvements in verbal

✉ Samantha Ridout

[1] University of Oslo, Oslo, Norway

[2] Greensboro, USA

[3] Oslo Metropolitan University, Oslo, Norway

Published online: 06 May 2023

developmental quotients (Vivanti & Dissanayake, 2016). However, gains differ between individuals, and several factors may influence treatment outcomes such as: milder symptom severity and intellectual functioning at intake (Ben-Itzchak & Zachor, 2007; Fossum et al., 2018; Smith et al. 2015a, b; Zachor et al., 2007), age of treatment onset (Harris & Handleman, 2000), intensity of supervision (Eikeseth et al., 2009), and treatment intensity (Makrygianni & Reed, 2010).

Despite the growing body of literature supporting improved outcomes for children receiving early and intensive behavioral interventions, researchers lack a consensus regarding the selection of outcome measures. Chosen measures should demonstrate sensitivity, as they must detect any gains made over the course of treatment; reliability, in that they can be depended upon to deliver accurate measurement across different assessors, and different points in time; and should demonstrate validity, that is, assessments accurately measure what they proport to. Previous reviews in EIBI outcome research have identified a large volume of outcome measures used in ASD research (Bolte & Diehl, 2013; Stolte et al., 2016). The variety and inconsistencies found in these reviews could reflect frequent revision of measures, shifting administration requirements and the vast number of tools available in the market.

There has been some discussion as to how and what should be assessed as part of an initial diagnostic battery. For example, Ozonoff et al. (2005) suggested an initial assessment battery to include measures of autism severity, intellectual functioning, adaptive functioning, and a language assessment. Matson and Rieske (2014) extend this to include measures of challenging behavior, direct measures of targeted behavior (focused criterion referenced measurement), family or consumer satisfaction, and treatment side effects. A review of assessments by Goulde and colleagues (Gould et al., 2011) discusses what they determined to be critical assessment components for use in EIBI programs. They suggest assessments must be comprehensive, targeting all aspects of child development and human functioning. Assessments should also target early childhood development, that is, assessments should be useable for children from infancy until early childhood and should be age-normed and age-appropriate. Assessments should consider behavior function and not just the topography of the behavior. Finally, assessment should provide a direct link to specific targets or goals.

Considerations for Research

When selecting outcome measures, goals of the assessment must be considered. Standardized measures are used often in outcome research and may be important in evaluating large scale effects of treatment. However, these measures require a large degree of generalization and often measure tasks that are never directly addressed in treatment (Rogers & Vismara, 2014). Reassessments using standardized measures are not typically recommended in intervals of less than one year. Alternatively, criterion-referenced assessments measure individual performance against an objective criterion, identify specific skills and skill deficits, may aid in curriculum development, and detect moderate or specific gains of treatment (Granpeesheh et al., 2009; Lotfizadeh et al., 2020).

When evaluating treatment effectiveness, scoring and score interpretation should also be considered. Standard scores are the preferred method for reporting change, as they measure progress in comparison to same-age peers, represent statistically robust gains, and are prevalent in outcome research. Although small increases in raw scores may represent meaningful change, the corresponding standard scores may not increase and can even decrease over time. Reporting age-equivalents as an alternative to standard scores has been suggested in the literature, as standard scores may mask intervention effectiveness (Klintwall et al., 2013). Age-equivalents can be converted to learn rates, which may reflect progress of slower learners with greater accuracy and may better communicate outcomes to parents and stakeholders (Klintwall et al., 2013).

Finally, when selecting assessment tools, researchers must consider the available resources. Master-level clinicians and behavior analysts typically meet Pearson's qualifications at the B-Level (Qualifications Policy, n.d.), which require one or more of the following: master's degree in a filed closely related to the intended assessment, certification by applicable professional organizations, formal supervised training, license to practice in healthcare or allied health, or employment with an accredited institution. Several standardized and diagnostic assessments require additional intensive training, are time-consuming and costly, or require administration by a licensed professional, limiting their utility as feasible, quick and cost-effective methods of assessment.

As more states require EIBI programs to be funded through insurance, identifying psychometrically strong assessments for use within the ASD population to measure outcomes is critical and can contribute to improving both clinical and research-based evaluations.

Purpose

The goals of this study are to review the literature and identify outcome measures and published evidence of their psychometric properties. Our research questions are as follows: What measures have been used up until now to assess treatment outcomes in EIBI research? What are the current psychometric properties of these measures?

Are the identified instruments reliable? Is there published evidence of the validity of these measures as tools to assess treatment outcomes? Finally, are these measures sensitive enough to measure gains over time? Findings are aimed at providing brief recommendations for selecting appropriate assessment tools as part of a developing set of standards for EIBI research.

Methods

Inclusion Criteria

The selection criteria were determined a priori. In order to capture as much published literature as possible, the inclusion criteria were kept intentionally broad. Outcome studies were selected and appraised if (1) interventions were comprehensive and based on the principles of applied behavior analysis, including Lovaas-style EIBI programs (Lovaas, 1987), Pivotal Response Treatment (Koegel & Koegel, 2006) and the Early Start Denver Model (Rogers & Dawson, 2010); (2) participants received at least 5 h per week of 1:1 treatment; (3) participants were a maximum of 7 years of age at the onset of treatment; (4) children had a diagnosis of autism spectrum disorder or pervasive developmental disorder—not otherwise specified; (5) the study specified the use of at least one standardized measurement tool to assess treatment outcomes in one or more domains, such as adaptive functioning, intellectual functioning, or autism core symptom severity; (6) the study utilized group designs; and (7) the study was published in a peer-reviewed journal, in English, between 2006 and 2021.

Search and Search Strategy

The electronic search was performed between the 12th and 14th of January in 2021 in the databases PsycINFO and ERIC using a combination of the following keywords: autism and/or pervasive developmental disorders, children, EIBI or early intensive behavioral intervention, applied behavior analysis, and outcome measures or treatment outcomes. Guidance from a librarian at the University of Oslo was used to determine appropriate usage of Boolean search terms. The electronic search retrieved a total of 517 peer-reviewed articles; 383 articles were excluded for irrelevance, publication before 2006, incorrect diagnosis, and/or duplication. Of the remaining 135 articles, 104 articles were selected for full-text screening and more detailed coding. Studies were deemed eligible for inclusion and quality appraisal if they met all the inclusion criteria listed above. Thirty-five articles from the database search met inclusion criteria; an additional 8 studies were retrieved through hand search by examining the reference sections of the included articles, yielding a total of 43 articles included in the review. See Fig. 1 for search and selection procedure.

Quality Appraisal and Interrater Agreement

Articles were appraised for methodological rigor using the *Council for Exceptional Children Standards for Evidence-Based Practices in Special Education* (Lane et al., 2014). The *Standards for EBP* is a quality index matrix which appraises scientific publications based on eight domains. Quality indicators are met when raters agree the study satisfactorily addresses the content outlined in each indicator (CEC, 2014). All included studies were evaluated by the author and one independent rater. Raters worked together on the first 10 articles before scoring independently, disagreements were discussed, and interrater agreement was determined to be > 95%.

Analysis

Outcome measures were extracted and coded using a matrix of whether they assessed (1) intellectual functioning, (2) language ability, (3) adaptive functioning, (4) ASD symptom severity, (5) challenging behavior, (5) parental wellbeing, or (7) a criterion-referenced or direct observation measure. A total of 92 outcome measures were found across the 43 included articles in this review. This total reflects sequential revisions to instruments as separate measures (e.g., Vineland-2 and Vineland-3 are recorded as two independent measures). Measures of intellectual functioning (86%) and adaptive functioning (91%) were most prevalent in the literature, followed by measures of core symptom severity (67%). Measures of language ability and challenging behavior were found in 33% and 30% of the published papers, respectively. Measures to assess parental wellbeing were found in 14% of articles, and 6% of articles reported the use of manualized, criterion-based measures. A brief description of each measure, including cost, administration, reliability, validity, and frequency in which they appear in the literature, is reported in Appendix A. Although earlier editions to instruments will be referenced throughout the following sections, they will be cited in their most current edition for ease of reference and clarity. Psychometrics of measures reported in three or more articles are included below.

The reliability of the measures was evaluated based on the following coefficient scale: 0.00 to 0.59—very poor reliability, 0.60–0.69—low or poor reliability, 0.70–0.79—moderate to fair reliability, 0.80–0.89—good reliability, and 0.90–0.99—excellent reliability. The validity of the assessment was determined as satisfactory if we could find

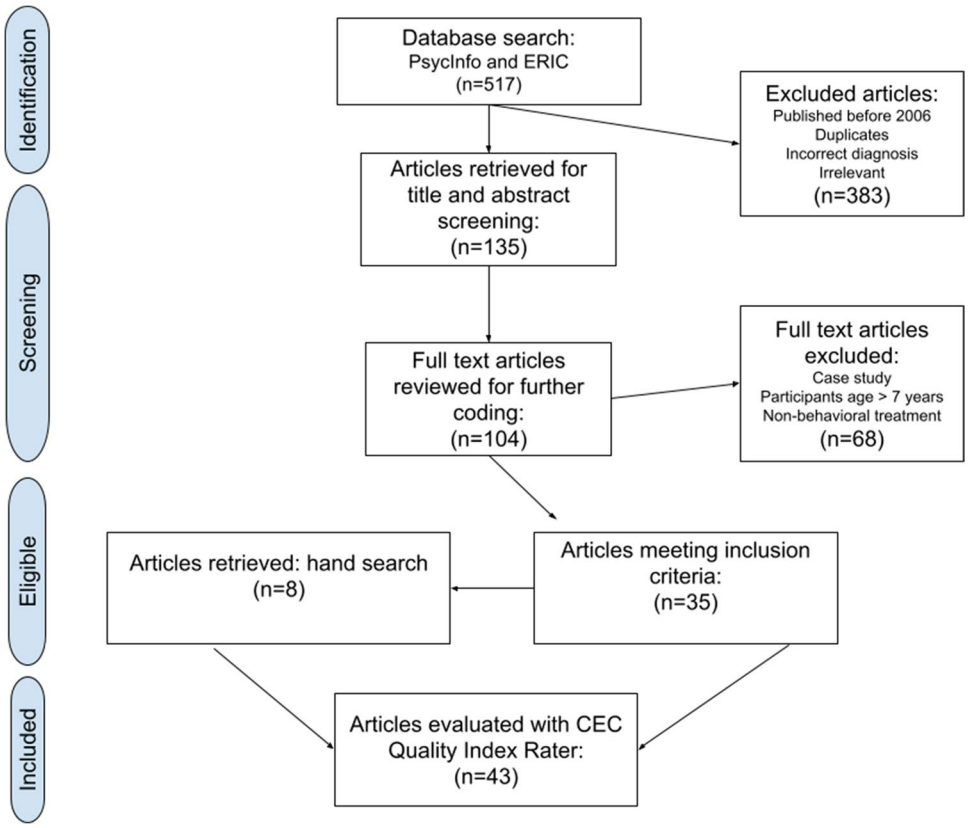

Fig. 1 Database search and selection procedure

Table 1 Matrix of reported scores in included articles for intellectual and adaptive functioning

Domain		Standard scores	Ratio score	Raw scores	Age equivalents
Intellectual	Composite	37	14	0	4
	Subscale	7	5	1	3
Adaptive	Composite	27	1	2	7
	Subscale	20	1	3	8

Intellectual and adaptive functioning are categorized into composite and subscale as many articles reported both composite scores and subscale scores

current-published evidence of criterion validity, concurrent validity, or construct validity.

Secondary variables such as score reporting methods and intervals between assessments were also examined. Table 1 provides frequencies of scores reported in standard scores (SS), age equivalents (AE), ratio scores (RA), or raw scores (RW).

Time between assessment administrations was determined as the interval between the initial assessment (T1) and outcome measurement (T2). If more than two assessments were provided, the time interval between each assessment was recorded (ex. T1: baseline, T2: after 3 months of treatment, T3: outcomes after 6 months of treatment = 3 month intervals between assessments). Table 2 describes measures used in assessment intervals of one year or less.

Measures of Intellectual Functioning

Measures of intellectual functioning appear frequently in the literature (Matson & Rieske, 2014). Thirty-seven of the forty-three articles, report at least one measure of intellectual functioning. Thirty different measures of intellectual functioning were reported. More than half (53%) of the articles reported the use of more than one measure of intellectual functioning, either across participants or across time. Forty percent (17 out of 43) of articles computed ratio IQ scores for at least some of their participants. Full Scale Measures of Intelligence (FSIQ) was reported in 74% (32 out of 43) of articles. Some articles used a mix of FSIQ and nonverbal intelligence tests (4 out of 43, 9%) and reported only the use of nonverbal tests (2 out of 43, 5%) to measure intellectual functioning. Measures of full scale intelligence include Bayley Scales of Infant Development (Bayley-4), Mullen Scales of Early Learning (MSEL), Wechsler Preschool and Primary Scale of Intelligence (WPPSI-IV), PsychoEducational Profile-Revised (PEP-3), Differential Abilities Scale (DAS-II),

Table 2 Measures used in intervals of one year or less

Author	3–6 months	9 months	1 year
Costanza et al., 2018	GMDS, VABS		
Cohen et al., 2006			BSID, WPPSI, NRDLS, M-P-R, VABS,
Dawson et al., 2010			MSEL, VABS, ADOS, RBS
Rogers et al., 2019			MSEL, VABS, ADOS
Eikeseth et al., 2012			VABS, CARS
Fava et al., 2011	ADOS, GMDS, CDI, VABS, CBCL, PSI, DO		
Howard et al., 2014			VABS, M-P-R, WPPSI, WISC, SB, DAS, NRDLS, ROWPVT, EOWPVT, SICDT, PPVT, EVT
Macdonald et al., 2014			ESAT
Peters-Scheffer et al., 2010		BSID, SON 2.5–7, VABS, CBCL, PDD-MRS	
Reed et al., 2007		GARS, PEP-R, BAS-EY, VABS	
Reed & Osborne, 2012		GARS, PEP-R, BAS-EY, VABS	
Remington et al., 2007			BSID, SB, NRDLS, VABS, NCBRF, DBC, ASQ, ESCS, HADS, QRS, KIPP
Smith et al., 2019a, b			VABS, SRS, SIB-R
Strauss et al., 2012	ADOS, GMDS, VABS, CDI, DO		
Vivanti et al., 2019		LENA, M-COSMIC, MSEL, VABS, PSI	
Waters et al., 2020			WPPSI, WISC, BSCID, DP-2, PEP-R, Leiter, SB, M-P-R, VABS, WIAT, WRAT
Zachor et al., 2007			ADOS, BSID, SB
Zachor & Ben Itzchak, 2010			ADOS, VABS, MSEL
Lin et al., 2020	MSEL, ADOS		
Paynter et al., 2018			MSEL, SCQ, VABS
Smith et al., 2015a, b	PLS, CELF, PPVT, M-P-R, WPPSI, VABS, SRS, CBCL, PSI-SF		
Fossum et al., 2018			PLS, CELF, VABS, SRS, M-P-R
Ben-Itzchak & Zachor, 2007			ADOS, BSID, SB
Smith et al., 2015b			MSEL, VABS, ADOS, ADI-R
Vivanti & Dissanayake, 2016			MSEL, ADOS, VABS
Vivanti et al., 2013			MSEL, ADOS
Eldevik et al., 2019			VABS, BSID, SB, CARS
Lotfizadeh et al., 2020	VB-MAPP		VABS

ADOS Autism Diagnostic Observation Schedule, *ADI-R* Autism Diagnostic Interview-Revised, *ASQ* Autism Spectrum Questionnaire, *BAS-EY* British Abilities Scales-Early Years, *BSID* Bayley Scales of Infant Development, *CARS* Childhood Autism Rating Scale, *CDI* Macarthur Bates Communicative Development Inventory, *CBCL* Child Behavior Checklist, *CELF* Clinical Evaluation of Language Fundamentals, *DAS* Differential Abilities Scale, *DP-2* Developmental Profile-2, *DO* Direct Observation Measure (author), *EOWPVT* Expressive One Word Picture Vocabulary Test, *ESAT* Early Skills Assessment Tool, *ESCS* Early Social Communication Scales, *EVT* Expressive Vocabulary Test, *GARS* Gilliam Autism Rating Scales, *GMDS* Griffith Mental Development Scales, *HADS* Hospital Anxiety and Depression Scale, *KIPP* Kansas Inventory of Parental Perceptions, *LENA* Language ENvironment Assessment, *Leiter* Leiter International Performance Scale, *M-COSMIC* Modified Classroom Observation Schedule to Measure Intentional Communication, *M-P-R* Merrill Palmer Scale of Mental Tests-Revised, *MSEL* Mullen Scales of Early Learning, *NRDLS* New Reynell Developmental Language Scales, *NCBRF* Nisonger Child Behavior Rating Form, *QRS* Questionnaire on Resources and Stress, *PEP-R* Psychoeducational Profile-Revised, *PDD-MRS* Scale of Pervasive Developmental Disorder in Mentally Retarded Persons, *PLS* Preschool Language Scales, *PPVT* Peabody Picture Vocabulary Test, *PSI/PSI-SF* Parental Stress Index/ Short Form, *RBS* Repetitive Behavior Scales, *ROWPVT* Receptive One Word Picture Vocabulary Test, *SB* Stanford Binet Intelligence Scales, *SCQ* Social Communication Questionnaire, *SIB-R* Scales of Independent Behavior-Revised, *SICDT* Sequenced Inventory of Communication Development-Revised, *SON 2.5–7* Snijders-Oomen Nonverbal Intelligence Test, *SRS* Social Responsiveness Scale, *VABS* Vineland Adaptive Behavior Scale, *VB-MAPP* Verbal Behavior-Milestones Assessment and Placement Program, *WIAT* Wechsler Individual Achievement Test, *WISC* Wechsler Intelligence Test for Children, *WPPSI* Wechsler Preschool and Primary Scales of Intelligence, *WRAT* Wide Range Achievement Test

Stanford Binet (SB-5), and the Wechsler Intelligence Scale for Children (WISC-V).

Both Wechsler tests (WPPSI-IV, WISC-V) are considered to have excellent internal consistency reliability and show satisfactory criterion validity, though tests are limited. SB-5 has excellent internal consistency and test–retest reliability; satisfactory concurrent validity and may be useful for older children with significant developmental delays (Klinger et al., 2018). DAS-II is considered to have excellent reliability and shows satisfactory concurrent validity. PEP-R has been reported to have good internal reliability (Reed et al., 2007) and has been found to correlate highly with measures like Childhood Autism Rating Scale and the original Vineland Adaptive Behavior Scales, Expanded Form (Naglieri et al., 2018). Bayley-4 is reported to have excellent internal consistency reliability and good test–retest reliability, correlates with similar developmental measures, and has a good degree of classification accuracy (convergent validity). Construct and convergent validity of the Mullen Scales of Early Learning has been demonstrated in young children with ASD (Swineford et al., 2015). Internal consistency reliability of the scales ranges from satisfactory to good and from good to excellent for the Early Learning Composite. Test–retest reliability is good for children ages 1 month to 24 months, but poorer reliability has been reported for children 25 to 56 months (Shank, 2018).

Measures of nonverbal intelligence were reported for some participants but were typically used as part of a comprehensive intellectual assessment. In two articles**,** the Merrill-Palmer Scale of Mental Tests Revised (M-P-R) was used in place of a FSIQ (Fossum et al., 2018; Smith et al., 2010). M-P-R has excellent reliability and has evidence of content and criterion-related validity, correlations to the Bayley Scales, and the abbreviated version of the SB-5.

Measures of Adaptive Functioning

Adaptive functioning was predominantly measured by the Vineland Adaptive Behavior Scales (Vineland-3). All 39 articles reporting a measure of adaptive functioning used either the first or the second edition of the Vineland to assess outcomes of adaptive functioning. In three articles, the Child Behavior Checklist was used as a supplement to the Vineland (Eikeseth et al., 2007; Fava et al., 2011; Peters-Scheffer et al., 2010), and in one case, the Developmental Profile 1 and 2 was used (Waters et al., 2020).

The Vineland has excellent internal consistency reliability. Test–retest reliability at the domain level ranges from moderate to excellent, while test–retest reliability for the adaptive behavior composite is considered good to excellent. The Vineland demonstrates satisfactory construct, content, and concurrent validity as reported by the Vineland-3 publication summary.

Measures of Autism Core Symptoms

Measures of autism core symptoms were identified in 33 articles. Of these articles, 15 assessment tools were identified. The original and revised versions of the Autism Diagnostic Interview (ADI-R), Autism Diagnostic Observation Schedule (ADOS-2), and Childhood Autism Rating Scale (CARS2-ST) were the most prevalent. Both the ADI-R and ADOS-2 are considered the "gold standard" in autism diagnosis and measurement (Ozonoff et al., 2005). The ADOS demonstrates excellent internal consistency, interrater and test–retest reliabilities, and excellent diagnostic validity in distinguishing individuals with autism and those without autism. ADI-R has good intraclass correlations (Ozonoff et al., 2005) and has been shown to correlate with the Social Communication Questionnaire (Naglieri et al., 2018 p. 43.). Although the ADI-R has empirical support for discriminating ASD from other developmental disorders, these findings are limited to children whose mental age is above 2 years (Ozonoff, 2005). The CARS2-ST demonstrates excellent internal reliability, and many studies demonstrate diagnostic and criterion- related validity (Ozonoff et al., 2005; Naglieri et al., 2018 p. 51).

The ADI-R and ADOS are limited in their utility of measures of change over time. However, it is possible to use parts of the ADI-R for measuring sensitivity across time using the ADOS as a guide to compare scores (Gotham et al., 2009). In general, the CARS is more suited for measuring change over time (CARS2-ST).

The Gilliam Autism Rating Scale (GARS-2; Gilliam, 2006) has internal consistency, and test–retest reliabilities are reported as good for the subscales and excellent for the Autism Indexes. Interrater reliability for the Autism Index is good. GARS has excellent sensitivity and specificity and correlates with other measures of ASD diagnostics, though specifics were not provided. Reliability and validity of the GARS were obtained from Pearson Assessments website (Pearson Assessments, nd). The Social Responsiveness Scale (SRS-2) was the final measure used in three or more articles. Internal consistency reports are in the range of excellence for all age ranges. Interrater reliability between parents and teachers for both school age and preschool forms was low to fair. Correlations between SRS-2 and Child Behavior Checklist were found by the authors to be moderate, noting SRS-2 was more sensitive to specific behaviors associated with ASD (Naglieri et al., 2018 p. 61–65). The Early Social Communication Scales (ESCS) is a manualized, direct-observation measure using video recordings to assess nonverbal communication skills in children with mental ages between 8 and 30 months and was the only criterion-referenced measure used to assess outcomes in core symptoms. Recently published reliability and validity of the ESCS could not be found. Reliability and validity of the author-created direct observation tools were not included.

Maladaptive Behavior

Thirteen articles reported a measure addressing either repetitive or challenging behavior (30%). Of these articles, 11 measures were reported. The Child Behavior Checklist (CBCL 1.5–5) ($n = 4$), Nisonger Child Behavior Rating Form (NCBRF) ($n = 2$), and Maladaptive Domain of the Vineland ($n = 2$) were reported more than once. Both articles reporting use of the NCBRF used only the Positive Social subscale to report outcomes of challenging behavior; a 10-item Likert scale provides general descriptions of prosocial behaviors and may not accurately reflect specific challenging behaviors. The Maladaptive Behavior subscale of the Vineland Adaptive Behavior Scales was reported in two articles (Eikeseth et al., 2007, 2012), though recent reliability and validity of this subscale could not be found. Test–retest reliabilities for the CBCL 1.5–5 are considered good, though interrater reliabilities between parents and teachers are low. Additionally, the manual provides evidence of construct, criterion, and content validities. The Repetitive Behavior Scale-Revised (RBS-R) was the only assessment tool used to measure restrictive and repetitive behaviors observed in individuals with autism. Outcomes related to the reduction of RRBs were reported in 3 out of 43 (7%) articles. RBS-R shows good internal consistency reliability has been validated in ASD populations though sample sizes were small (Hooker et al., 2019; Lam & Aman, 2007).

Language Assessment

Measures designed to assess language were found in 14 of 43 articles, and fourteen different measures were found. The following measures were reported in three or more articles: the Reynell Developmental Language Scales-3rd edition (Edwards et al., 1999) ($n = 5$), the third and fourth editions of the Peabody Picture Vocabulary Tests (PPVT-V; Dunn, 2019) ($n = 4$), Macarthur Bates Communicative Developmental Inventories (CDI) ($n = 3$), Expressive One Word Picture Vocabulary Tests (EOWPVT-R) ($n = 3$), and Preschool Language Scales-fourth edition (PLS-5; Zimmerman et al., 2011) ($n = 3$). Thirteen of the 14 measures used to report language functioning focus exclusively on receptive and expressive vocabulary. All reliability and validity measures for the PPVT-5 indicate good to excellent reliability, good clinical validity in autism populations, and moderate correlations to similar measures (Dunn, 2019). Internal consistency reliability of the EOWPVT is reported as acceptable, with excellent test–retest reliability. Additionally, the EOWPVT has been shown to correlate with other measures of vocabulary such as the WISC-4 VCI and WISC-4 FSIQ (Frauwirth et al., 2018). Most recent psychometrics were not available for the PLS-5, NRDLS, or the Macarthur Bates CDI. Outcomes assessed using the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP) are included here as it primarily measures language functioning in young children. The Verbal Behavior-Milestones Assessment and Placement Program (VB-MAPP) is a criterion-referenced assessment and curriculum development tool designed to measure and develop skills in language and related skills. Interrater reliability for the Total Milestones was reported as good (0.87), though low to poor (0.62) reliability for the Barriers Assessment was reported (Montallana et al., 2019). Content validity of the VB-MAPP was recently examined by national experts. Domain relevance, age appropriateness, method of measurement, and domain representativeness were considered to be moderate to strong (Padilla & Akers, 2021).

Parent or Caregiver Wellbeing

Parent or caregiver wellbeing was measured in 6 out of the 43 articles. The Short-Form of the Parenting Stress Index (PSI-4 SF) was used in 3 out of the 6 articles reporting a measure of parental well-being. The PSI-4 provides a measure of 120 items designed to quantify parent and child characteristics, as well as situational and demographic information which may be influencing familial stress. Internal reliability for the two domains and the Total Stress scale reported as excellent, though test–retest reliabilities were mixed and ranged from poor to good. Validation in families of children with autism was not reported. The Hospital and Depression Scale (HADS), Questionnaire on Resources and Stress-Short Form, and Kansas Inventory of Parental Perceptions were reported once, though psychometrics for these instruments could not be found.

Other Measures

Academic achievement ($n = 2$) was measured by the Wide Range Achievement Test 3rd and 4th edition (WRAT) or the Wechsler Individual Achievement Test-II (WIAT-4) ($n = 1$). Play was assessed using the Symbolic Play Test and the Test of Pretend Play in one article. None of these measures were reported more than once in this review.

Discussion

Core Findings by Domain

Adaptive Functioning

The Vineland Adaptive Behavior Scales was indicated as the measure of choice when reporting outcomes in adaptive functioning, used in 39 out of the 43 published articles (91%). Due to its strong psychometric properties, ease of

administration, and developmental comprehensiveness, the Vineland is considered the gold standard when selecting measures of adaptive functioning. Standards scores for the Vineland were reported most frequently, though age equivalents, raw scores, and ratio scores were also reported. Cost and qualifications to administer the Vineland were compared with other standardized measures of adaptive functioning, such as the Adaptive Behavior Assessment System (ABAS-3; Harrison & Oakland, 2013) or the Scales of Independent Behavior-Revised. The Vineland is typically assessed in intervals of one year, indicating that it is a robust and sensitive measure when evaluating outcomes over time.

Measures of Autism Core Symptoms

Although it may be unreasonable to expect changes in diagnostic status over time (Reichow et al., 2018; Vivanti & Dissanayake, 2016), measures of core symptoms of ASD are a critical component of a comprehensive assessment. In this review, the ADOS was most frequently reported to evaluate the effects of treatment on autism core symptom. The ADI-R was used to compare outcomes in some articles but was used primarily to confirm an autism diagnosis. Both the ADOS and ADI-R are considered the gold-standard in autism diagnostic measurement, given their excellent sensitivity and specificity to determine an autism diagnosis, but may be limited when measuring changes in scores over time. However, guidelines for how to use the ADOS for measuring change over time have been published (Gotham, et al., 2009). Clinical limitations to the ADOS include a licensed professional to administer and are time-consuming and costly. However, unlike many measures of core symptoms which rely on parent or caregiver report, the ADOS modules use direct testing and observation. An alternative to these measures could be the CARS-2ST which relies on both direct and indirect measures to evaluate symptom severity, may be sensitive to changes in severity over time, and does not require a licensed professional. Finally, restricted and repetitive behaviors were assessed by the Repetitive Behavior Scale-Revised, a continuous measurement tool rating the frequency and severity of common behaviors in ASD (Lam & Aman, 2007). The RBS-R has good reliability, but published validity studies indicate mixed results.

Intellectual Functioning

Providing measurement of full-scale intelligence presents challenges within a clinical context. Lengthy assessment times and stringent qualifications create practical challenges to repeated administrations necessary for determining outcomes. However, there is research that suggests intellectual functioning at intake is a predictor of treatment outcomes (Smith et al., 2015a, b) and therefore should be considered as part of a comprehensive assessment in research. Although ASD is not defined by intellectual functioning, it is an important variable measure to better describe the sample and any variations within the group. The DSM-V categorizes the disorder by level of severity and level or support required. Increases in intellectual functioning following intervention will often decrease the level of support needed. Measures like the Bayley Scales and Weschler Preschool and Primary Scales have strong psychometrics and were represented frequently within the literature (see Appendix A). Nonverbal intelligence tests such as the Merrill-Palmer-Revised have attractive stimuli which may retain the interests of some children and were somewhat prevalent in the literature but may inflate intelligence scores in young children (Eldevik et al., 2006) and are therefore not recommended as a primary measure of intellectual functioning. Though used less frequently, the Psychoeducational Profile and Differential Abilities Scales may more accurately reflect intellectual functioning in individuals who do not reach basal levels or have aged-out of measures like the Bayley, Weschler tests, or the Mullen Scales. Because full-scale intelligence testing requires significant time and high levels of qualifications to administer, these instruments may not be feasible or practical for applications at the agency level; however, they should be included when used in outcome research.

Language Assessment

Language outcomes were primarily measured by standardized assessments of receptive and expressive language. Ten different measures were used, with the Reynell Developmental Language Scales being reported in 5 of the10 papers. Although frequently used, current published reliability and validity data could not be found. In addition, a speech pathologist credential is required for administration. The PPVT, EVT, ROPVT, and EOPVT all utilize direct testing and observations, have good psychometric properties, and are norm referenced. Only one study used the Verbal Behavior Milestone Assessment and Placement Program as a measure of language ability (Lotfizadeh et al., 2020). The VB-MAPP uses direct observations to measure language and related skills such as play, social, and motor skills. From a clinical standpoint, criterion referenced measurement tools like the VB-MAPP can be readministered in shorter intervals and help guide moment-to-moment treatment decisions (Granpeesheh et al., 2009). The VB-MAPP has promising psychometrics and has been shown to correlate with other behavioral measures. However, the VB-MAPP Barriers Assessment was found to have poor reliability and should be used with caution. As is the recommendation for

assessment in general, the use of the VB-MAPP should be used in conjunction with other measurement tools (Montallana et al., 2019; Padilla & Akers, 2021).

Parent and Caregiver Wellbeing

Parental outcomes were primarily assessed by the Parental Stress Index. Parents of children receiving EIBI make a considerable time, financial, and emotional contribution (Matson & Rieske, 2014); thus, stress and parents' perceived relationships with their children are good indicators of the family's well-being. The PSI-4 demonstrates excellent reliability, but research to determine the validity of the PSI-4 and PSI-4 SF in families of children with autism is needed.

Maladaptive Behavior

Maladaptive behavior was largely measured by informant-based checklists and rating scales. While measures like the Maladaptive Behavior domain on the Vineland-3 or the Child Behavior Checklist may give some indication of frequency and/or severity of the behavior, they do not provide an accurate description of the function or context of the behavior and may primarily serve as screeners to a more extensive assessment such as a functional analysis. Measures like the Questions about Behavior Function do provide an indication of function but may not capture a reduction in maladaptive behaviors over time. Measures with published psychometrics within this domain were difficult to find, with the Maladaptive Behavior domain of the Vineland demonstrating the strongest evidence of good psychometrics.

Limitations and Future Research

This paper reviewed 43 articles reporting outcomes for children who received early and intensive behavioral intervention. The majority of outcome measures fell within the domains of adaptive functioning (91%), intellectual functioning (86%), and core symptoms of ASD (77%). This review extends the existing body of knowledge by pooling together both standardized and criterion-referenced measures toward standardization of measurement selection in outcome research.

There are some limitations to the current study. Although efforts were made to ensure as much of the published EIBI outcome literature was captured, due to the timing of the search, some relevant papers may have been missed. Single case designs, which are frequently used in educational, and behavior analytic research were excluded, and therefore, it is possible some measures, especially criterion-referenced measures, may be missing or under-represented. These criterion-referenced assessments as measures of treatment effectiveness should be further explored.

This review touched on the intervals at which measures are administered, but more research into the sensitivity of these instruments to detect change over shorter periods of time is warranted. Further research into the prevalence and validity of measures of social validity, treatment side effects, and quality of life would extend findings from this review. Finally, future research may be able to discern the frequency to which the identified measures are being used in clinical practice and whether a gap between research and practice exists.

Conclusion and Brief Recommendations for Research

This review attempted to identify assessments used to measure treatment outcomes within EIBI outcome research; however, research-informed practice is a hallmark of applied behavior analytic treatment, and these findings may be of interest to practitioners and insurance payors alike. No longer considered "experimental" treatments, EIBI interventions are now required to be funded through private insurance in almost all 50 states (Zhang & Cummings, 2020). Mandates to provide documentation and measures of treatment outcomes are certainly considered by many to be a positive movement in the field. Measures that are sensitive to change over relatively small periods of time (e.g., 6 month intervals), inexpensive, easy to administer, and psychometrically strong are likely to appeal to insurance funders. Two organizations, the Behavioral Health Center of Excellence (BHCOE) and the International Consortium for Health Outcomes Measurement (ICHOM), have recently published frameworks for selecting appropriate instruments to measure treatment outcomes for children with ASD (BHCOE, 2019; Kazemi et al., 2023; ICHOM, 2023). These frameworks seem to be well aligned with the research literature, are comprehensive, and provide associated costs of recommended tools. The recommendations can be accessed via the ICHOM and BHCOE websites respectively. In alignment with current suggestions from the literature (measures should have representative norms and strong psychometrics; multiple measures should be used and address core symptomology of ASD) and the more comprehensive BHCOE and ICHOM frameworks, the following brief recommendations for assessing outcomes in research are provided:

When possible, a measure of full-scale intelligence should be considered, at least, at the on and offset of treatment. Measures like the Bayley-4 and Mullen Scales and WPPSI are prevalent in the literature, have strong psychometrics, and are based on direct observation and testing. When unable to reach basal or aging out the Psychoeducational Profile or Differential Abilities Scales may be suitable alternatives. Another alternative is to compute a ratio IQ based on

the mental age scores from the Bayley-4 or similar. The Vineland-3 provides a representative measure of adaptive functioning, has been validated for use in ASD populations, and can be administered by most service providers; therefore, it should be considered the gold-standard for assessment of adaptive functioning. Core symptoms may be accurately represented by the CARS2-ST as it is based on both informed report and direct observation of the child. SRS-2 or SCQ may be considered supplementary or additional measures when necessary. As a newer measure, the Autism Impact Measure has recently gained interest as a measure of core symptoms (Kanne et al., 2014), though more research is necessary to determine if it is both sensitive and psychometrically valid. Finally, direct observation or criterion-based measures such as the VB-MAPP or the ABLLS-R may be a useful and sensitive measure of treatment outcomes (Granpeesheh et al., 2009; Titlestad & Eldevik, 2019).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40489-023-00355-9.

Acknowledgements The primary author would like to acknowledge Martine Bjerke for her contribution to interrater agreement.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital)

Data Availability The data that supports the findings of this study can be obtained from the corresponding author upon reasonable request.

Declarations

Conflicts of Interest The authors declare no competing interests.

References

Behavioral Health Center of Excellence. (2021). *Selecting appropriate measurement instruments to assess treatment outcomes of individuals with autism spectrum disorder: guidelines for practitioners, payors, patients, and other stakeholders*. Author.

Ben-Itzchak, E., & Zachor, D. A. (2007). The effects of intellectual functioning and autism severity on outcome of early behavioral intervention for children with autism. *Research in Developmental Disabilities, 28*(3), 287–303.

Bolte, E. E., & Diehl, J. J. (2013). Measurement tools and target symptoms/skills used to assess treatment response for individuals with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 43*(11), 2491–2501. https://doi.org/10.1007/s10803-013-1798-7

CDC. (2020). *Basics About Autism Spectrum Disorder (ASD) | NCBDDD | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/facts.html

Cohen, H., Amerine-Dickens, M., & Smith, T. (2006). Early intensive behavioral treatment: Replication of the UCLA model in a community setting. *Journal of Developmental and Behavioral Pediatrics, 27*(2), S145–S155. https://doi.org/10.1097/00004703-200604002-00013

Costanza, C., Narzisi, A., Ruta, L., Cigala, V., Gagliano, A., Pioggia, G., Siracusano, R., Rogers, S. J., & Muratori, F. (2018). Implementation of the early start Denver model in an Italian community. *Autism: The International Journal of Research and Practice, 22*(2), 126–133.

Dawson, G., Jones, E. J., Merkle, K., Venema, K., Lowy, R., Faja, S., Kamara, D., Murias, M., Greenson, J., Winter, J., Smith, M., Rogers, S. J., & Webb, S. J. (2012). Early behavioral intervention is associated with normalized brain activity in young children with autism. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(11), 1150–1159. https://doi.org/10.1016/j.jaac.2012.08.018

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: The early start Denver model. *Pediatrics (evanston), 125*(1), e17–e23. https://doi.org/10.1542/peds.2009-0958

Dunn, D. M. (2019). *Peabody Picture Vocabulary Test* (5th ed.). Pearson.

Edwards, S., Fletcher, P., Garman, M., Highes, A., Letts, C., & Sinka, I. (1999). *Reynell Developmental Language Scales-III*. NFER-Nelson.

Eikeseth, S., Hayward, D., Gale, C., Gitlesen, J.-P., & Eldevik, S. (2009). Intensity of supervision and outcome for preschool aged children receiving early and intensive behavioral interventions: A preliminary study. *Research in Autism Spectrum Disorders, 3*(1), 67–73. https://doi.org/10.1016/j.rasd.2008.04.003

Eikeseth, S., Klintwall, L., Jahr, E., & Karlsson, P. (2012). Outcome for children with autism receiving early and intensive behavioral intervention in mainstream preschool and kindergarten settings. *Research in Autism Spectrum Disorders, 6*(2), 829–835. https://doi.org/10.1016/j.rasd.2011.09.002

Eikeseth, S., Smith, T., Jahr, E., & Eldevik, S. (2007). Outcome for children with autism who began intensive behavioral treatment between ages 4 and 7: A comparison controlled study. *Behavior Modification, 31*(3), 264–278. https://doi.org/10.1177/0145445506291396

Eldevik, S., Titlestad, K. B., Aarlie, H., & Tønnesen, R. (2019). Community implementation of early behavioral intervention: higher intensity gives better outcome. *European Journal of Behavior Analysis*. https://doi.org/10.1080/15021149.2019.1629781

Eldevik, S., Eikeseth, S., Jahr, E., & Smith, T. (2006). Effects of low-intensity behavioral treatment for children with autism and mental retardation. *Journal of Autism and Developmental Disorders, 36*(2), 211–224. https://doi.org/10.1007/s10803-005-0058-x

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*(3), 439–450. https://doi.org/10.1080/15374410902851739

Fava, L., Strauss, K., Valeri, G., D'Elia, L., Arima, S., & Vicari, S. (2011). The effectiveness of a cross-setting complementary

staff- and parent-mediated early intensive behavioral intervention for young children with ASD. *Research in Autism Spectrum Disorders, 5*(4), 1479–1492. https://doi.org/10.1016/j.rasd.2011.02.009

Fletcher-Watson, S., & McConachie, H. (2017). The search for an early intervention outcome measurement tool in autism. *Focus on Autism and Other Developmental Disabilities, 32*(1), 71–80. https://doi.org/10.1177/1088357615583468

Fossum, K.-L., Williams, L., Garon, N., Bryson, S. E., & Smith, I. M. (2018). Pivotal response treatment for preschoolers with autism spectrum disorder: Defining a predictor profile. *Autism Research, 11*(1), 153–165. https://doi.org/10.1002/aur.1859

Frauwirth, S., Michalec, D., Henninger, N. (2018). Expressive one-word picture vocabulary test. In Kreutzer, J.S., DeLuca, J., Caplan, B. (eds) *Encyclopedia of Clinical Neuropsychology*. Springer. https://doi-org.ezproxy.uio.no/10.1007/978-3-319-57111-9_1544

Gilliam, J. (2006). *Gilliam Autism Rating Scale-Second Edition*. PRO-ED.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(5), 693–705. https://doi.org/10.1007/s10803-008-0674-3

Gould, E., Dixon, D. R., Najdowski, A. C., Smith, M. N., & Tarbox, J. (2011). A review of assessments for determining the content of early intensive behavioral intervention programs for autism spectrum disorders. *Research in Autism Spectrum Disorders, 5*(3), 990–1002.

Granpeesheh, D., Dixon, D. R., Tarbox, J., Kaplan, A. M., & Wilke, A. E. (2009). The effects of age and treatment intensity on behavioral intervention outcomes for children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 3*(4), 1014–1022. https://doi.org/10.1016/j.rasd.2009.06.007

Harris, S. L., & Handleman, J. S. (2000). Age and IQ at intake as predictors of placement for young children with autism: A four- to six-year follow-up. *Journal of Autism and Developmental Disorders, 30*, 137–142.

Harrison, P. L., & Oakland, T. (2013). *Adaptive Behavior Assessment System* (3rd ed.). Western Psychological Services.

Hooker, J., Dow, D., Morgan, L., Schatschneider, C., & Wetherby, A. (2019). Psychometric analysis of the repetitive behavior scale-revised using confirmatory factor analysis in children with autism: psychometric analysis of the RBS-R. *Autism Research*, *12*. https://doi.org/10.1002/aur.2159

Howard, J. S., Stanislaw, H., Green, G., Sparkman, C. R., & Cohen, H. G. (2014). Comparison of behavior analytic and eclectic early interventions for young children with autism after three years. *Research in Developmental Disabilities, 35*(12), 3326–3344. https://doi.org/10.1016/j.ridd.2014.08.021

International Consortium for Health Outcome Measures. (2023). *Patient-centered outcome measures: autism spectrum disorder*. Retrieved January 2023. https://connect.ichom.org/patient-centered-outcome-measures/autism-spectrum-disorder/

Kanne, S. M., Mazurek, M. O., Sikora, D., Bellando, J., Branum-Martin, L., Handen, B., Katz, T., Freedman, B., Powell, M. P., & Warren, Z. (2014). The autism impact measure (AIM): initial development of a new tool for treatment outcome measurement. *Journal of Autism and Developmental Disorders, 44*(1), 168–179. https://doi.org/10.1007/s10803-013-1862-3

Kazemi, E., Eldevik, S., Adzhyan, P., Cox, D., & Litvak, S. (2023). Selecting appropriate assessment instruments to measure behavior analytic treatment outcomes for individuals with autism spectrum disorder (ASD). Autism: The International Journal of Research and Practice. Submitted

Klinger, L. G., Mussey, J. L., & O'Kelly, S. (2018). Assessment of intellectual functioning in autism spectrum disorder. In S. Goldstein & S. Ozonoff (Eds.), *Assessment of Autism Spectrum Disorders* (2nd ed., pp. 215–262). Guilford Press.

Klintwall, L., Eldevik, S., & Eikeseth, S. (2013). Narrowing the gap: effects of intervention on developmental trajectories in autism. *Autism: The International Journal of Research and Practice*, *19*. https://doi.org/10.1177/1362361313510067

Koegel, R. L., & Koegel, L. K. (2006). *Pivotal response treatments for autism: communication, social, & academic development*. Brookes Publishing Co.

Lam, K. S. L., & Aman, M. G. (2007). The repetitive behavior scale-revised: Independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 37*(5), 855–866. https://doi.org/10.1007/s10803-006-0213-z

Lane, K. L., Common, E. A., Royer, D. J., & Muller, K. (2014). Group comparison and single case research design quality indicator matrix using Council for Exceptional Children 2014 standards. Unpublished tool. Retrieved from https://www.ci3t.org/practice

Lin, T.-L., Chiang, C.-H., Ho, S. Y., Wu, H.-C., & Wong, C.-C. (2020). Preliminary clinical outcomes of a short-term low-intensity early start Denver model implemented in the Taiwanese public health system. *Autism: THe International Journal of Research and Practice, 24*(5), 1300–1306. https://doi.org/10.1177/1362361319897179

Lotfizadeh, A. D., Kazemi, E., Pompa-Craven, P., & Eldevik, S. (2020). Moderate effects of low-intensity behavioral intervention. *Behavior Modification, 44*(1), 92–113. https://doi.org/10.1177/0145445518796204

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55*(1), 3–9. https://doi.org/10.1037/0022-006X.55.1.3

MacDonald, R., Parry-Cruwys, D., Dupere, S., & Ahearn, W. (2014). Assessing progress and outcome of early intensive behavioral intervention for toddlers with autism. *Research in Developmental Disabilities, 35*(12), 3632–3644. https://doi.org/10.1016/j.ridd.2014.08.036

Makrygianni, M. K., & Reed, P. (2010). A meta-analytic review of the effectiveness of behavioural early intervention programs for children with autistic spectrum disorders. *Research in Autism Spectrum Disorders, 4*(4), 577–593.

Matson, J. L., & Rieske, R. D. (2014). Are outcome measures for early intensive treatment of autism improving? *Research in Autism Spectrum Disorders, 8*(3), 178–185. https://doi.org/10.1016/j.rasd.2013.11.006

Montallana, K. L., Gard, B. M., Lotfizadeh, A. D., & Poling, A. (2019). Inter-rater agreement for the milestones and barriers assessments of the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP). *Journal of Autism and Developmental Disorders, 49*(5), 2015–2023. https://doi.org/10.1007/s10803-019-03879-4

Naglieri, J. A., Chambers, K. M., McGoldrick, K. D., & Goldstein, S. (2018). Psychometric issues and current scales for assessing autism spectrum disorder. In S. Goldstein & S. Ozonoff (Eds.), *Assessment of Autism Spectrum Disorders* (2nd ed., pp. 26–71). Guilford Press.

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology*, American Psychological Association, Division 53, *34*(3), 523–540. https://doi.org/10.1207/s15374424jccp3403_8

Paynter, J., Trembath, D., & Lane, A. (2018). Differential outcome subgroups in children with autism spectrum disorder attending early intervention: ASD outcome subgroups. *Journal of*

Intellectual Disability Research, 62(7), 650–659. https://doi.org/10.1111/jir.12504

Padilla, K. L., & Akers, J. S. (2021). Content validity evidence for the verbal behavior milestones assessment and placement program. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-020-04864-y

Peters-Scheffer, N., Didden, R., Korzilius, H., & Sturmey, P. (2011). A meta-analytic study on the effectiveness of comprehensive ABA-based early intervention programs for children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 5*(1), 60–69. https://doi.org/10.1016/j.rasd.2010.03.011

Peters-Scheffer, N., Didden, R., Mulders, M., & Korzilius, H. (2010). Low intensity behavioral treatment supplementing preschool services for young children with autism spectrum disorders and severe to mild intellectual disability. *Research in Developmental Disabilities, 31*(6), 1678–1684. https://doi.org/10.1016/j.ridd.2010.04.008

Qualifications Policy. (n.d.). Retrieved May 8, 2021, from https://www.pearsonassessments.com/professional-assessments/ordering/how-to-order/qualifications/qualifcations-policy.html

Reed, P., Osborne, L. A., & Corness, M. (2007). Brief report: Relative effectiveness of different home-based behavioral approaches to early teaching intervention. *Journal of Autism and Developmental Disorders, 37*(9), 1815–1821. https://doi.org/10.1007/s10803-006-0306-8

Reed, P., & Osborne, L. (2012). Impact of severity of autism and intervention time-input on child outcomes: Comparison across several early Iinterventions. *British Journal of Special Education, 39*(3), 130–136. https://doi.org/10.1111/j.1467-8578.2012.00549.x

Reichow, B., Hume, K., Barton, E. E., & Boyd, B. A. (2018). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *Cochrane Database of Systematic Reviews*, *5*. https://doi.org/10.1002/14651858.CD009260.pub3

Remington, B., Hastings, R. P., Kovshoff, H., degli Espinosa, F., Jahr, E., Brown, T., Alsford, P., Lemaic, M., & Ward, N. (2007). Early intensive behavioral intervention: outcomes for children with autism and their parents after two years. *American Journal of Mental Retardation, 112*(6), 418–438. https://doi.org/10.1352/0895-8017(2007)112[418:EIBIOF]2.0.CO;2

Rogers, S. J., & Dawson, G. (2010). *Early start Denver model for young children with autism: Promoting language, learning and engagement*. Guilford Press.

Rogers, S., & Vismara, L. (2014). Interventions for infants and toddlers at risk for autism spectrum disorder. In F. Volkmar, S. Rogers, R. Paul, & K. Pelphrey (Eds.), *Handbook of Autism and Pervasive Developmental Disorders* (4th ed., Vol. 2). Wiley.

Rogers, S. J., Estes, A., Lord, C., Munson, J., Rocha, M., Winter, J., Greenson, J., Colombi, C., Dawson, G., Vismara, L. A., Sugar, C. A., Hellemann, G., Whelan, F., & Talbott, M. (2019). A multisite randomized controlled two-phase trial of the early start Denver model compared to treatment as usual. *Journal of the American Academy of Child and Adolescent Psychiatry, 58*(9), 853–865. https://doi.org/10.1016/j.jaac.2019.01.004

Shank, L. (2018). Mullen Scales of Early Learning. In Kreutzer, J.S., DeLuca, J., Caplan, B. (Eds.), Encyclopedia of Clinical Neuropsychology. Springer. https://doi-org.ezproxy.uio.no/10.1007/978-3-319-57111-9_1570

Smith, D. P., Hayward, D. W., Gale, C. M., Eikeseth, S., & Klintwall, L. (2019a). Treatment Gains from Early and Intensive Behavioral Intervention (EIBI) are maintained 10 years later. *Behavior Modification*, 14544551988289–145445519882895. https://doi.org/10.1177/0145445519882895

Smith, I. M., Flanagan, H. E., Ungar, W. J., D'Entremont, B., Garon, N., den Otter, J., Waddell, C., Bryson, S. E., Tsiplova, K., Léger, N., Vezina, F., & Murray, P. (2019b). Comparing the 1-year impact of preschool autism intervention programs in two Canadian provinces. *Autism Research, 12*(4), 667–681. https://doi.org/10.1002/aur.2072

Smith, I. M., Flanagan, H. E., Garon, N., & Bryson, S. E. (2015a). Effectiveness of community-based early intervention based on pivotal response treatment. *Journal of Autism and Developmental Disorders, 45*(6), 1858–1872. https://doi.org/10.1007/s10803-014-2345-x

Smith, I. M., Koegel, R. L., Koegel, L. K., Openden, D. A., Fossum, K. L., & Bryson, S. E. (2010). Effectiveness of a novel community-based early intervention model for children with autistic spectrum disorder. *American Journal on Intellectual and Developmental Disabilities, 115*(6), 504–523. https://doi.org/10.1352/1944-7558-115.6.504

Smith, T., Klorman, R., & Mruzek, D. W. (2015b). Predicting outcome of community-based early intensive behavioral intervention for children with autism. *Journal of Abnormal Child Psychology, 43*(7), 1271–1282. https://doi.org/10.1007/s10802-015-0002-2

Stolte, M., Hodgetts, S., & Smith, V. (2016). A critical review of outcome measures used to evaluate the effectiveness of comprehensive, community based treatment for young children with ASD. *Research in Autism Spectrum Disorders, 23*, 221–234.

Strauss, K., Vicari, S., Valeri, G., D'Elia, L., Arima, S., & Fava, L. (2012). Parent inclusion in early intensive behavioral intervention: The influence of parental stress, parent treatment fidelity and parent-mediated generalization of behavior targets on child outcomes. *Research in Developmental Disabilities, 33*(2), 688–703. https://doi.org/10.1016/j.ridd.2011.11.008

Swineford, L. B., Guthrie, W., & Thurm, A. (2015). Convergent and divergent validity of the Mullen Scales of Early Learning in young children with and without autism spectrum disorder. *Psychological Assessment, 27*(4), 1364–1378. https://doi.org/10.1037/pas0000116

Titlestad, K. B., & Eldevik, S. (2019). Brief report: Modest but clinically meaningful effects of early behavioral intervention in twins with Rett syndrome-a case study. *Journal of Autism and Developmental Disorders, 49*(12), 5063–5072.

Vismara, L. A., & Rogers, S. J. (2010). Behavioral treatments in autism spectrum disorder: What do we know? *Annual Review of Clinical Psychology, 6*(1), 447–468. https://doi.org/10.1146/annurev.clinpsy.121208.131151

Vivanti, G., Dissanayake, C., Zierhut, C., & Rogers, S. J. (2013). Brief report: Predictors of outcomes in the early start Denver model delivered in a group setting. *Journal of Autism and Developmental Disorders, 43*(7), 1717–1724. https://doi.org/10.1007/s10803-012-1705-7

Vivanti, G., & Dissanayake, C. (2016). Outcome for children receiving the early start Denver model before and after 48 months. *Journal of Autism and Developmental Disorders, 46*(7), 2441–2449. https://doi.org/10.1007/s10803-016-2777-6

Vivanti, G., Dissanayake, C., Duncan, E., Feary, J., Capes, K., Upson, S., Bent, C. A., Rogers, S. J., Hudry, K., Jones, C., Bajwa, H., Marshall, A., Maya, J., Pye, K., Reynolds, J., Rodset, D., & Toscano, G. (2019). Outcomes of children receiving group-early start Denver model in an inclusive versus autism-specific setting: A pilot randomized controlled trial. *Autism: THe International Journal of Research and Practice, 23*(5), 1165–1175. https://doi.org/10.1177/1362361318801341

Waters, C. F., Amerine Dickens, M., Thurston, S. W., Lu, X., & Smith, T. (2020). Sustainability of early intensive behavioral intervention for children with autism spectrum disorder in a community

setting. *Behavior Modification, 44*(1), 3–26. https://doi.org/10.1177/0145445518786463

Zachor, D. A., Ben-Itzchak, E., Rabinovich, A.-L., & Lahat, E. (2007). Change in autism core symptoms with intervention. *Research in Autism Spectrum Disorders, 1*(4), 304–317. https://doi.org/10.1016/j.rasd.2006.12.001

Zachor, D. A., & Ben Itzchak, E. (2010). Treatment approach, autism severity and intervention outcomes in young children. *Research in Autism Spectrum Disorders, 4*(3), 425–432. https://doi.org/10.1016/j.rasd.2009.10.013

Zhang, Y. X., & Cummings, J. R. (2020). Supply of certified applied behavior analysts in the United States: implications for service delivery for children with autism. *Psychiatric Services (Washington, D.C.), 71*(4), 385–388. https://doi.org/10.1176/appi.ps.201900058

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales* (5th ed.). Pearson.