Review Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s40489-020-00205-y

REVIEW PAPER

A Systematic Review of Behaviour-Related Outcome Assessments for Children on the Autism Spectrum with Intellectual Disabilities in Education Settings

Melanie Howell[1] · Jill Bradshaw[1] · Peter E. Langdon[2,3]

Received: 22 September 2019 / Accepted: 26 March 2020

Abstract
A systematic review was completed to identify assessments used with children with intellectual disabilities to assess adaptive behaviour, challenging behaviour and autism-related behaviour and consider their appropriateness for use by special education teachers with autistic pupils. The findings of this review led to the recommendation that the Pervasive Development Disorder Behavior Inventory, Aberrant Behavior Checklist, Autism Treatment Evaluation Checklist and the Teacher Autism Progress Scale are currently the most appropriate assessments for these purposes, although some limitations of these assessments exist. Additional recommendations included teacher input during the development of robust assessments to show progress for autistic children with intellectual disabilities and further evaluation of commonly used assessments with an appropriate sample in a relevant education setting.

Keywords Autism · Intellectual disabilities · Education · Assessment · Measurement · Measurement properties

Autism spectrum disorder affects between 1 and 2% of the population (Baron-Cohen et al. 2009) with an estimated 55% of this population also having intellectual disabilities (Knapp et al. 2007). Autism is a lifelong condition characterised by difficulties in social communication and interaction and the presence of restricted and repetitive behaviours or interests (American Psychiatric Association 2013). Coexisting intellectual disabilities are also associated with an increase in challenging behaviour and difficulties with social skills (Walton and Ingersoll 2013). Around 25% of this population are pre- or minimally verbal (Arnold and Reed 2016) and the possible requirement for special education provision is one of the reasons autistic individuals with intellectual disabilities are considered a group with distinct needs compared with those with autism alone (Matson and Shoemaker 2009).

✉ Melanie Howell
msh46@kent.ac.uk

[1] Tizard Centre, University of Kent, Canterbury, Kent, UK

[2] Centre for Educational Development, Appraisal and Research (CEDAR), University of Warwick, Coventry, UK

[3] Coventry and Warwickshire Partnership NHS Trust, Coventry, UK

Assessment in Special Education

Much research in the area of educational assessment has focused on mainstream schools. Although only around 29% of autistic children are educated in special schools (Arnold and Reed 2016), it is important to ensure that school-based assessments are appropriate for these pupils. Aside from the statutory assessment of core curriculum subjects (English, maths and science), there is no specification of how other areas of learning are assessed, leaving the format of assessment to individual schools (Office for Standards in Education 2019).

'School functioning' has been suggested as a quality of life domain for children and, in special schools, abilities and behaviours affecting and limiting school functioning are often key areas targeted by special educational needs (SEN) teachers (Burgess and Gutstein 2007). Small improvements of these basic skills may be a priority, especially for autistic children with severe to profound intellectual disabilities (Pellicano et al. 2014). Parent and professional groups have suggested that progress and outcomes should be measured in a broad repertoire of behaviours such as social skills, adaptive skills and coexisting problems (McConachie et al. 2018). For autistic individuals with intellectual disabilities, good outcomes may include improvements in adaptive behaviour

Published online: 29 April 2020

(including functional communication, independence and daily living skills), appropriate social behaviour (including reductions in challenging and problem behaviour) and autism-related behaviours (such as restricted and repetitive behaviours and interests). These areas of difficulty can also result in barriers to accessing learning and stagnating progress in curriculum areas (Steer 2009), adversely affecting school functioning and, consequently, quality of life (Burgess and Gutstein 2007). For these reasons, assessment of these areas will be the focus of this review.

Much assessment research in the area of SEN and autism has addressed diagnosis and screening rather than outcomes and progress (Wigelsworth et al. 2015). Few assessments specific to populations with coexisting autism and intellectual disabilities are in common use in schools and, routinely, autistic pupils are assessed using generic assessments developed for all pupils regardless of diagnosis (Arnold and Reed 2016). In light of the specific needs of this group discussed above, this review aimed to identify assessments which measure outcomes for school aged children on the autism spectrum with coexisting intellectual disabilities in areas of adaptive behaviour, challenging or problem behaviour and autism-related behaviour. The term *autism-related behaviour* was preferred to language such as 'autism symptomatology' but includes measures pertaining to these areas as well as restricted and repetitive behaviours (RRBs) and sensory behaviour. The measurement properties of the assessments were considered and discussed in light of their potential for use by teachers in schools.

Previous Reviews

Three previous systematic reviews have been conducted with aims particularly relevant to this review. McConachie et al. (2015) conducted the first comprehensive review of the quality and appropriateness of progress and outcome measures for children on the autism spectrum. Strong evidence was found for 12 assessments, the majority of which assessed autism characteristics and problem behaviour, and the importance of measuring 'functioning in everyday life' was highlighted (McConachie et al. 2015, p.121). However, only assessments used with children under 6 years old were considered and the review did not focus on tools that can be used within special schools.

Hanratty et al. (2015) conducted a systematic review of behaviour problem assessments for children on the autism spectrum under 6 years old. Six assessments were evaluated and the measurement properties of the Child Behavior Checklist and the Home Situations Questionnaire (HSQ) were found to be the most robust. Evidence for the measurement properties of tools was found to be patchy and it was noted that responsiveness was often not considered in studies evaluating assessments, even though it is particularly relevant when measuring progress (Hanratty et al. 2015).

A recent systematic review by Provenzani et al. (2019) identified assessments used to measure outcomes and found 327 outcome measures, 69% of which were only used within the literature once. Only seven assessments were used in over 5% of the studies. They also outlined the regular use of non-specific assessments for autism and noted that many of the assessments were not developed as outcome measures.

To the best of our knowledge, no published research has been conducted on the use of assessments by teachers in special education settings for autistic pupils with coexisting intellectual disabilities.

Research Questions

This systematic review was completed to rigorously identify relevant assessment tools and consider their appropriateness for use within special education settings. This review also aimed to synthesise information on the measurement properties of the assessments and present the information in an accessible way (Higgins and Green 2011).

Whilst McConachie et al. (2015) and Hanratty et al. (2015) identified a gap in the research on autism assessment, this review differs in a number of ways. Firstly, this review included assessments measuring adaptive behaviour, problem or challenging behaviour and autism-related behaviour; it is narrower in scope than the review by McConachie et al. (2015) but broader than Hanratty et al. (2015). Secondly, this review extended the age range of previous reviews by including assessments appropriate for school aged children. Finally, this review included assessments devised for individuals with intellectual disabilities, such as may be used in special schools, and considered them in relation to individuals on the autism spectrum with coexisting intellectual disabilities.

This review addressed two primary questions:

1. Which assessment tools can be used by teachers within special education settings to measure adaptive behaviour, problem or challenging behaviour or autism-related behaviour of children with intellectual disabilities?
2. Which of those assessment tools are appropriate for measuring the progress and outcomes of children on the autism spectrum with coexisting intellectual disabilities within a special education setting?

As part of the evaluation of the appropriateness of identified assessments, a secondary aim was to evaluate their measurement properties in order to judge their likely utility.

Methods

A search was conducted for studies which report primary data on the measurement properties of assessments used to measure adaptive behaviour, problem or challenging behaviour or behaviour related to autism. The exact definition of what constitutes 'adaptive behaviour' is unclear and the behaviours measured by different assessments may vary (Kramer et al. 2012). For the purposes of this review, adaptive behaviour assessments are those which focused on assessing functional, applied or generalised skills including independence. To address the fact that there is no clear distinction between measures of 'participation' and adaptive behaviour, measures of participation were included if they were appropriate for a school setting, could be used by teachers and the focus was on skills or abilities relevant to participation as opposed to measuring levels of participation. To ensure all relevant tools were identified, all assessments used with individuals with intellectual disabilities were considered and evaluated in light of their application to pupils on the autism spectrum.

Searches

Searches were conducted of a number of electronic databases using EBSCOhost including Academic Search Complete; British Education Index; ERIC; MEDLINE; PsychArticles; PsychInfo and CINAHL. A separate search was conducted using PubMed. Table 1 shows the key search terms used including combinations, spelling variations and truncation.

The search yielded 3497 results and was repeated with PubMed finding 323 articles. Automatic removal of duplicates resulted in 2397 articles and a hand removal of duplicates left 2270 articles for consideration.

Different combinations of the search terms were also used to search the grey literature using opengrey.eu but no relevant results were found. The above databases were also used to search for assessments commonly used in special schools by name (e.g. Early Years Foundation Stage Profile, P Scales, B Squared) but, again, no relevant results were found.

Eligibility Criteria

Articles were first screened by title and abstract according to the inclusion and exclusion criteria in Table 2. Eligibility was not restricted by year of publication.

Screening

Following the first title and abstract screening, 2196 studies were excluded leaving 74 studies included for a second screening.

The second screening determined whether the assessment met criterion 6 of the exclusion criteria. Criterion 6 excluded assessments which are subject to publisher qualification codes for purchase (e.g. by a qualified psychologist) and are not freely available or able to be purchased, administered and scored and the results interpreted by a qualified teacher. Twenty-eight studies were excluded at this stage because the assessment had a publisher qualification code which required a clinical psychology qualification in order to purchase or use (e.g. Pearson Clinical codes CL1, CL2; WPS publishing Level N) and therefore could not be used by a teacher. Some education systems (e.g. some US states) require a master's degree or further training in SEN in order to teach this population. However, this is not a requirement for SEN teachers in England who can teach in special schools with an ordinary teaching qualification. A number of US-based publishers have intermediary qualification codes which reflect this requirement of further qualification in order to purchase and use specific assessments; these measures were included and their utility will be considered within the discussion.

The full text article was obtained for 46 studies.

Table 1 Searches

Search term blocks combined with AND							
Search terms (OR)	AB(autis* OR "pervasive development* disorder*" OR "learning disab*" OR "learning difficult*" OR "intellectual* impair*" OR "intellectual* disab*" OR "mental* disab*" OR "abnormal development" OR "developmental* disab*" OR retard* OR handicap* OR "special needs" OR "special educational needs")	AB(assess OR assessment)	tool OR instrument OR scale OR checklist OR questionnaire	progress OR progression OR attainment OR skill* OR outcome*	measure OR evaluate OR test	Development OR functioning OR social OR emotional OR communication OR behav* OR motor OR sensory OR math* OR numeracy OR English OR literacy OR language OR interaction OR play	NOT(elderly OR dementia OR concussion OR dizziness OR dyslexia OR midwifery OR "drug therapy" OR "health related quality of life" OR "HRQOL" OR "life satisfaction" OR cancer)

Table 2 Eligibility criteria

Inclusion criteria	Exclusion criteria
1. The focus of the study was to evaluate one or more properties of an assessment tool (either a newly developed or existing assessment tool).	1. The study described the sample as having Asperger's syndrome, 'high-functioning autism', specific learning disabilities (e.g. dyslexia) without coexisting IDs or specified that all participants had an IQ > 70. If a study reported that the average IQ of participants was > 70 (without a range) or if it was not possible to determine whether the participants had IDs, the study was excluded.
2. The assessment was intended for use with and was evaluated using a sample of which some individuals had IDs (IQ < 70), or both autism and IDs. Where general intellectual functioning of the sample could not be determined, the authors were contacted for clarification. Where no response was received, if the study specified that the participants attended a specialist school/education setting or had educational needs potentially unsuitable for a mainstream school (e.g. nonverbal/limited language) it was included.	2. The assessment was evaluated primarily for its properties relating to diagnosis, screening for likely diagnosis, its ability to identify individuals with IDs/autism or predict a future diagnosis.
3. The assessment was evaluated using a sample which overlapped with school age range (≥ 4 years and ≤ 16 years).	3. The assessment tool was developed to be part of or used to assess a specific curriculum or covered the areas of interest but focused on a broad range of developmental outcomes, limiting its generalisability for use in other contexts.
4. The focus of the assessment was to measure outcomes related to adaptive behaviour, challenging behaviour or behaviour related to autism.	4. The assessment was devised to be used specifically by health professionals.
5. The study was a primary quantitative or qualitative study with 5 or more participants.	5. The focus of the assessment is on the participation of the individual rather than their skills, abilities, performance or behaviour.
6. The study was a full text study published in a peer-reviewed journal and published in English.	6. The assessment is subject to publisher qualification codes for purchase (e.g. by a qualified psychologist) and is not freely available or able to be purchased, administered and scored and the results interpreted by a qualified teacher.
	7. The assessment tool was used to measure outcomes for offenders with IDs.
	8. The assessment was used to measure outcomes for a sample with mental health conditions (e.g. depression, anxiety, psychosis, attention deficit hyperactivity disorder, reactive attachment disorder, anger) or, if a sample with coexisting IDs and mental health conditions was used, the mental health condition was the focus of the study.
	9. The assessment was evaluated using a sample specifically with traumatic brain injury.
	10. The assessment was used to assess the quality of life, the function of challenging behaviour or to assess risk.
	11. The assessment tool was developed as an outcome measure for a particular study and was based on observation of characteristics specific to that study (e.g. on task behaviour, parent-child interaction, coded observation) and could not be searched for by name.
	12. The assessment tool was predominantly employment, vocation or training based.
	13. The focus of the study was predominantly to evaluate a translated tool.
	14. The article reported a case study, conference abstract or test review.

On full text screening, 19 studies were excluded. Seven were excluded due to sample age or absence of intellectual disabilities and four were excluded on methodology. A number of authors were contacted for further information about the study. Three studies were excluded for which IQ or intellectual disabilities could not be discerned, which contained no suggestion that the sample required any special education provision and with no response from authors were excluded. Five studies were excluded for not measuring relevant specific domains. The total number of studies included from the search was 27. The search was updated in March and May, 2019, and two additional studies matching the inclusion criteria were

found, bringing the number of included studies from the search to 29.

The first author became aware of an assessment from the wider literature which was particularly relevant for the purposes of this review but was not found through the search. This assessment, the School Function Assessment (SFA), was searched for by name using the same databases and combined with the first block of search terms shown in Table 1 (variations of 'intellectual disabilities'). Two studies were included as per the inclusion criteria.

The ancestry method identified 11 further studies eligible for inclusion. Although including a large number of extra studies through a manual search may signify limiting search terms, eight of these studies commented on assessments already included through the search. In total, 42 articles reporting on 26 assessment tools were included. The search is outlined in Fig. 1.

Risk of Bias and Study Quality

The quality appraisal was guided by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) manual for systematic reviews of patient-reported outcome measures (PROMs). The COSMIN checklist was created for use in systematic reviews to evaluate the risk of bias in studies on measurement properties of PROMs (Prinsen et al. 2018; Mokkink et al. 2018; Terwee et al. 2018). Although this review did not specifically consider PROMs, the COSMIN checklist is a robust and valid evaluation tool for studies reporting on outcome measures and was chosen as a framework to assess both the methodological quality of studies and the reported properties of the assessments in this review. The COSMIN checklist covers aspects of development, validity, reliability and responsiveness with the methodological quality rated separately for each measurement property as 'very good', 'adequate', 'doubtful' or 'inadequate'. The lowest rating in each box is taken to be the overall rating of each measurement property. The COSMIN manual also contains criteria for sufficient measurement properties.

Whilst much of the COSMIN checklist is highly relevant to our appraisal of tools for use in an education setting, this review was concerned with the appropriateness of outcome measures in education settings rather than clinical practice. As suggested by the COSMIN manual, the checklist was adapted for the purposes of this study and therefore some qualifications to the checklist must be made. Doubtful or inadequate ratings were, in some cases, due to missing information where studies did not provide sufficient detail required by the COSMIN checklist for a high rating. As different categories of assessment were considered and many assessments measure slightly different aspects of behaviour, it was more informative for our purposes to consider convergent validity in relation to each hypothesis rather than the criteria specified for criterion validity. Correlations with other measures for comparison were therefore appraised with reference to convergent validity. Similarly, studies which used different versions of assessments with different respondents (e.g. teacher and parent forms) were evaluated with reference to convergent validity as opposed to inter-rater reliability (IRR). Furthermore, studies on different modes of administration (e.g. telephone administration by interview with written reports) were considered relevant and coded with reference to convergent validity. Minimally important change (MIC) is not considered an adequate measure of responsiveness according to the COSMIN checklist so in order to consider MIC for our purposes, as reported in Chatham et al. (2018), this was rated with reference to hypothesis testing. As the purpose of this review is to report on outcome measures, correlations with IQ and diagnostic tools were not considered in light of the evaluations of measurement properties. Although the COSMIN manual suggests addressing scores individually for each subscale, this review reported on assessments' properties overall where possible with comments on subscale results as necessary. The COSMIN manual allows for additional criteria to be used when assessing results from exploratory factor analysis (EFA) and this review used the criteria outlined in McConachie et al. (2015) for sufficient construct validitya: that factors explain > 50% of the variance. The COSMIN manual outlines a way of pooling or summarising results per measurement property, per assessment. Even where assessments had more than one study reporting on their properties, only a small number of studies considered the same measurement properties and these often utilised different versions of the assessment. As further information is unlikely to be provided by summarising the quality of the assessment as a whole, the results, therefore, were reported and discussed separately for each study as well as an overview provided in table format.

Therefore, the COSMIN checklist was used as a guide and the quality ratings of studies, the evaluation of measurement properties of the assessments and the appropriateness of their use by teachers in schools were considered and discussed in the Results section below. Twenty-one percent of the studies were rated by a second blind rater. The ratings were then reviewed and errors were corrected. The inter-rater agreement was 94%, and $k = 0.85$.

Results

Description of Included Studies

In total, 26 assessments were evaluated by 42 included studies and participant numbers ranged from 14 to 9067. Where studies evaluated different versions of an assessment (e.g. parent

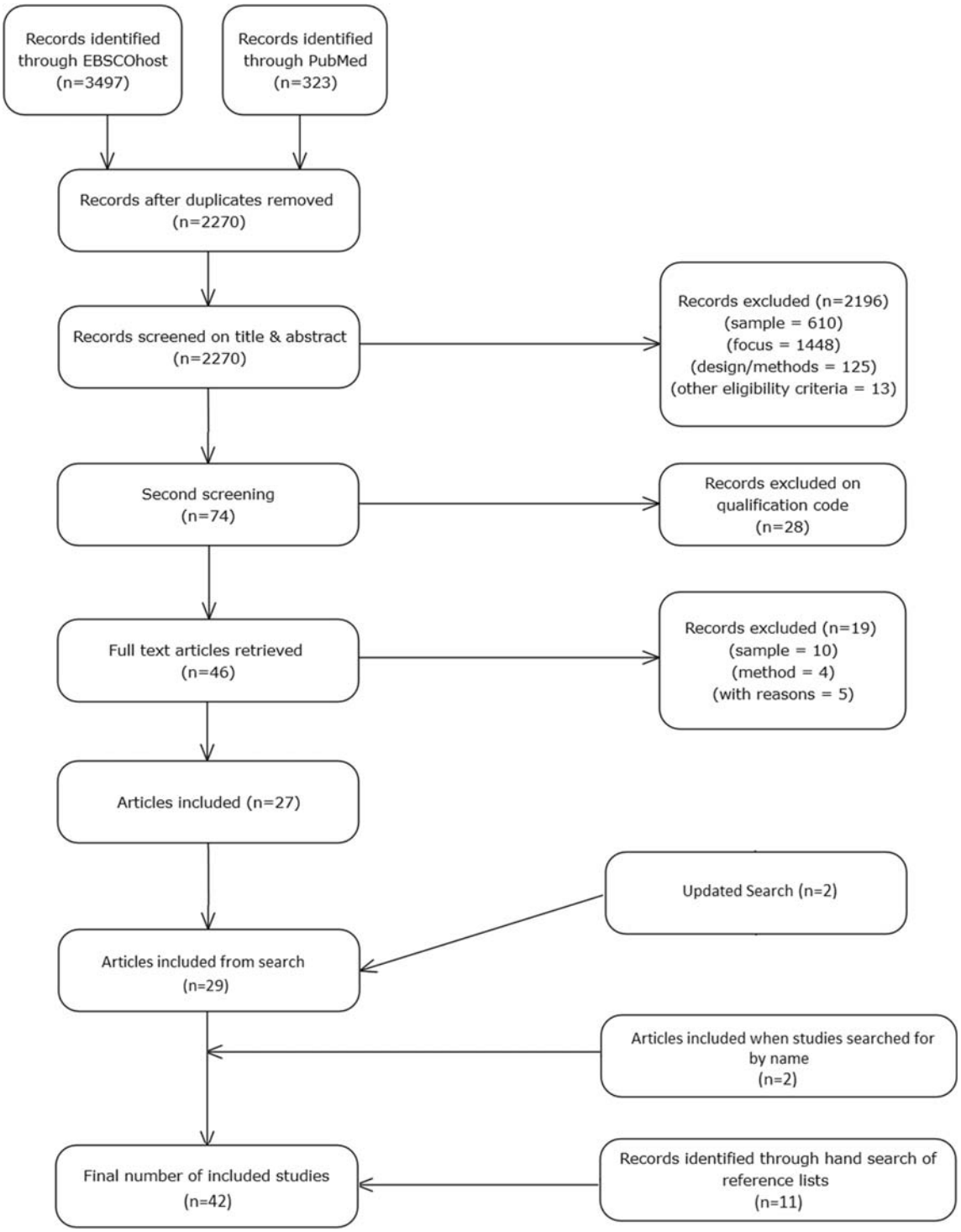

Fig. 1 PRISMA flow diagram

and teacher versions) or used separate samples (e.g. neurotypical sample and sample with intellectual disabilities or adults and children), only data relevant to the eligibility criteria were considered unless comparisons were relevant (e.g. comparisons between parent and teacher responses). The data extracted from the studies are combined in Table 3.

Measurement properties of the majority of the assessments ($n = 16$) were reported in a single study. Seven studies evaluated versions of the Vineland Adaptive Behavior Scales (VABS) (e.g. Charman et al. 2004; Harris et al. 1995). Four studies considered the Aberrant Behavior Checklist (ABC) (e.g. Brown et al. 2002; Kaat et al. 2014; Siegel et al. 2013)

Table 3 Data from included studies

Assessment	Study/year	Country	Category	Sample	Respondents	Assessment type	Newest/ updated version?
Adaptive and Maladaptive/Problem Behaviour Assessments							
AAMD Adaptive Behaviour Scale - Public School Version Part 1	Mayfield, Forman and Nagle 1984	USA	Part 1 Adaptive behaviour (independence and social skills) and Part 2 maladaptive behaviour	31 – intellectual disabilities 7–12 years IQ 55–70	Teachers (SEN and regular), parents and independent researcher	Interview or questionnaire	AAMR ABS- II
AAMD Adaptive Behaviour Scale - School Edition	Perry and Factor 1989	Canada	Part 1 Adaptive behaviour (independence and social skills) and Part 2 maladaptive behaviour	15 – autism 8–18 years IQ 18–82	Parents and group home workers	Interview or questionnaire	AAMR ABS- II
AAMD Adaptive Behaviour Scale - Part 2	Spreat 1982	USA	Part 1 Adaptive behaviour (independence and social skills) and Part 2 maladaptive behaviour	221 – intellectual disabilities 9–57 years IQ 5–76	Direct care technicians (staff)	Interview or questionnaire	AAMR ABS- II
AAMR Adaptive Behavior Scales-School 2nd Edition - Part 1 (revision of AAMD school edition)	Wells et al. 2009	Canada	Part 1 Adaptive behaviour (independence and social skills) and Part 2 maladaptive behaviour	50 – autism and intellectual disabilities 5–-37 years	Staff	Questionnaire or structured interview	Unknown
Behavior Assessment System for Children – 2	Ellison et al. 2016	USA	Adaptive functioning and internalising and externalising behaviours	67–26 months to 18 years IQ 38–128	Teachers and parents	Rating scale	3rd edition
Behavior Assessment System for Children - 2	Lane, Paynter and Sharman 2013	Aus	Adaptive functioning and internalising and externalising behaviours	39 – autism in early intervention 2–5 years	Teachers and parents	Rating scale	3rd edition
Nisonger Child Behavior Rating Form	Aman et al. 1996	USA	Adaptive (social) and problem behaviour	326 – estimated IQ 55–70 3–16 years 260 of these rated by teachers	Teachers and parents	Rating scale	–
PDD-Behavior Inventory	Cohen 2003	USA	Adaptive and maladaptive behaviour	311 – parents 298 – teachers children 1–17 years attending specialist provision	Teachers and parents	Rating scale	–
PDD-Behavior Inventory	Cohen et al. 2003	USA	Adaptive and maladaptive behaviour	84 – autism mild/borderline Griffith GC scores 3–-6 years	Teachers and parents	Rating scale	-
Vineland Adaptive Behavior Scales-II Screener	Charman et al. 2004	UK	Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	125 – 3–6 years autism from schools/units with dedicated autism provision	Caregivers	Caregiver interview	New 3rd version
Vineland Adaptive Behavior Scales-II Survey Interview Form	Chatham et al. 2018	USA/Europe	Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	9067 – autism 28% of sample with intellectual disabilities 1–55 years	Teachers and parents	Semi structured interview/rating scale	New 3rd version
Vineland Adaptive Behavior Scales - Interview Edition, Survey Form	Harris et al. 1995	USA		20 – autism specialised classes	Parents	Semi structured interview/rating scale	New 3rd version

Table 3 (continued)

			Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	2–5 years			
Vineland Adaptive Behavior Scales-II Parent and Teacher Forms	Lane, Paynter and Sharman 2013	Aus	Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	39 – autism in early intervention 2–5 years	Teachers and parents	Rating scale	New 3rd version
Vineland Adaptive Behavior Scales - Survey Form	Perry and Factor 1989	Canada	Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	15 – autism IQ 18–82 8–18 years	Parent or support worker	Semi structured interview	New 3rd version
Vineland Adaptive Behavior Scales - Survey Edition and Classroom Edition	Voelker et al. 2000	Canada	Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	90 – various impairments 3–12 years	Teachers and parents	Survey interview (parents) and rating scale (teachers)	New 3rd version
Vineland Adaptive Behavior Scales-Classroom version	Wells et al. 2009	Canada	Adaptive behaviour (personal and social life skills) with optional Maladaptive Behavior Index	50 – autism and intellectual disabilities 6–30 years	Staff	Rating scale	New 3rd version
Wider Outcomes Survey for Teachers	Wigelsworth et al. 2015	UK	Wider outcomes with behaviour difficulties, relationships, exposure to bullying	6164 – SEND, some statements, all in receipt of SEND provision 5–6 years, 9–10 years, 11–12 years, 14–15 years	Teachers	Questionnaire	–
Adaptive Behaviour Assessments							
Adaptive Behavior Assessment System-II	Aricak and Oakland 2010	USA	Adaptive behaviour	1690 – TD, intellectual disabilities, autism, etc. 2 age groups (5–12 years and 13–21 years)	Teachers or family members	Rating scale	ABAS III
Behavior Rating Inventory for the Retarded	Sparrow and Cicchetti 1978	USA?	Adaptive behaviour	45 – intellectual disabilities 6–15 years IQ 2–53	Staff	Interview	Unknown
Children's Adaptive Behavior Scales	Kicklighter and Bailey 1980	USA	Adaptive behaviour	60 – IQ 70–89 60 – IQ 50–69 all 7–10 years	No mention	Direct assessment	Unknown
Great Outcomes for Kids Impacted by Severe Developmental Disabilities-Brief Adaptive Scale	Pan et al. 2019	UK	Adaptive skills/behaviour	361 - SLD school 38% autism 4–19 years	Caregivers/teachers	Report scale	Yes
Great Outcomes for Kids Impacted by Severe Developmental Disabilities-Brief Adaptive Scale	Perry et al. 2015	Canada	Adaptive skills/behaviour	432 parents of children with intellectual disabilities 3–20 years	Parents	Report scale	Yes
Independent Behaviour Assessment Scale	Munir et al. 1999	Bangladesh	Adaptive behaviour	Stage 1: 154 1–10 years Stage 2: 1404 – typically developing	Trained psychologists conducted assessment	Direct observation supplemented by interview	Unknown

Table 3 (continued)

				Third sample of 22 2–9 years with intellectual disabilities			
Minnesota Developmental Programming System Behavioural Scales-Alternate Form C	Silverman et al. 1983	USA	Adaptive behaviour	3487 - various levels of intellectual disabilities 0–70 years	No info	Rating scale	Unknown
Pediatric Evaluation of Disability Inventory-Computer Adaptive Test	Dumas et al. 2010	USA	Adaptive behaviour (personal self-sufficiency and engagement in community)	Focus group - 6 parents of children with disabilities Focus groups - 4-5 hospital clinicians Cognitive Interviews - 11 parents of children with disabilities (32 interviews)	Parents/clinicians	Computer adaptive test	–
Pediatric Evaluation of Disability Inventory-Computer Adaptive Test	Kramer et al. 2012	USA	Adaptive behaviour (personal self-sufficiency and engagement in community)	20 autism professionals (incl. 2 teachers) 18 parents of 21 autistic children, 3–17 years (varying communication)	Professionals/parents	Computer adaptive test	–
Pediatric Evaluation of Disability Inventory-Patient Reported Outcome	Kramer and Schwartz 2018	USA	Adaptive behaviour (functional performance of functional tasks)	Phase I: 8–14-21 years (incl.6 with intellectual disabilities) Phase II: 43–15-21 years with developmental disabilities incl. some with autism	Self-report scale	Self-report	Continuing development
Pediatric Evaluation of Disability Inventory-Patient Reported Outcome	Kramer and Schwartz 2017	USA	Adaptive behaviour (functional performance of functional tasks)	37–68% with intellectual disabilities, some with autism 8–21 years	Self-report	Self-report	Continuing development
School Function Assessment	Coster, Mancini and Ludlow 1999	USA	Adaptive behaviour (performance of functional tasks/skills)	266 – typically developing and varying needs 5–14 years	Teacher or therapist	Questionnaire	–
School Function Assessment	Davies et al. 2004	USA	Adaptive behaviour (performance of functional tasks/skills)	11 – autism	Teachers	Questionnaire	–

Table 3 (continued)

				15 - learning disabilities 6–11 years			
School Function Assessment	Hwang et al. 2002	USA	Adaptive behaviour (performance of functional tasks/skills)	29 – typically developing 18 – learning disabilities 17 – cerebral palsy 6–14 years	Teachers	Questionnaire	–
Street Survival Skills Questionnaire	Janniro, Sapp and Kohler 1994	USA	Adaptive behaviour	18–9-17 years Mean IQ = 45	No info	Direct Assessment	–
Challenging/Problem Behaviour Assessments							
Aberrant Behavior Checklist	Brown, Aman and Havercamp 2002	USA	Problem behaviour	601 – intellectual disabilities 6–22 years	Parents	Checklist	2nd Edition
Aberrant Behavior Checklist	Kaat, Lecavalier and Aman 2014	USA and Canada	Problem behaviour	1130 (calibration sample) 763 (validation sample) autism 47% intellectual disabilities 2–18 years	Parents	Checklist	2nd Edition
Aberrant Behavior Checklist	Marshburn and Aman 1992	USA	Problem behaviour	666 – intellectual disabilities 6–21 years	Teachers	Checklist	2nd Edition
Aberrant Behavior Checklist - Irritability Subscale	Siegel et al. 2013	USA	Irritability	14 – autism and/or intellectual disabilities 5–18 years	Parents	Telephone administration and checklist	2nd Edition
Challenging Behaviour Interview	Oliver et al. 2003	UK	Severity of challenging behaviour	47 – severe intellectual disabilities 4–12 years	Teachers for child sample	Interview	–
Developmentally Delayed Children's Behaviour Checklist	Einfeld and Tonge 1991	Aus	Behavioural and emotional problems	5000 – range of intellectual disabilities unknown age	Parents, care workers, teachers, teacher aides	Questionnaire/rating scale	Unknown
Home Situations Questionnaire-PDD	Chowdhury et al. 2010	USA	Severity of disruptive behaviour and behavioural non compliance	124 – PDD/autism IQ > 35, some in SEN education 4–13 years	Parents	Rating scale	Also HSQ-AS-D
Assessments of Behaviour Related to Autism							
Autism Behavior Inventory	Bangerter et al. 2017	USA	Core/associated symptoms in autism	Phase 1: 353 – autism – varying communication 3–21 yearsPhase 2: 23 – ASD 3- > 18 years	Caregivers	Rating scales (web based)	–
Autism Impact Measure	Kanne et al. 2014	USA	Core autism symptoms	440 – autism, IQ 45–148 2–17 years	Caregivers	Rating scale	–
Autism Treatment Evaluation Checklist	Charman et al. 2004	UK	Autism symptoms	Cohort 1–79 Cohort 2–55 (134 total)	Parents	Short checklist	–

Table 3 (continued)

Autism Treatment Evaluation Checklist	Magiati et al. 2011	UK	Autism symptoms	57 - 1 year follow-up pupils from autism provision < 6 years 22 parents of children, autism, various IQs, early intervention FU1–2 years later (age 4–6 years) FU2–5-6 years later (age 9–12 years)	Parents and teachers	Short checklist	–
Repetitive Behavior Scale-Revised	Lam and Aman 2007	USA	Severity of restricted and repetitive behaviours	307 – autism, varying severity, some at SEN school 3–48 years	Caregivers and staff	Rating scale	
Teacher Autism Progress Scale	Dang et al. 2017	USA	Outcomes for children on autism spectrum	30 – autism 6–24 years	Teachers	Rating scale	–
Sensory Behaviour Questionnaire	Neil, Green and Pellicano 2017	UK	Sensory behaviours for moderate-severe intellectual disabilities	66 – autism 70 – typically developing 6–17 years IQ 57–130	Parents	Questionnaire	–

Available?	Appropriate for teachers?	Appropriate for school?	Intended population and purpose
Adaptive and Maladaptive/Problem Behaviour Assessments			
Unknown	Yes	School version evaluated with teachers	
Unknown	Yes	School edition evaluated	
Unknown	Yes	No mention of teachers/use in schools	
Abledata out of print 2016	Yes	School version evaluated	Intellectual disabilities/developmental disabilities/autism Screening and evaluating skills
Pearson clinical	Level B	Teacher rating scales evaluated	Various Intellectual disabilities/developmental disabilities Screening/diagnosis and evaluating/measuring behaviour
Pearson clinical	Level B	Teacher and parent version evaluated	Various Intellectual disabilities/developmental disabilities Screening/diagnosis and evaluating/measuring behaviour
Available online	Yes	Teachers evaluated	Intellectual disabilities Assessing behaviour and emotional problems
WPS publishing /PAR Inc.	Yes	Teacher and parent version evaluated	PDD Outcome measure for responsiveness to intervention
WPS publishing /PAR Inc.	Yes	Teacher and parent version evaluated	PDD Outcome measure for responsiveness to intervention
Pearson	Teacher version but code B	Teacher version but code B	SEN populations Support diagnosis and intervention planning and tracking progress
Pearson	Teacher version but code B	Teacher version but code B	
Pearson	Teacher version but code B	Teacher version but code B	
Pearson	Teacher (and parent) version evaluated	Teacher (and parent) version evaluated	
Pearson	Teacher version but code B	Teacher version but code B	
Pearson	Teacher (and parent) version evaluated	Teacher (and parent) version evaluated	
Pearson	Classroom version used	Classroom version used	
Article	Yes	Teachers evaluated	SEND

Table 3 (continued)

			outcome measure
Adaptive Behaviour Assessments			
Pearson	CLR2	School version/teacher form used	Useful for intellectual disabilities Screening and intervention planning
Unknown	No mention of teachers	No mention of school	intellectual disabilities Measuring behaviour
Not in print? Publishers 1920–2008	Mention of teachers and assessment being 'educationally useful'	Mention of teachers and assessment being 'educationally useful'	intellectual disabilities/developmental disabilities Screening and measuring behaviour
Yes	Yes	Yes	SLD Research purposes, not recommended for clinical use
Yes	Suggests not	Suggests not	SLD Research purposes, not recommended for clinical use
Author	Mentions teacher manual	Mentions teacher manual - also mentions training required	Screening/identification of disabilities
Article	No mention	No mention	Severe intellectual disabilities Diagnostic and intervention planning
Website	Yes	Yes	Identifying delay, assessing improvement, monitoring and evaluating interventions Autism module in development
Website	Yes	Yes	Identifying delay, assessing improvement, monitoring and evaluating interventionsAutism module in development
Continuing development	n/a	Mention of schools	Disabilities Planning and evaluating outcomes of education and healthcare
Continuing development	n/a	Mention of schools	Disabilities Planning and evaluating outcomes of education and healthcare
Pearson	Yes	Yes	Disabilities Screening, planning, identifying and reporting progress
Pearson	Yes	Yes	Disabilities Screening, planning, identifying and reporting progress
Pearson	Yes	Yes	Disabilities Screening, planning, identifying and reporting progress
McCarron Dial Systems	No mention	No mention but participants noted as students attending public schools	Intellectual disabilities Screening, diagnosis and intervention planning
Challenging/Problem Behaviour Assessments			
Stoeltling	Has been evaluated with teachers	Has been evaluated with teachers	Intellectual disabilities Outcome Measure
Stoeltling	Has been evaluated with teachers	Has been evaluated with teachers	Intellectual disabilities Outcome Measure
Stoeltling	Teachers used as respondents	Teachers used as respondents	Intellectual disabilities Outcome Measure
Stoeltling	Has been evaluated with teachers	Has been evaluated with teachers	Intellectual disabilities Outcome Measure
Author	Teacher respondents - suggests training to administer	Teachers respondents - suggests training to administer	Intellectual disabilities Outcome measure of challenging behaviour
Unknown	Teacher version but level C	Teacher version evaluated	Intellectual disabilities/developmental disabilities Supports clinical assessment and progress
Article	Items relevant to home but some may be relevant to school	School Situations Questionnaire available	PDD Outcome measure
Assessments of Behaviour Related to Autism			
Article and JanssenMD	No mention of teachers/use in schools	No mention of teachers/use in schools	Autism Designed to assess and capture change
Soon to be available from WPS	No mention of teachers/use in schools	No mention of teachers/schools but author correspondence suggested may be useful Teacher version in development	Autism Outcome measure
Autism Research Institute website	Yes	Completed by parents, teachers or caregivers	Autism Outcome measure
Autism Research Institute website	Yes	Teachers evaluated	Autism Outcome measure
Unknown	Teacher respondents used in studies elsewhere	Teacher respondents used in studies elsewhere	Autism Assessing a variety of RRBs
Article	Yes	Designed for and evaluated with teachers	Autism Outcome measure
Freely available	May be 'useful' for educators	May be 'useful' for educators	Autism Frequency and impact of sensory behaviours

and current or previous versions of AAMR Adaptive Behavior Scale-II (AAMR ABS-II) (e.g. Mayfield et al. 1984; Spreat 1982; Wells, Condillac, Perry and Factor 1989). Three studies evaluated the SFA (Coster et al. 1999; Davies et al. 2004; Hwang et al. 2002) and six assessments had two studies evaluating them (e.g. PDD-Behavior Inventory; Autism Treatment Evaluation Checklist). Four studies (Charman et al. 2004; Lane et al. 2013; Perry and Factor 1989; Wells et al. 2009) reported on more than one assessment.

The focus of the included assessments fell into four categories: adaptive behaviour including adaptive functioning ($n = 10$), problem or challenging behaviour ($n = 4$), autism-related behaviour ($n = 6$) and both adaptive and problem or maladaptive behaviour ($n = 6$).

Even though studies were only included if the assessment was evaluated as an outcome measure, the Great Outcomes for Kids Impacted by Severe Developmental Disabilities-Brief Adaptive Scale (GO4KIDDS) was developed specifically for research purposes (Perry et al. 2015) and the Independent Behaviour Assessment Scale (IBAS) was developed for diagnostic or screening purposes (Munir et al. 1999). Ten further assessments were reported to be useful for both diagnosis or screening and as outcome measures (e.g. Adaptive Behavior Assessment System-II; VABS-II), whilst the remaining 14 assessments were developed specifically as outcome measures (e.g. Autism Treatment Evaluation Checklist; Challenging Behaviour Interview; Teacher Autism Progress Scale).

Domains

The 10 adaptive behaviour assessments considered a number of different areas of functioning including social skills, communication, independence or self-help and physical skills. The Pediatric Evaluation of Disability Inventory-Computer Adaptive Test (PEDI-CAT), Pediatric Evaluation of Disability Inventory-Patient Reported Outcome (PEDI-PRO) and SFA, although addressing elements of participation, focused on functional skills of children in schools and therefore were included in the adaptive behaviour category. The Minnesota Developmental Programming System Behavioural Scales–Alternate Form C (MDPS–C) also included a domain labelled *eating behaviours* (Silverman et al. 1983). The Street Survival Skills Questionnaire (SSSQ) included skills relevant to teenagers and adolescents, for example *health and safety*, *public services* and *time*, *money and measurement* (Janniro et al. 1994).

Four assessments focused on problem or challenging behaviour: the ABC (e.g. Brown et al. 2002; Marshburn and Aman 1992), Challenging Behaviour Interview (CBI) (Oliver et al. 2003), Developmentally Delayed Children's Behaviour Checklist (DDCBCL) (Einfeld and Tonge 1991) and the HSQ–PDD (Chowdhury et al. 2010). These assessments measured behaviour such as physical aggression, stereotypic behaviours and non-compliance. The HSQ–PDD and CBI both addressed aspects of severity of problem behaviour (Chowdhury et al. 2010; Oliver et al. 2003). The DDCBCL yielded scores relating to deviant behaviour, distress to carers and impairment of adaptive functioning (Einfeld and Tonge 1991).

Six assessments considered both adaptive and maladaptive or problem behaviour including the AAMR ABS-II (Wells et al. 2009 plus previous versions from Mayfield et al. 1984; Perry and Factor 1989; Spreat 1982), Behavior Assessment System for Children-2 (BASC-2) (Ellison et al. 2016; Lane et al. 2013), Pervasive Developmental Disorder-Behavior Inventory (PDD-BI) (Cohen et al. 2003; Cohen 2003), VABS (e.g. Harris et al. 1995; Perry and Factor 1989; Voelker et al. 2000), Nisonger Child Behavior Rating Form (NCBRF) (Aman et al. 1996) and the Wider Outcomes Survey for Teachers (WOST) (Wigelsworth et al. 2015). Assessments such as the BASC-2 and VABS considered a wide variety of adaptive and maladaptive behaviour. The PDD-BI was developed specifically for use with children on the autism spectrum and included both autism-specific and broader, more generic skills and behaviours (Cohen et al. 2003). The WOST assessed behaviour difficulties, social relationships and experiences of bullying.

Six assessments measured autism-related behaviour: the Autism Behavior Inventory (ABI) (Bangerter et al. 2017), Autism Impact Measure (AIM) (Kanne et al. 2014), Autism Treatment Evaluation Checklist (ATEC) (Charman et al. 2004; Magiati et al. 2011), Repetitive Behavior Scale-Revised (RBS-R) (Lam and Aman 2007), Teacher Autism Progress Scale (TAPS) (Dang et al. 2017) and the Sensory Behaviour Questionnaire (SBQ) (Neil et al. 2017). The ABI, AIM, ATEC and the TAPS all considered the assessments' abilities to capture progress and change (Bangerter et al. 2017; Charman et al. 2004; Dang et al. 2017; Kanne et al. 2014). The RBS-R was devised specifically to assess RRBs of autistic individuals and also suggested potential usefulness in measuring intervention outcomes (Lam and Aman 2007). The SBQ assessed the frequency and impact of 25 different sensory behaviours and Neil et al. (2017) suggested it may also be useful in measuring outcomes.

Samples

Diagnosis

As per the inclusion criteria, all studies included at least some participants with intellectual disabilities. Eight assessments were devised specifically for autism including pervasive developmental disorders (e.g. PDD-BI; AIM; ATEC). A number of other assessments were described as appropriate for a

variety of intellectual disabilities or developmental disabilities including autism (e.g. BASC-II; NCBRF; VABS-II).

The numbers or percentage of participants with intellectual disabilities or who were on the autism spectrum varied; in two studies all or nearly all participants had intellectual disabilities and coexisting autism (e.g. Wells et al. 2009), while in other studies only some of the sample had intellectual disabilities (e.g. Ellison et al. 2016). Some studies reported the Full-Scale IQ of participants but nine studies were included on the basis that it was described or inferred that some or all of the participants needed educational provision above that which could be provided by a mainstream school (e.g. Bangerter et al. 2017; Hwang et al. 2002; Wigelsworth et al. 2015). Some studies specified the numbers of participants with each diagnosis (e.g. Chatham et al. 2018), whilst others did not provide the exact number of participants with intellectual disabilities within their sample (e.g. Kanne et al. 2014). The samples will be taken into account when discussing the appropriateness of the assessments in the Discussion section below.

Age

Five studies used samples 6 years old or younger (e.g. Charman et al. 2004; Cohen 2003), whilst three studies used primary school aged samples (e.g. Aricak and Oakland 2010; Davies et al. 2004) and Munir et al. (1999) involved participants aged 2–9 years. Most other studies used samples of children and adolescents spanning school age ($n = 12$), children up to age 18 ($n = 7$) or a broad age range that included children and adults ($n = 14$). Where samples were split into children and adults (e.g. Oliver et al. 2003), only results from the child sample were considered in this review.

Methods of Assessment

As would be expected, assessment methods of the included measures varied. This included direct assessment or observation (e.g. Children's Adaptive Behavior Scale; SSSQ), interviews with parents, caregivers or teachers (e.g. CBI; VABS-II) or a mixture of methods (e.g. observation and interview in the IBAS). The majority of measures were rating scales, checklists or questionnaires filled out by professionals or parents (e.g. ABC; PDD-BI; RBS-R) with only one self-report measure (PEDI-PRO).

Use by Teachers in Educational Settings

Sixteen assessments were either designed specifically for use by teachers, developed for use in schools or evaluated using teacher respondents in the included studies. The HSQ–PDD has a school form of its original version available but studies reporting on this version were not found in the systematic search. A teacher version of the AIM is in development (S. Kanne, personal communication, December 2018). Studies which evaluated the Children's Adaptive Behavior Scales (CABS) and the IBAS did not use teacher respondents; however, Kicklighter and Bailey (1980) suggested that the CABS may be 'educationally useful' (p.169) and Munir et al. (1999) mentioned a 'teacher's manual', although there was also a suggestion that those administering the assessment in the study needed 'extensive training' (p.246). Studies which discussed seven of the assessments did not mention or imply the possibility of use in schools or by teachers (e.g. ABI; MDPS–C). Whilst this does not mean that these assessments may not be useful with teachers, it is more likely that these were designed for clinical use.

Fourteen assessments were evaluated using teacher respondents; however, four of these are subject to intermediary qualification codes upon purchase requiring teachers to have a master's degree or further qualification in assessment which may restrict access or use by ordinary SEN teachers (VABS-II; ABAS-II; BASC-2; SFA). The implications of these assessments subject to intermediary qualification codes will be considered in the discussion on the utility of the assessments by teachers in schools.

Availability and Year of Study

As no date limit was specified for inclusion, 13 studies (30%) were conducted prior to 2000. Some of these studies may have used methods which have since been revised and updated and this must be a consideration when judging the evidence and potential uses of these assessments. In addition, older assessments may not comprehensively address adaptive behaviour involving modern technologies (Floyd et al. 2015). Furthermore, current information on some assessments proved difficult to find. Two assessments were out of print or appeared unavailable from publishers (AAMR ABS-II; CABS). Four assessments have more recent or updated versions than those considered in the included studies (ABAS-III; ABC-2; BASC-3; VABS-III). Eight assessments (or their most recent versions) are available from publishers (ABAS-II; ABC-2; ABI; BASC-3; PDD-BI; SFA; SSSQ; VABS-III). The ATEC is available from the Autism Research Institute website and the PEDI-CAT and PEDI-PRO are available directly from their respective websites or universities, although the PEDI-PRO is still under continuing development. Nine assessments were included in the article or suggest they are available from the author (CBI; GO4KIDDS; HSQ–PDD; IBAS; MDPS–C; NCBRF; SBQ; TAPS; WOST). No information could be found in regard to accessing three of the assessments and/or there were no replies when the authors were contacted (Behavior Rating Inventory; DDCBCL; RBS-R). The AIM is soon to be available for purchase.

Measurement Properties and Quality Assessment

Content Validity

The COSMIN manual considers content validity to be 'the most important measurement property' (Prinsen et al. 2018; Mokkink et al. 2018; Terwee et al. 2018; p.36) and it is relevant for the purposes of this review that assessments used by teachers were developed with teachers' input. Many studies briefly described content validity when discussing the test development (e.g. Aman et al. 1996; Kanne et al. 2014) but only four studies discussed assessment development or content validity in sufficient detail to be rated here (Dumas et al. 2010; Kramer et al. 2012; Kramer and Schwartz 2017; Kramer and Schwartz 2018).

Kramer et al. (2012) reported sufficient content validity of the PEDI-CAT with very good and adequate quality evidence. Dumas et al. (2010) also found sufficient content validity during assessment development; however, the methodological quality of the comprehensiveness study was rated inadequate due to lack of information on data coding methods, whilst the relevance and comprehensibility studies were rated as doubtful due to unclear data analysis procedures and for not including a range of professionals in the sample. Kramer and Schwartz (2017) and (2018) showed sufficient content validity of the PEDI-PRO but the evidence was rated doubtful due to lack of information on the skill or experience of the moderators.

Assessments with Sufficient Measurement Properties

A number of assessments showed sufficient evidence over a number of measurement properties. The HSQ–PDD had good responsiveness and good internal consistency with very good quality ratings (Chowdhury et al. 2010). Construct validity was sufficient but with evidence rated inadequate due to unsatisfactory sample size. Correlations with subscales of the ABC were modest to moderate and significant. Correlations with VABS subscales were inverse but non-significant (Chowdhury et al. 2010).

Wigelsworth et al. (2015) showed very good quality evidence of sufficient internal consistency of the WOST. Structural validity was insufficient according to the COSMIN criteria even though the model was close to ideal fit (CFI = 0.858, TLI = 0.838).

GO4KIDDS showed very good quality evidence of internal consistency and convergent validity with the VABS and Scales of Independent Behaviour-Revised (Pan et al. 2019; Perry et al. 2015). Pan et al. (2019) found one principal component measuring adaptive behaviour with adequate quality evidence.

Magiati et al. (2011) provided very good quality evidence of sufficient internal consistency in all subscales of the ATEC, initially and on both follow-up periods. ATEC total scores significantly correlated with the ADI-R total raw score and inversely with the VABS Composite age equivalent score at both follow-up periods. These correlations, however, became insufficient by COSMIN standards when controlling for IQ. Total ATEC scores remained stable over time with large individual differences; however, the methodological quality for responsiveness was rated as inadequate. Charman et al. (2004) also evaluated responsiveness and reported change in one of the three ATEC subscales but this evidence was also rated inadequate.

The TAPS measured improvements over time as compared with the ABC and Social Responsiveness Scale with adequate evidence (Dang et al. 2017). Only three other studies evaluated responsiveness of assessments. Charman et al. (2004) did not find significant change in Adaptive Behaviour Composite Score of the VABS-screener over time, whilst Harris et al. (1995) showed significant change of the VABS-Survey Interview Form at the first follow-up but not the second. The methodological quality of these two evaluations was rated inadequate due to the COSMIN manual considering paired *t* tests an inappropriate measure of responsiveness.

The PDD-BI showed varied data on IRR and convergent validity; teacher IRR across subscales ranged from moderate to high (range 0.55 to 0.93) and was more strongly correlated than parent-teacher IRR. Test-retest reliability for the teacher scale was sufficient with very good quality evidence (range 0.73 to 0.97). Convergent validity with the Childhood Autism Rating Scale was moderate but significant (0.50), NCBRF was low to moderate (range 0.16 to 0.66) and VABS subscales significant with a range from 0.31 to 0.81.

Lam and Aman (2007) provided very good quality evidence of sufficient internal consistency of the RBS-R. IRR for the different subscales ranged from 0.57 to 0.73 for the younger sample and − 0.24 to 0.95 for the older sample. A five-factor solution for the RBS-R was adopted from the EFA which accounted for 47.5% of the variance, below the cutoff of 50% for good structural validity. Adequate evidence was found of a close to ideal fit with a root mean square error of approximation (RMSEA) of 0.061, just outside of the COSMIN level for sufficient structural validity.

The SFA showed moderate to good convergent validity with the VABS-Classroom version for the learning disabilities group with very good quality evidence (Hwang et al. 2002). Davies et al. (2004) found sufficient IRR between teachers and therapists for only two of the three scales. A two-factor solution was indicated by Coster et al. (1999) although this evidence was of inadequate quality due to a small sample size.

Kaat et al. (2014) provided very good quality evidence for sufficient internal consistency of the ABC and appropriate convergent and divergent validity with the CBCL and VABS. Construct validity varied across studies. An EFA by Marshburn and Aman (1992) found that a four-factor solution

accounted for 52% of the variance. A confirmatory factor analysis (CFA) of a five-factor solution by Brown et al. (2002) yielded a sufficient RMSEA according to COSMIN criteria (< 0.06); however, this threshold was not reached by Kaat et al. (2014). Siegel et al. (2013) found very good quality evidence for no significant difference between written and telephone administrated ABC scores.

The VABS showed reasonable convergent validity with the AAMD Adaptive Behavior Scale (AAMD ABS) (Perry and Factor 1989) and a significant moderate inverse correlation with the CARS (Wells et al. 2009).

Teacher and Parent Rating Scales

Ratings of parents and teachers on the BASC-2 were significant on the externalising composite (Lane et al. 2013) but with adaptive skills rated significantly lower by parents than teachers on the composite and adaptive subscales (Ellison et al. 2016; Lane et al. 2013). Lane et al. (2013) showed that parent-teacher correlations on VABS-II domains were all significant with no significant differences. These studies were of very good methodological quality. Voelker et al. (2000) compared parent-teacher ratings on the VABS and found that correlations were high for the summary score and all domains apart from the socialisation domain. Very good quality evidence showed, again, that parents consistently and significantly reported lower adaptive behaviour skills than teachers. However, when the 169 overlapping items from the VABS classroom and survey form were analysed for IRR, parents reported higher skill level on 70% of comparisons with 93% of correlations significant. This evidence was considered indeterminate and of inadequate quality due to the use of the phi correlation coefficient. Aman et al. (1996) considered teacher and parent agreement on the NCBRF. They found that correlations were significant but ranged from 0.22 to 0.54, indicating differences between teacher and parent ratings of a child's adaptive skills or problem behaviour.

Other Assessments

Twenty-one studies only reported on one measurement property for the sample of interest, and for seven of the 26 assessments there was information on only one measurement property from only one study (e.g. ABAS-II; CABS; IBAS; MDPS–C). Most other assessments had three or more measurement properties evaluated. Neil et al. (2017) found good internal consistency for the SBQ and convergent validity with the Short Sensory Profile. Spreat (1982) found no significant differences between weighted and non-weighted items on previous versions of the AAMD ABS and Chatham et al. (2018) estimated minimal clinically important differences of the Composite Score of the VABS to be 2–2.5 points for the relevant sample.

Of the studies conducted before the year 2000, it was noted that six studies had inadequate ratings for one or more measurement property (Aman et al. 1996; Coster et al. 1999; Kicklighter and Bailey 1980; Harris et al. 1995; Mayfield et al. 1984; Sparrow and Cicchetti 1978). Weaknesses shown in older scales may be an indication of progress made in scale development and validation over time and/or improved reporting within peer-reviewed studies (Floyd et al. 2015). The evaluation of measurement properties and the quality assessment of each study are summarised in Table 4.

Discussion

Twenty-six assessments were found with potential for use in school settings to measure progress of adaptive behaviour, problem or challenging behaviour or autism-related behaviour of children on the autism spectrum with coexisting intellectual disabilities. When considering the appropriateness of these assessments for use by teachers in special schools, there are a number of factors that need consideration: (a) the purpose of the assessment, (b) the usability of the assessment, e.g. whether consideration of use by teachers had been made during development, (c) the applicability of use alongside the school curriculum and (d) the measurement properties of the assessment.

Assessment Purpose and Intended Population

It is necessary to take into account the original purpose for which the assessment was developed when considering the appropriateness of an assessment in a specific context. Even though use as an outcome measure was necessary for inclusion in this review, 42% ($n = 11$) of the included assessments also support screening or diagnosis with one further assessment developed for research purposes. Assessments which either attempt to serve multiple purposes or are used for purposes for which they were not intended may be less effective at measuring for a specific purpose. It must not be assumed that 'an assessment is appropriate and interpretable for a particular context of use without determining if there is evidence regarding the validity of such assumptions within the context' (Pellegrino 2014, p.68). Similarly, evaluations of an assessment for one purpose are not necessarily generalisable to the use of the assessment for other purposes (Haynes et al. 1995). As an example, the authors of GO4KIDDS specified that it was not recommended for contexts other than research (Perry et al. 2015) and, although Pan et al. (2019) found some initial promise for its use by teachers in special schools, further validation would be needed before it could be considered an appropriate measure for use in schools (McConachie et al. 2015). Assessments developed specifically to measure outcomes and progress (e.g. AIM; ATEC; CBI; PDD-BI; RBS-

Table 4 Risk of bias and measurement properties

Assessment	Author	Content validity				Construct validity		Reliability			
		Development	Parents/experts								
		Prom design	Relevance	Comprehensiveness	Comprehensibility	Structural validity	Hypothesis testing (convergent)	Internal consistency	Inter-rater reliability	Test-retest reliability	Responsiveness
Adaptive and Maladaptive/Problem Behaviour Assessments											
AAMD Adaptive Behaviour Scale - Public School Version Part 1	Mayfield et al. 1984								Inadequate -	Inadequate +	
AAMD Adaptive Behaviour Scale - School Edition	Perry and Factor 1989						Very good +			Doubtful +	
AAMD Adaptive Behaviour Scale - Part 2	Spreat 1982						Adequate +				
AAMR Adaptive Behavior Scales-School 2nd Edition Part 1 (revision of AAMD school edition)	Wells et al. 2009						Adequate +				
Behavior Assessment System for Children −2	Ellison et al. 2016						Very good +				
Behavior Assessment System for Children - 2	Lane et al. 2013						Very Good +				
Nisonger Child Behavior Rating Form	Aman et al. 1996					Inadequate -	Adequate +		Inadequate -		
PDD-Behavior Inventory	Cohen 2003					Inadequate ?		Very Good +	Adequate +		
PDD-Behavior Inventory	Cohen et al. 2003						Adequate +				
Vineland Adaptive Behavior Scales II - Screener	Charman et al. 2004										Inadequate +
Vineland Adaptive Behavior Scales II - Survey Interview Form	Chatham et al. 2018						Very good +				
Vineland Adaptive Behavior Scales - Interview Edition, Survey Form	Harris et al. 1995										Inadequate +
Vineland Adaptive Behavior Scales II -Parent and Teacher Forms	Lane, Paynter and Sharman 2013						Very good +				
Vineland Adaptive Behavior Scales -Survey Form	Perry and Factor 1989						Adequate +				

Table 4 (continued)

Assessment	Author	Content validity				Construct validity		Reliability			Responsiveness
		Development	Parents/experts								
		Prom design	Relevance	Comprehensiveness	Comprehensibility	Structural validity	Hypothesis testing (convergent)	Internal consistency	Inter-rater reliability	Test-retest reliability	
Vineland Adaptive Behavior Scales - Classroom version	Wells et al. 2009						Adequate +				
Vineland Adaptive Behavior Scales - Survey Edition and Classroom Edition	Voelker et al. 2000						Very good +		Inadequate ?		
Wider Outcomes Survey for Teachers	Wigelsworth, Oldfield and Humphrey 2015					Very good -		Very good +			
Adaptive Behaviour Assessments											
Adaptive Behavior Assessment System	Aricak and Oakland 2010					Very good -					
Behavior Rating Inventory for the Retarded	Sparrow and Cicchetti 1978					Inadequate +	Inadequate +		Adequate +		
Children's Adaptive Behavior Scales	Kicklighter and Bailey 1980							Inadequate total score +			
Great Outcomes for Kids Impacted by Severe Developmental Disabilities-Brief Adaptive Scale	Pan et al. 2019					Adequate +	Very good +	Very good +			
Great Outcomes for Kids Impacted by Severe Developmental Disabilities-Brief Adaptive Scale	Perry et al. 2015						Very Good +	Very good +			
Independent Behaviour Assessment Scale	Munir et al. 1999							Very good +			
Minnesota Developmental Programming System Behavioural Scales-Alternate Form C	Silverman et al. 1983					Adequate ?					
Pediatric Evaluation of Disability Inventory-Computer Adaptive Test	Dumas et al. 2010	Relevance Doubtful + Comprehensiveness Inadequate - Comprehensibility Doubtful +									
Pediatric Evaluation of Disability	Kramer et al. 2012		Adequate +	Adequate +	Adequate +						

Table 4 (continued)

Assessment	Author	Content validity				Construct validity		Reliability			Responsiveness
		Development	Parents/experts								
		Prom design	Relevance	Comprehensiveness	Comprehensibility	Structural validity	Hypothesis testing (convergent)	Internal consistency	Inter-rater reliability	Test-retest reliability	
Inventory-Computer Adaptive Test Parents Professionals			Very good +	Very good +							
Pediatric Evaluation of Disability Inventory-Patient Reported Outcome	Kramer and Schwartz 2018	Relevance Very good + Comprehensiveness Doubtful + Comprehensibility Doubtful +									
Pediatric Evaluation of Disability Inventory-Patient Reported Outcome	Kramer and Schwartz 2017		Doubtful +		Doubtful +						
School Function Assessment	Coster, Mancini and Ludlow 1999					Inadequate +					
Performance Scales											
School Function Assessment Performance Scales	Davies et al. 2004								Adequate + -		
School Function Assessment Performance Scales	Hwang et al. 2002						Very good +				
Street Survival Skills Questionnaire	Janniro, Sapp and Kohler 1994						Doubtful + -				
Challenging/Problem Behaviour Assessments											
Aberrant Behavior Checklist	Brown, Aman and Havercamp. 2002					Very good -					
Aberrant Behavior Checklist	Kaat, Lecavalier and Aman 2014					Very good -	Very good +	Very good +			
Aberrant Behavior Checklist	Marshburn and Aman 1992					Adequate +					
Aberrant Behavior Checklist - Irritability Subscale	Siegel et al. 2013						Very good +				
Challenging Behaviour Interview	Oliver et al. 2003						Adequate +				
Developmentally Delayed Children's Behaviour Checklist	Einfeld and Tonge 1991						Adequate +	Doubtful +	Dubtful Teachers + Teacher-TAs -	Doubtful +	
Home Situations Questionnaire-PDD	Chowdhury et al. 2010					Inadequate +	Very good +	Very good +			Very good +

Table 4 (continued)

Assessment	Author	Content validity				Construct validity		Reliability			
		Development	Parents/experts								
		Prom design	Relevance	Comprehensiveness	Comprehensibility	Structural validity	Hypothesis testing (convergent)	Internal consistency	Inter-rater reliability	Test-retest reliability	Responsiveness
Assessments of Behaviour Related to Autism											
Autism Behavior Inventory	Bangerter et al. 2017					Inadequate ?	Very good +	Very good +		Inadequate ?	
Autism Impact Measure	Kanne et al. 2014					Adequate +	Very good +		Doubtful +	Doubtful +	
Autism Treatment Evaluation Checklist	Charman et al. 2004										Inadequate -
Autism Treatment Evaluation Checklist	Magiati et al. 2011						Very good +	Very good +			Inadequate -
Repetitive Behavior Scale-Revised	Lam and Aman 2007					Adequate -		Very good +	Very good +		
Teacher Autism Progress Scale	Dang et al. 2017										Adequate +
Sensory Behaviour Questionnaire	Neil, Green and Pellicano 2017						Adequate +	Very good +			

Measurement properties ratings: + sufficient; - insufficient;? indeterminate; +- inconsistent. Evidence quality ratings: very good, adequate, doubtful or inadequate

R; SBQ; TAPS) are likely to be more effective, valid and reliable for this purpose than those which were developed for multiple purposes.

Another consideration is the population for which the assessment was intended and, further, the population with which the assessment has been evaluated. Eight assessments were developed specifically for use with individuals on the autism spectrum but only three of these were intended for use in schools or evaluated using teacher respondents (PDD-BI; ATEC; TAPS). Of the 14 assessments evaluated by teachers, only Wells et al. (2009) and Perry and Factor (1989) who evaluated the AAMR ABS-II and previous AAMD version specified that all or nearly all participants were on the autism spectrum with coexisting intellectual disabilities. The ATEC, PDD-BI, SFA, TAPS and the WOST used samples who needed some special educational provision but the level of intellectual disabilities amongst participants is likely to have varied (e.g. Charman et al. 2004; Dang et al. 2017; Magiati et al. 2011). The NCBRF contained samples with various levels of intellectual disabilities without specifying autism (Aman et al. 1996). Considering the often-complex educational needs of this specific population, it would be beneficial for further studies on these assessments to be carried out using a sample of children on the autism spectrum with coexisting intellectual disabilities and to consider whether revised versions of these assessments specific to this population would be useful.

Usability of Assessment by Teachers

Only 12 of the 26 assessments were developed specifically for use by teachers or in schools and four further assessments (ABAS-II; VABS-II; BASC-2, SFA) were subject to publishers' qualification codes. As mentioned during the description of the screening process, the intermediary qualification codes may restrict the use of these assessments by a large number of special needs teachers in education systems where no further qualifications are required to teach this population. It may be that these assessments require specialist knowledge in terms of scoring or interpreting the results and potentially any interventions or support resulting from the outcome of the assessments. Therefore, effective use of these assessments would require supervision by members of school leadership with extra qualifications or even external professionals, limiting their use for the purposes of this review. However, this does not mean that teachers cannot inherently develop and implement interventions that are designed to improve functioning in areas relevant to these assessments (e.g. adaptive behaviour).

Although under half of included assessments were developed for use by teachers in schools, it is encouraging that assessments are being developed specifically for this purpose and that use in an education setting is considered during development. Some recently evaluated and available assessments such as the ABI, the RBS-R and the SBQ may also have potential for assessing particular areas of difficulty in schools; however, it is necessary for research to be conducted using teacher respondents to further assess applicability and appropriateness for use in education settings. Only four studies provided enough information to assess content validity and, of the other included assessments which gave brief information about the development process, only the WOST reported input from teachers at the development stage (Wigelsworth et al. 2015). Content validity is a vital consideration and, in this context, requires input from teachers during development in the areas identified in the COSMIN checklist including relevance of the items and comprehensiveness of the assessment as well as comprehensibility of the assessment instructions, items and response options.

The ATEC, WOST, CBI and GO4KIDDS were the only four assessments tested by teachers in the UK. Although results of studies conducted in one country may be applicable to another, it is useful to consider the appropriateness of these assessments in schools of the country in question. In this case, it may be particularly beneficial for the TAPS, which was specifically devised to be used by teachers and showed sensitivity to progress over time, to be evaluated within special schools in England. Similarly, although mainly involving clinicians in its development, the PEDI-CAT received some input from teachers during the development process and, with further evaluation in schools, may be appropriate for use in education settings. Initial evaluation of the PEDI-PRO suggests it may be useful as a pupil report measure but, again, further research on its use in schools is needed.

Measurement Properties

Although a number of assessments showed sufficient evidence for various different measurement properties, few were evaluated with a relevant sample or in an appropriate setting. The ABC and the RBS-R showed promise for use by teachers with autistic children with intellectual disabilities to assess challenging behaviour and repetitive behaviour respectively, as did the PDD-BI. However, these assessments need further evaluation of their responsiveness to change and their use in schools in England. Both studies with a focus on the ATEC reported on responsiveness, which is an important measurement property when determining how well an assessment measures progress. Studies evaluating the responsiveness of the ATEC reported either some change on some subscales during the time period (mean 11 months) or scores remaining relatively stable but with different individual patterns of change (Charman et al. 2004; Magiati et al. 2011). This appears in line with expectations for a heterogeneous condition such as autism. These results suggest it is unlikely the ATEC would show progress and change over shorter periods of time (e.g. termly or half-termly) although the tool may be useful for

teachers to show longer term progress. Both of these evaluations, however, were rated as having inadequate methodological quality. The WOST showed high internal consistency and the CFA indicated a close to ideal model fit (Wigelsworth et al. 2015) but further testing may be needed to determine the responsiveness of the scale. Some items on this scale are more relevant to individuals with mild intellectual disabilities rather than moderate to severe intellectual disabilities so it may be useful for further research to be conducted using this sample. It is also necessary for further studies to be carried out on the content validity of these assessments with the specific population considered in this study, as mentioned above.

Assessments Appropriate for Schools

Twenty-six relevant assessments were identified in this review. When taking account of the factors considered above, there are few, if any, assessments which have been evidenced to be entirely appropriate for teachers to show progress of autistic pupils with intellectual disabilities without need for further evaluation. Many of these assessments were originally developed for other purposes or are limited by qualification codes and, therefore, may need to be adapted in order to be appropriate for use in schools. Many do not have sufficient evidence of a number of robust measurement properties, particularly responsiveness, when using teacher respondents. Most have not been evaluated with individuals on the autism spectrum with intellectual disabilities when used by teachers in education settings in England. Considering the various factors discussed above, the ABC, PDD-BI, ATEC and TAPS may have potential for use in special schools to show progress of pupils on the autism spectrum. However, further evaluation is necessary. The teacher version of the AIM could also be a useful addition to the pool of current available assessments upon completion. In light of the discussion and evaluation of the identified assessments, there is a clear need to develop robust assessments for use by special needs teachers to measure progress and outcomes of autistic children with intellectual disabilities outside of curriculum areas.

Limitations

This systematic review, to our knowledge, is the first to consider the educational appropriateness of assessments which measure progress in adaptive behaviour, challenging or problem behaviour and autism-related behaviour for children on the autism spectrum with intellectual disabilities. It has systematically identified relevant assessments, summarised and reviewed evidence pertaining to measurement properties and examined the assessments in respect of their use by teachers in special schools. It has also devised some adaptations to the COSMIN checklist for these purposes. This review provides a resource for teachers which summarises the potential uses of included assessments with different pupils as well as reporting on their measurement properties.

There are, however, a number of limitations of the current systematic review. Firstly, some notable assessments were not included in this study. This may be for a number of reasons. Relevant assessments may have been used in studies for diagnostic/screening purposes, to discriminate between groups or may have used a sample of individuals without intellectual disabilities and would therefore have been excluded. Some assessments were also excluded on qualification code whilst newer assessments and recent versions may not have been included if there have not yet been studies of their measurement properties published. It is, therefore, important to acknowledge that this review is not an exhaustive list of assessments appropriate for use with children on the autism spectrum with intellectual disabilities but an example of those that can be used by SEN teachers.

Furthermore, evaluations of assessments' measurement properties outside of peer-reviewed literature, for example in books, were not included. Properties containing potentially helpful information such as ceiling and floor effects were not considered here. The COSMIN checklist guided the quality assessment but was adapted to suit the specific aims of this review and it is necessary to interpret the methodological quality and summary of measurement properties with caution if considering the results in a broader context than is specified here.

In addition, there are also a number of school assessments which are notably missing from the literature. Those widely used in special schools in England include the B Squared assessment software, the Early Years Foundation Stage Profile and the TEACCH Transition Assessment Profile. Not only did these assessments not appear in this systematic search or further searches of peer-reviewed journals, grey literature searches specifically for these assessments also yielded no information on evaluation of their properties. B Squared was also contacted for information on their measurement properties but no reply was received. In a similar way, measurement approaches such as Goal Attainment Scaling are less likely to be included when considering evaluations of the measurement properties of these tools. McConachie et al. (2015) mentioned that criterion-referenced assessment and other assessment approaches are often not examined for their measurement properties in research. Persicke et al. (2014) recognised that, due to a lack of expertise around measurement properties, limited information on measurement properties of assessments are available in fields such as education. Teachers may intrinsically 'know' which assessments are helpful for them and their pupils and not rely on further academic evaluations of assessments which they find useful. With school wide assessment policies often chosen and developed by individuals predominantly working outside of the classroom, it is important that

the gap between robust and sound assessments and their effective use by teachers is bridged.

With these limitations in mind, systematic reviews and further research replicating and evaluating the results here are recommended to address the lack of research in this area.

Conclusions and Recommendations

This systematic review has addressed the first of the two research questions by identifying and listing assessments which can be used by special needs teachers to assess the progress and outcomes of pupils on the autism spectrum with coexisting intellectual disabilities. The review summarised the assessment information and identified the assessment methods, previous uses of the assessments and the populations they have assessed. In addressing the second research question and determining which assessments are appropriate for these purposes, factors considered included the availability of the assessment, accessibility and ease of use by teachers, whether the assessments had been evaluated with a relevant population and with teacher respondents, and the outcome of the evaluation in relation to their measurement properties. The findings of this systematic review lead to the recommendation that the ABC, PDD-BI, ATEC and TAPS are currently the most appropriate assessments of outcomes for pupils on the autism spectrum in education settings in the areas of adaptive behaviour, challenging behaviour and autism-related behaviour. These recommendations are made whilst accepting some limitations of these assessments and with the understanding that their appropriateness may vary depending upon the unique purposes of assessment and needs of the pupils. All but the ATEC were evaluated in the USA and therefore this may be a consideration regarding their uses in other education systems. Furthermore, all of these studies require additional evaluation of various measurement properties in relevant contexts and settings.

There are a number of further recommendations as a result of these findings. Firstly, many assessments used in schools have not been evaluated in peer-review literature and it is recommended that widely used assessments in special schools have their measurement properties evaluated. Secondly, as recommended by McConachie et al. (2015), it is critical that stakeholders are involved in the development of new assessments; specifically, that teachers are included in the development process of teacher assessments and that they support decisions on skills and behaviour which are most useful to assess. Thirdly, it is important for responsiveness to be evaluated including measuring small amounts of progress over shorter periods of time (e.g. termly or half-termly) for the purpose of showing progress in schools. Finally, it is recommended that assessments are developed with and for teachers to show progress for children on the autism spectrum with intellectual disabilities outside of curriculum areas. Evaluation with an appropriate sample in a relevant education setting is also recommended in order to address the need for robust assessment tools for these purposes.

Compliance with Ethical Standards

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

Ethical Approval This article does not contain any studies by human participants or animals performed by any of the authors.

References

Aman, M. G., Tassé, M. J., Rojahn, J., & Hammer, D. (1996). The Nisonger CBRF: A child behavior rating form for children with developmental disabilities. *Research in Developmental Disabilities, 17*(1), 41–57.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington: American Psychiatric Publishing.

Aricak, O. T., & Oakland, T. (2010). Multigroup confirmatory factor analysis for the teacher form, ages 5 to 21, of the adaptive behavior assessment system-II. *Journal of Psychoeducational Assessment, 28*(6), 578–584.

Arnold, S., & Reed, P. (2016). Reading assessments for students with ASD: A survey of summative reading assessments used in special education schools in the UK. *British Journal of Special Education, 43*(2), 122–141.

Bangerter, A., Ness, S., Aman, M. G., Esbensen, A. J., Goodwin, M. S., Dawson, G., & Harris, A. (2017). Autism behavior inventory: A novel tool for assessing core and associated symptoms of autism spectrum disorder. *Journal of Child and Adolescent Psychopharmacology, 27*(9), 814–822.

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., & Brayne, C. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *The British Journal of Psychiatry, 194*, 500–509.

Brown, E. C., Aman, M. G., & Havercamp, S. M. (2002). Factor analysis and norms for parent ratings on the Aberrant Behavior Checklist-Community for young people in special education. *Research in Developmental Disabilities, 23*(1), 45–60.

Burgess, A., & Gutstein, S. (2007). Quality of life for people with autism: Raising the standard for evaluating successful outcomes. *Child and Adolescent Mental Health, 12*(2), 80–86.

Charman, T., Howlin, P., Berry, B., & Prince, E. (2004). Measuring developmental progress of children with autism spectrum disorder on school entry using parent report. *Autism, 8*(1), 89–100.

Chatham, C. H., Taylor, K. I., Charman, T., Liogier D'ardhuy, X., Eule, E., Fedele, A., & San Jose Caceres, A. (2018). Adaptive behavior in autism: Minimal clinically important differences on the Vineland-II. *Autism Research, 11*(2), 270–283.

Chowdhury, M., Aman, M. G., Scahill, L., Swiezy, N., Arnold, L. E., Lecavalier, L., & Sukhodolsky, D. (2010). The Home Situations Questionnaire-PDD version: Factor structure and psychometric properties. *Journal of Intellectual Disability Research, 54*(3), 281–291.

Cohen, I. L. (2003). Criterion-related validity of the PDD Behavior Inventory. *Journal of Autism and Developmental Disorders, 33*(1), 47–53.

Cohen, I. L., Schmidt-Lackner, S., Romanczyk, R., & Sudhalter, V. (2003). The PDD Behavior Inventory: A rating scale for assessing response to intervention in children with pervasive developmental disorder. *Journal of Autism and Developmental Disorders, 33*(1), 31–45.

Coster, W. J., Mancini, M. C., & Ludlow, L. H. (1999). Factor structure of the school function assessment. *Educational and Psychological Measurement, 59*(4), 665–677.

Dang, K., Bent, S., Lawton, B., Warren, T., Widjaja, F., McDonald, M. G., & Hendren, R. L. (2017). Integrating autism care through a school-based intervention model: A pilot study. *Journal of Clinical Medicine, 6*(10), 97.

Davies, P. L., Soon, P. L., Young, M., & Clausen-Yamaki, A. (2004). Validity and reliability of the school function assessment in elementary school students with disabilities. *Physical & Occupational Therapy in Pediatrics, 24*(3), 23–43.

Dumas, H., Fragala-Pinkham, M., Haley, S., Coster, W., Kramer, J., Kao, Y. C., & Moed, R. (2010). Item bank development for a revised pediatric evaluation of disability inventory (PEDI). *Physical & Occupational Therapy in Pediatrics, 30*(3), 168–184.

Einfeld, S. L., & Tonge, B. J. (1991). Psychometric and clinical assessment of psychopathology in developmentally disabled children. *Australia and New Zealand Journal of Developmental Disabilities, 17*(2), 147–154.

Ellison, K. S., Bundy, M. B., Wygant, D. B., & Gore, J. S. (2016). Don't forget about the teachers! Parent and teacher BASC-2 reports of children on the DSM-5 Autism Spectrum. *Journal of Mental Health Research in Intellectual Disabilities, 9*(3), 119–132.

Floyd, R. G., Shands, E. I., Alfonso, V. C., Phillips, J. F., Autry, B. K., Mosteller, J. A., & Irby, S. (2015). A systematic review and psychometric evaluation of adaptive behavior scales and recommendations for practice. *Journal of Applied School Psychology, 31*(1), 83–113.

Hanratty, J., Livingstone, N., Robalino, S., Terwee, C. B., Glod, M., Oono, I. P., & McConachie, H. (2015). Systematic review of the measurement properties of tools used to measure behaviour problems in young children with autism. *PLoS One, 10*(12), e0144649.

Harris, S., Handleman, J., Belchic, J., & Glasberg, B. (1995). The Vineland Adaptive Behavior Scales for young children with autism. *Special Services in the Schools, 10*(1), 45–54.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238.

Higgins, J. P. T. & Green, S. (eds). (2011). Cochrane handbook for systematic reviews of interventions version 5.1.0. The Cochrane Collaboration. Available from www.handbook.cochrane.org. Accessed: 27/12/2017

Hwang, J. L., Davies, P. L., Taylor, M. P., & Gavin, W. J. (2002). Validation of school function assessment with elementary school children. *OTJR: Occupation, Participation and Health, 22*(2), 48–58.

Janniro, F., Sapp, G. L., & Kohler, M. P. (1994). Validating the Street Survival Skills Questionnaire. *Psychological Reports, 74*(1), 191–194.

Kaat, A. J., Lecavalier, L., & Aman, M. G. (2014). Validity of the aberrant behavior checklist in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 44*(5), 1103–1116.

Kanne, S. M., Mazurek, M. O., Sikora, D., Bellando, J., Branum-Martin, L., Handen, B., & 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.

Kicklighter, R. H., & Bailey, B. S. (1980). A direct measure of adaptive behavior. *School Psychology Review, 9*(2), 168–173.

Knapp, M., Romeo, R. & Beecham, J. (2007). *The Economic Consequences of Autism in the UK.* London: Foundation for People with Learning Disabilities.

Kramer, J. M., Coster, W., Kao, Y., Snow, A., & Orsmond, G. (2012). A new approach to the measurement of adaptive behavior: Development of the PEDI-CAT for children and youth with autism spectrum disorders. *Physical & Occupational Therapy in Pediatrics, 32*(1), 34–47.

Kramer, J. M., & Schwartz, A. E. (2017). Refining the Pediatric Evaluation of Disability Inventory–Patient-Reported Outcome (PEDI-PRO) item candidates: Interpretation of a self-reported outcome measure of functional performance by young people with neurodevelopmental disabilities. *Developmental Medicine and Child Neurology, 59*(10), 1083–1088.

Kramer, J. M., & Schwartz, A. E. (2018). Development of the Pediatric Disability Inventory-Patient Reported Outcome (PEDI-PRO) measurement conceptual framework and item candidates. *Scandinavian Journal of Occupational Therapy, 25*(5), 335–346 1-12.

Lam, K. S., & 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.

Lane, B. R., Paynter, J., & Sharman, R. (2013). Parent and teacher ratings of adaptive and challenging behaviours in young children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 7*(10), 1196–1203.

Magiati, I., Moss, J., Yates, R., Charman, T., & Howlin, P. (2011). Is the Autism Treatment Evaluation Checklist a useful tool for monitoring progress in children with autism spectrum disorders? *Journal of Intellectual Disability Research, 55*(3), 302–312.

Marshburn, E. C., & Aman, M. G. (1992). Factor validity and norms for the Aberrant Behavior Checklist in a community sample of children with mental retardation. *Journal of Autism and Developmental Disorders, 22*(3), 357–373.

Matson, J., & Shoemaker, M. (2009). Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities, 30*, 107–1114.

Mayfield, K. L., Forman, S. G., & Nagle, R. J. (1984). Reliability of the AAMD adaptive behavior scale-public school version. *Journal of School Psychology, 22*(1), 53–61.

McConachie, H., Livingstone, N., Morris, C., Beresford, B., Le Couteur, A., Gringras, P., Garland, D., Jones, G., Macdonald, G., Williams, K., & Parr, J. R. (2018). Parents suggest which indicators of progress and outcomes should be measured in young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 48*(4), 1041–1051.

McConachie, H., Parr, J. R., Glod, M., Hanratty, J., Livingstone, N., Oono, I. P., Robalino, S., Baird, G., Beresford, B., Charman, T., Garland, D., Green, J., Gringras, P., Jones, G., Law, J., Le Couteur, A. S., Macdonald, G., McColl, E. M., Morris, C., Rodgers, J., Simonoff, E., Terwee, C. B., & Williams, K. (2015). Systematic review of tools to measure outcomes for young children with autism spectrum disorder. *Health Technology Assessment, 19*(41).

Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171–1179.

Munir, S. Z., Zaman, S., & McConachie, H. (1999). Development of an independent behaviour assessment scale for Bangladesh. *Journal of Applied Research in Intellectual Disabilities, 12*(3), 241–252.

Neil, L., Green, D., & Pellicano, E. (2017). The psychometric properties of a new measure of sensory behaviors in autistic children. *Journal of Autism and Developmental Disorders, 47*(4), 1261–1268.

Office for Standards in Education, Children's Services and Skills (Ofsted). (2019). *School Inspection Handbook.* Available at www.gov.uk/government/organisations/ofsted . Accessed: 07/07/2019

Oliver, C., McClintock, K., Hall, S., Smith, M., Dagnan, D., & Stenfert-Kroese, B. (2003). Assessing the severity of challenging behaviour: Psychometric properties of the challenging behaviour interview. *Journal of Applied Research in Intellectual Disabilities, 16*(1), 53–61.

Pan, X., Totsika, V., Nicholls, G., & Paris, A. (2019). Validating GO 4 KIDDS as a brief measure of adaptive skills in special education settings for children with severe intellectual disability. *Journal of Applied Research in Intellectual Disabilities, 32*(2), 280–287.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65–77.

Pellicano, E., Dinsmore, A., & Charman, T. (2014). What should autism research focus upon? Community views and priorities from the United Kingdom. *Autism, 18*(7), 756–770.

Perry, A., & Factor, D. C. (1989). Psychometric validity and clinical usefulness of the Vineland Adaptive Behavior Scales and the AAMD Adaptive Behavior Scale for an autistic sample. *Journal of Autism and Developmental Disorders, 19*(1), 41–55.

Perry, A., Taheri, A., Ting, V., & Weiss, J. (2015). The GO 4 KIDDS Brief Adaptive Scale. *Journal of Applied Research in Intellectual Disabilities, 28*(6), 594–597.

Persicke, A., Bishop, M. R., Coffman, C. M., Najdowski, A. C., Tarbox, J., Chi, K., & Ranick, J. (2014). Evaluation of the concurrent validity of a skills assessment for autism treatment. *Research in Autism Spectrum Disorders, 8*(3), 281–285.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157.

Provenzani, U., Fusar-Poli, L., Brondino, N., Damiani, S., Vercesi, M., Meyer, N., & Politi, P. (2019). What are we targeting when we treat autism spectrum disorder? A systematic review of 406 clinical trials. *Autism, 24*(2), 274-284.

Siegel, M., Milligan, B., Stein, H., Teer, O., & Smith, K. A. (2013). Telephone administration of the aberrant behavior checklist: A pilot study of feasibility in children with intellectual disability and autism. *Journal of Intellectual Disabilities, 17*(3), 265–271.

Silverman, W. P., Silver, E. J., Lubin, R. A., & Sersen, E. A. (1983). Structure of the Minnesota Developmental Programming System Behavioral Scales, Alternate Form C. *American Journal of Mental Deficiency, 88*(2), 170–176.

Sparrow, S., & Cicchetti, D. (1978). Behavior rating inventory for moderately, severely, and profoundly retarded persons. *American Journal of Mental Deficiency, 82*(4), 365–374.

Spreat, S. (1982). An empirical analysis of item weighting on the adaptive behavior scale. *American Journal of Mental Deficiency, 87*(2), 159–163.

Steer, A. S. (2009). *Learning behaviour: Lessons learned. A review of behaviour standards and practices in our schools*. Nottingham: DCSF Publications.

Terwee, C. B., Prinsen, C. A., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*(5), 1159–1170.

Voelker, S. L., Shore, D. L., Lee, C. H., & Szuszkiewicz, T. A. (2000). Congruence in parent and teacher ratings of adaptive behavior of low-functioning children. *Journal of Developmental and Physical Disabilities, 12*(4), 367–376.

Walton, K. M., & Ingersoll, B. R. (2013). Improving social skills in adolescents and adults with autism and severe to profound intellectual disability: A review of the literature. *Journal of Autism and Developmental Disorders, 43*(3), 594–615.

Wells, K., Condillac, R., Perry, A., & Factor, D. C. (2009). A comparison of three adaptive behaviour measures in relation to cognitive level and severity of autism. *Journal on Developmental Disabilities., 15*(3), 55.

Wigelsworth, M., Oldfield, J., & Humphrey, N. (2015). Validation of the Wider Outcomes Survey for Teachers (WOST): A measure for assessing the behaviour, relationships and exposure to bullying of children and young people with special educational needs and disabilities (SEND). *Journal of Research in Special Educational Needs, 15*(1), 3–11.