BMJ Open

Psychometric properties of gross motor assessment tools for children: a systematic review

Alison Griffiths,[1,2,3] Rachel Toovey,[3,4] Prue E Morgan,[1] Alicia J Spittle[3,4]

To cite: Griffiths A, Toovey R, Morgan PE, *et al*. Psychometric properties of gross motor assessment tools for children: a systematic review. *BMJ Open* 2018;**8**:e021734. doi:10.1136/bmjopen-2018-021734

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-021734).

Received 15 January 2018
Revised 23 August 2018
Accepted 31 August 2018

[1]Department of Physiotherapy, School of Primary and Allied Health Care, Monash University, Frankston, Victoria, Australia
[2]Department of Physiotherapy, The Royal Children's Hospital, Parkville, Victoria, Australia
[3]Murdoch Children's Research Institute, Parkville, Victoria, Australia
[4]Department of Physiotherapy, The University of Melbourne, Parkville, Victoria, Australia

Correspondence to
Dr Alicia J Spittle;
aspittle@unimelb.edu.au

ABSTRACT

Objective Gross motor assessment tools have a critical role in identifying, diagnosing and evaluating motor difficulties in childhood. The objective of this review was to systematically evaluate the psychometric properties and clinical utility of gross motor assessment tools for children aged 2–12 years.

Method A systematic search of MEDLINE, Embase, CINAHL and AMED was performed between May and July 2017. Methodological quality was assessed with the COnsensus-based Standards for the selection of health status Measurement INstruments checklist and an outcome measures rating form was used to evaluate reliability, validity and clinical utility of assessment tools.

Results Seven assessment tools from 37 studies/manuals met the inclusion criteria: Bayley Scale of Infant and Toddler Development-III (Bayley-III), Bruininks-Oseretsky Test of Motor Proficiency-2 (BOT-2), Movement Assessment Battery for Children-2 (MABC-2), McCarron Assessment of Neuromuscular Development (MAND), Neurological Sensory Motor Developmental Assessment (NSMDA), Peabody Developmental Motor Scales-2 (PDMS-2) and Test of Gross Motor Development-2 (TGMD-2). Methodological quality varied from poor to excellent. Validity and internal consistency varied from fair to excellent (α=0.5–0.99). The Bayley-III, NSMDA and MABC-2 have evidence of predictive validity. Test–retest reliability is excellent in the BOT-2 (intraclass correlation coefficient (ICC)=0.80–0.99), PDMS-2 (ICC=0.97), MABC-2 (ICC=0.83–0.96) and TGMD-2 (ICC=0.81–0.92). TGMD-2 has the highest inter-rater (ICC=0.88–0.93) and intrarater reliability (ICC=0.92–0.99).

Conclusions The majority of gross motor assessments for children have good-excellent validity. Test–retest reliability is highest in the BOT-2, MABC-2, PDMS-2 and TGMD-2. The Bayley-III has the best predictive validity at 2 years of age for later motor outcome. None of the assessment tools demonstrate good evaluative validity. Further research on evaluative gross motor assessment tools are urgently needed.

Strengths and limitations of this study

- This systematic review comprehensively assesses methodological quality of included studies using the COnsensus-based Standards for the selection of health status Measurement INstruments checklist.
- Results of this systematic review can provide guidance to clinicians when choosing gross motor assessment tools based on test psychometric properties and clinical utility.
- Areas for future research are identified including improving the evidence of inter-rater and intrarater reliability and responsiveness to change as well as the ascertainment of predictive validity over a longer period of time.
- Only articles or test manuals written in English were included.
- Only one reviewer screened titles and abstracts for inclusion.

INTRODUCTION

Motor function promotes cognitive and perceptual development in children and contributes to their ability to participate in their home, school and community environments.[1] Motor impairment can negatively affect activity and participation levels of children,[2] which may lead to lower levels of physical activity, fitness and health into adulthood.[3] While severe motor deficits are usually diagnosed before 2 years of age, mild motor deficits may not become evident until children are in preschool and primary school environments where they are exposed to increasingly complex tasks and compared with their peers.[3] Identification of motor difficulties is an important step towards support and intervention for the child and their family.

Healthcare professionals and researchers require standardised assessment tools to identify, classify and diagnose motor problems in children.[4] Furthermore, assessment tools are essential to monitor the effects of interventions.[4] There is no gold standard of motor assessment for children and the available tests vary in their ease of use and interpretability in clinical and research settings, and whether they are norm or criterion referenced.[5] Criterion referenced tests are designed to be scored as items or criteria are demonstrated; meaning that the score

is a reflection of a child's competence on the test items. Most available assessments however, are norm referenced, meaning that a child's results are reported in relation to a specific population.[4] The characteristics of the normed population should be taken into consideration when interpreting test results as environmental and cultural differences have been found to affect motor development.[6]

Healthcare professionals should be aware of the validity and reliability of assessment tools to assist in their instrument selection and interpretation of results. Validity refers to 'the degree to which (an instrument) is an adequate reflection of the construct to be measured'.[7] If an instrument does not have adequate construct or content validity then it may not be assessing the skills that it purports to. Reliability refers to 'the degree to which the measurement is free from measurement error',[7] which is significant when interpreting results. If a child is assessed as being significantly delayed in their gross motor skills, the reliability of that tool indicates the likelihood that a result is due to error.

A systematic review in 2010 by Slater *et al*[8] evaluated performance-based gross motor tests for children with developmental coordination disorder; however, it did not include the second and most recent version of the Movement Assessment Battery for Children-2 (MABC-2), which is widely used. Brown and Lalor[9] suggested that as a result of the changes to the original MABC in age range, age bands, materials and tasks, the MABC-2 requires independent reliability and validity assessment. Over the past 8 years, there has also been a significant increase in the number of papers assessing the psychometric properties of motor assessment tools in children. A systematic review of these and previous papers is warranted, in order to add to our understanding of the psychometrics of standardised gross motor assessment tools.

The primary aim of this systematic review is to identify and evaluate the clinical utility and psychometric properties of gross motor assessment tools appropriate for use in preschool and school age children from 2 to 12 years by assessing the methodological quality of the included studies. The secondary aim of this review is to identify any areas for further research.

METHOD

A comprehensive search strategy was completed in databases OVID Medline (1996 to May 2017), CINAHL plus (1937 to July 2017), Embase (1974–May 2017) and AMED (1985–July 2017) (see online supplementary tables 1-4). The search strategy used MeSH terms and text words for ('child' or 'paediatric') and ('motor skills' or 'motor activity' or 'gross motor' or 'psychomotor' or 'developmental coordination disorder') and ('questionnaires' or 'outcome assessment' or 'instrument' or 'task performance') and ('reliability' or 'validity' or 'psychometrics'). Reference lists of included articles were also screened to identify any additional papers. If full texts were unavailable or further information was required regarding availability of manuals, the authors were contacted.

Assessment tools were included if they were (1) discriminative, predictive or evaluative of gross motor skills, (2) assessed ≥two gross motor (eg, balance, jumping, etc) items, (3) able to extract a meaningful gross motor subscore, (4) applicable to children aged 2–12 years, (5) criterion or norm referenced test with a standardised assessment procedure and (6) instructional manuals are published or commercially available.

Articles describing use of the assessment tool were included if; ≥90% of the study population were within 2-12 years of age, it was available in English and if validity and/or reliability of the assessment tool was reported.

Assessment tools were excluded if they met any of the following criteria: (1) questionnaires or screening tools, (2) only applicable to children with a specific diagnosis (eg, cerebral palsy, Down's syndrome), (3) test manuals not available in English and (4) the version of the test has been superseded.

Titles and abstracts were screened by the first author with any studies that clearly did not meet inclusion criteria excluded. The remaining papers were obtained in full text and reviewed by two authors (AG, RT or PM) with selection based on inclusion and exclusion criteria. Papers and assessment tools were included after discussing with both raters, with conflicting decisions discussed until a consensus was reached.

Methodological assessment of the papers was completed using the four-point scale of the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist.[10] The COSMIN incorporates three quality domains: validity, reliability and responsiveness consisting of seven measurement properties: content, construct and criterion validity, internal consistency, reliability, measurement error and responsiveness[7] (see online supplementary table 5). Cross-cultural validity, structural validity and hypothesis testing are all considered to be a component of construct validity.[7] While predictive validity is considered to be a component of content validity, it is reported separately in this paper for interpretability of results.[7]

The overall score for each measurement property on the COSMIN checklist is determined by a 'worse score counts' approach.[10] Each property is rated as excellent, good, fair or poor methodological quality based on descriptive criteria. Data extraction and assessment of methodological quality was performed independently by two assessors (AG and RT). In the case of any uncertainty, a third reviewer (AS) performed a COSMIN assessment and disagreement was resolved through discussion.

A data extraction form for each assessment tool was adapted from the CanChild Outcome Measures Rating Form to collate information on clinical utility, validity, reliability and responsiveness.[11] Items chosen to represent the clinical utility of the assessment tools were the cost

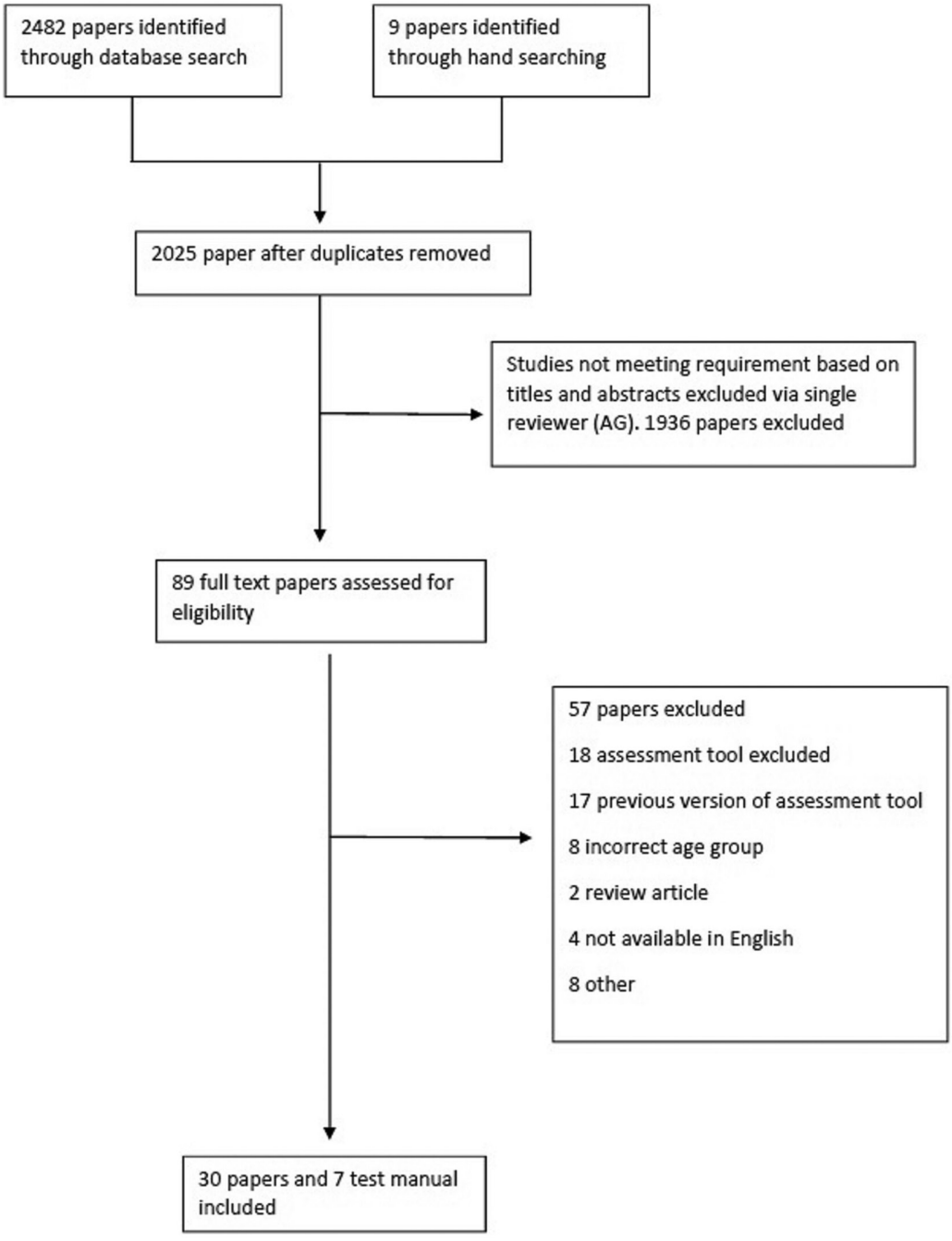

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram detailing study selection.

of manuals, kits, training requirements, time to administer the assessment and the ease of scoring. All reported values for reliability were collected; however, only those papers reporting intraclass correlation coefficient (ICC) were directly compared.

Patient and public involvement

As this was a systematic review of existing papers, there was no patient or public involvement.

RESULTS

Figure 1 provides details of study selection. Seven assessment tools were identified for inclusion: Bayley Scale of Infant and Toddler Development III (Bayley-III), Bruininks-Oseretsky Test of Motor Proficiency 2 (BOT-2), MABC-2, McCarron Assessment of Neuromuscular Development (MAND), Neurological Sensory Motor Developmental Assessment (NSMDA), Peabody Developmental Motor Scales 2 (PDMS-2) and Test of Gross Motor Development 2 (TGMD-2). The corresponding manuals were then added to the final yield resulting in 30 papers and 7 manuals. Twenty assessment tools were excluded (see online supplementary table 6).

The majority of assessment tools identified in this review are discriminative and most lend themselves towards use in a research setting. All norm referenced tools are

Table 1 Gross motor assessment tool characteristics

Assessment tool	Domains tested	Gross motor components tested	Age range	Diagnostic criteria	Primary purpose	Secondary purpose	Type of test	Normative sample (year)
Bayley-III[31]	Gross motor, fine motor, cognitive, communication, social/emotional, adaptive	Static postures, dynamic movement, balance	1 month to 3 years	Developmental delay: <25th centile or below 2SD*	Discriminative	Predictive, evaluative, research tool	Norm	1700 children from the USA (2000)
BOT-2[12]	Gross motor, fine motor	Coordination, balance, running speed and agility, strength	4–21 years	*	Discriminative Evaluative	Research tool	Norm	1520 children from the USA (2005)
MABC-2[29]	Gross motor, fine motor, balance	Aiming and catching, static and dynamic balance	3–16 years	Traffic light system: green=normal, amber='at risk' and red=definite motor impairment (<15%)*	Discriminative Evaluative	Intervention planning, research tool	Norm	1172 children from the UK (2006)
MAND[32]	Gross and fine motor	Coordination, jumping, static and dynamic balance	3–25 years	NDI 70–85=mild 55–69=moderate <55=severe disability*	Evaluative	Research tool	Norm	2000 3–35 years from the USA (1970s)
NSMDA[33]	Gross motor, fine motor, neurological, postural development, infant patterns of movement, sensory motor†	Sitting, kneeling, walking, balance, running, hopping, jumping, catching, motor planning	1 month to 6 years	Total score 6–8 normal, 9–11 minimal, 12–14 mild, 15–19 moderate, 20–25 severe, >25 profound disability*	Evaluative Discriminative	Predictive, Research tool	Criterion	NA
PDMS-2[34]	Gross motor, fine motor	Stationary (standing balance, sit-ups, push-ups), locomotion (walking, running, jumping, hopping, etc), object manipulation (kick, throw, hit, catch)	Birth to 5 years	*	Discriminative Evaluative	Predictive, research tool	Norm	2003 USA and Canada (1997–1998)
TGMD-2[15]	Gross motor	Locomotion (run, gallop, hop, leap, jump, slide) and object control (batting, dribbling, catch, kick, throw, roll)	3–10 years	*	Discriminative Evaluative	Outcome measure, research tool, intervention planning	Norm	1208 USA children (1997–1998)

*Advisable to use clinical reasoning.
†Requires some manual handling.[31]
Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; MABC-2, Movement Assessment Battery for Children second edition[29]; MAND, McCarron Assessment of Neuromuscular Development[32]; NA, not available; NDI, Neurodevelopmental Index; NSMDA, Neurological Sensory Motor Developmental Assessment[33]; PDMS-2, Peabody Developmental Motor Scales second edition[34]; TGMD-II, Test of Gross Motor Development second edition.[15]

from western countries and each identified test covers a different age range as shown in table 1.

The TGMD-2 is the only tool that assesses gross motor skills in isolation and that focusses on quality of performance. The other gross motor assessments were either in conjunction with assessment of fine motor and/or balance (MAND, MABC-2, BOT-2 and PDMS-2) or as a component of a developmental assessment (NSMDA, Bayley-III).

Despite the variability in test structures, there is some consistency of items included within the gross motor skill

subsets between tests. Most include a locomotion task such as walking, running or stair climbing; an object control or manipulation task such as throwing or catching a ball and a static or dynamic balance task such as standing on one leg or hopping. The PDMS-2, BOT-2 and the MAND also include strength assessments (the PDMS-2 only in some age groups).

The number of gross motor items for assessment vary both within and between the tools (table 1). For example, the number of items tested in the Bayley-III and the PDMS-2 depends on the age and ability of the child. Several assessments report criteria for describing gross motor delay, although all test manuals warn against diagnosing delay based on a single assessment.

The PDMS-2 is notable for the inclusion of credit towards incomplete skills in the scoring system. Most other tests award a point or credit towards a skill only if it is demonstrated to the full satisfaction of the stated criteria (score of 0 or 1). The PDMS-2 however is scored 0–2 allowing for 1 mark to be allocated as a child progresses towards a skill without mastering it. The TGMD-2 is also notable for its marking system, in which points are awarded for the quality of the action performed, instead of satisfactory completion of the task only. These actions include preparatory movements prior to running and jumping, or arm position during movements. The NSMDA marking criteria is somewhat more complicated with a system of scores 1–4 with a symbol of '+' denoting hyperactive response and '–' a hyporeactive response. The PDMS-2, MABC-2, BOT-2, MAND, TGMD-2 and Bayley-III all require raw scores to be converted to a standard (or scaled) score based on tables supplied in the manuals. For the BOT-2, this is a multiple step process which can then be converted to both sex-specific or combined standard scores and percentile ranks. A summary of assessment tool characteristics can be found in table 1.

Clinical utility

The clinical utility of the assessment tools is summarised in table 2, while scoring and administration is detailed in online supplementary table 7. The shortest administration time is 15–20 min for the TGMD-2 and the MAND, while most manuals report 20–60 min is required to complete an assessment. These times are not inclusive of equipment set up, pack up and scoring, which varies depending on the amount of equipment and complexity of the scoring process. All assessments require the user to be familiar with the test before administration and to possess a high level of understanding of child movement and development. The MABC-2 and PDMS-2 are the only assessments that come with supporting material to guide intervention postassessment (when the complete kit is purchased).

Methodological quality

All articles were assessed using the COSMIN checklist to determine methodological quality. Several studies were marked down for failing to report missing data, small sample sizes and for using inappropriate statistical methods. A summary of the articles and corresponding COSMIN methodology rating is provided in table 3.

Validity

The content and construct validity of the included assessment tools are summarised in table 4. Most assessments were developed by or with input from experts in the field, with most also performing literature reviews. Bruininks and Bruininks[12] performed comprehensive surveys, pilot, tryout and standardisation studies before finalising the BOT-2, providing the most comprehensively reported content validity.

Construct validity was confirmed with factor analysis (either exploratory or confirmatory) in most assessment tools. The TGMD-2 has the most evidence for construct validity with several papers performing confirmatory and exploratory factor analysis.[13–18] The MABC-2, BOT-2, Bayley-III, MAND and PDMS-2 had factor analysis performed only in one paper. The MABC-2 was shown to require changes to remain valid in the Chinese-speaking and Dutch-speaking populations.[19 20] The BOT-2, MABC-2 and TGMD-2 all provide evidence of the ability to discriminate between particular age or diagnosis groups, which can be considered to support their content validity. The NSMDA has minimal assessment of construct validity in children over 2 years. The Bayley-III, NSMDA and MABC-2 are the only assessments that provide evidence of predictive validity (table 5). Concurrent validity between the MABC-2, PDMS-2 and BOT-2 is moderate to high, while the TGMD-2 is only weakly correlated with the MABC-2[5] (table 5). The PDMS-2, TMGD-2 and NSMDA report correlations with other criteria such as paediatrician diagnosis, physical fitness or psychomotor/intelligence tests.

Reliability

Internal consistency of assessments are summarised in table 6. The high internal consistency of the BOT-2 is well supported, including for children with an intellectual disability.[21 22] The MABC-2 appears to have lower internal consistency than the BOT-2, which may relate to the limited number of test items (eight) on the MABC-2. The highest values for internal consistency for the MABC-2 were obtained in specific populations (intellectual disability and developmental coordination disorder) with poor to fair methodology only. Conversely, the highest quality articles reported the lowest values, although it should be noted that these assessed age band 1 (3–6 years) only. Internal consistency is reported to be high for the PDMS-2, while the Bayley-III is shown to have excellent internal consistency in children aged 24–42 months. The TGMD-2 is reported by two good quality (and four poor to fair quality) articles to have excellent internal consistency, including for children with vision impairment and intellectual disability. The MAND is the only assessment tool included in this review without published data of internal consistency or reliability in this age group.

Table 2 Clinical utility of gross motor assessment tools

Assessment tool	Time to administer (min)	Test procedure	Target examiner population	Training	Equipment/manual
Bayley-III[31]	30–90	Therapist administers in standardised order	Paediatric health professionals early childhood specialists	Formal training not required. DVD, webinars and workshops available	Comprehensive manual/kit: £1089 Test kit provides most equipment
BOT-2[12]	40–60	Therapist administered in standardised order	Paediatric health professionals early childhood specialists	Formal training not required	Comprehensive manual/kit: £961 Test kit provides most equipment
MABC-2[29]	20–40	Therapist administers items in standardised order. Some flexibility allowed	Research psychologists, OT, PT, paediatricians	Formal training not required.	Comprehensive manual/kit: £1191 Test kit provides most equipment
MAND[32]	15–20	Therapist administers items in standardised order	Professionals, eg, education, neurology, OT, PT, psychology, etc	Formal training not required	Manual and test kit: £1366 (includes equipment)
NSMDA[33]	20–45	Observation followed by therapist administration of test items	PT, OT	Formal training not required (but is available)	Comprehensive manual: £35 Equipment not included
PDMS-2[34]	45–60 (20–30 for GM only)	Standardised procedure	Paediatric health professionals, PE teachers, early intervention specialists	Formal training not required	Comprehensive manual/kit: £553 Includes some but not all equipment required
TGMD-2[15]	15–20	Standardised procedure	Teachers, health professionals (OT, PT, doctors)	Formal training not required	Kit includes manual and record form: £128 Equipment not included

Bayley-III, Bayley Scale of Infant and Toddler Development third edition[31]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; GM, gross motor; MABC-2, Movement Assessment Battery for Children second edition[29]; MAND, McCarron Assessment of Neuromuscular Development[32]; NSMDA, Neurological Sensory Motor Developmental Assessment[33]; OT, occupational therapy; PDMS-2, Peabody Developmental Motor Scales second edition[34]; PE, physical education; PT, physiotherapy; TGMD-II, Test of Gross Motor Development second edition.[15]

The reliability findings are summarised in table 6 and in figures 2 and 3. Test–retest reliability was excellent in the Bayley-III (table 6), BOT-2 and PDMS-2; and was good to excellent in the MABC-2 and TGMD-2 (figure 2). Intrarater reliability was rarely investigated or reported for most tools, with the TGMD-2 demonstrating better results than the MABC-2 (figure 3). Only the TGMD-2 and MABC-2 report inter-rater reliability values using an ICC (figure 3).[23 24] Inter-rater reliability is also supported in the BOT-2 with Pearson's correlation coefficient and Kappa, respectively. The studies referred to in the test manuals for the TGMD-2, Bayley-III, BOT-2 and MABC-2 all report reliability findings using Pearson's correlation, which is less ideal than an ICC or weighted kappa for statistical analysis.[25 26] Only studies reporting ICCs are visually represented in figure 2 (test–retest) and figure 3 (inter-rater and intra-rater). The TGMD-2 test–retest reliability results from Houwen *et al*[16] were believed to contain an error as the reported ICC was outside of the reported CIs (ICC 0.92, 95% CI 0.82 to 0.91). This data set was therefore excluded from figure 2.

Responsiveness was reported for the Bayley-III, BOT-2, MABC-2 and PDMS-2 with minimal detectable change (MDC) or a SE of measurement (SEM).[21] Sensitivity and specificity for detecting change was shown to be satisfactory in the MABC-2, PDMS-2 and MABC-2[21] (table 6). There have been no studies to date on the responsiveness of the TGMD-2, NSMDA or MAND.

DISCUSSION

This review identified seven gross motor assessment tools appropriate for use in clinical or research settings, each with their own strengths and limitations. Interestingly, only one of the seven assessments (TGMD-2) measured gross motor skills in isolation. This is likely a reflection on current

Table 3 Methodological quality of included articles

Test	Study	Country	Population (age, diagnosis)	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsiveness
Bayley-III	Bayley[31]	USA	1–42 months	Fair	Fair	Good	Excellent	Good	Good	–	Good	–
	Spittle *et al*[4]	Australia	2, 4 years, Ex prem	–	–	–	–	–	–	–	Good	–
	Visser *et al*[35]	The Netherlands	2.2–10.8 years, GDD, LI	–	–	–	Excellent	Poor	–	–	–	–
BOT-2	Wuang and Su[36]	Taiwan	4–12 years ID	Excellent	Excellent	Excellent	–	–	–	–	–	Fair
	Wuang *et al*[21]	Taiwan	3–6 years ID	Fair	Good	Good	–	–	–	–	Good	Fair
	Bruininks and Bruininks[12]	USA	4–21 years	Good	Fair (inter-rater) Fair (test–retest)	Good	Excellent	Good	–	–	Good	–
MABC-2 (AB 1)	Ellinoudis *et al*[37]	Greece	3–5.5 years	Excellent	Good	–	–	–	–	–	–	–
	Hua *et al*[19]	China	3–6 years	Excellent	Good	–	Excellent	Excellent	–	Poor	Excellent	–
	Logan *et al*[5]	USA	3–6 years	–	–	–	–	–	Fair	–	Fair	–
	Smits-Engelsman *et al*[28]	Belgium	3–4 years	Poor	Poor	Poor	–	–	–	–	–	–
MABC-2 (AB 2)	Holm *et al*[23]	Norway	7–9 years	–	Fair (inter-rater) Poor (intrarater)	Poor	–	–	–	–	–	–
	Kita *et al*[38]	Japan	7–10 years	Excellent	–	–	–	–	–	Poor	–	–
MABC-2	Griffiths *et al*[39]	Australia	4–8 years	–	–	–	–	–	–	–	Good	–
	Henderson *et al*[29]	UK	3–16 years	–	Fair	Good	Excellent	–	–	–	–	–
	Niemeijer *et al*[20]	The Netherlands+ Belgium	–	–	–	–	–	–	–	Poor	–	–
	Schulz *et al*[40]	UK	3–16 years	–	–	–	Excellent	Good	–	–	–	–
	Valentini *et al*[41]	Brazil	3–13 years	Fair	Fair	–	Fair	Poor	–	Poor	Poor	–
	Wuang *et al*[21]	Taiwan	3–6 years, ID	Fair	Good	Good	–	–	–	–	Good	Fair
	Wuang *et al*[42]	Taiwan	6–12 years DCD	Poor	Fair	Good	–	–	–	–	–	Fair
MAND	Hands *et al*[43]	Australia	10–17 years	–	–	–	–	Excellent	–	–	–	–
	McCarron[32]	USA	7 years	–	–	–	Fair	Poor	–	–	Poor	–
NSMDA	Danks *et al*[44]	Australia	2+4 years ELBW	–	–	–	–	–	–	–	Fair	–
	MacDonald and Burns[45]	Australia	2+4 years CP	–	–	–	–	Fair	–	–	Poor	–
	Burns *et al*[46]	Australia	1–24 months VLBW	Poor	–	–	Poor	–	–	–	–	–
	Burns *et al*[47]	Australia	1 month VLBW	–	–	–	–	Poor	–	–	Fair	–

Continued

Table 3 Continued

Test	Study	Country	Population (age, diagnosis)	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsiveness
PDMS-2	Hua *et al*[19]	China	3–6 years.	Excellent	Good	–	Excellent	Excellent	–	Poor	Excellent	–
	Wuang *et al*[21]	Taiwan	3–6 years ID	Fair	Good	Good	–	–	–	–	Good	Fair
	Folio and Fewell[34]	USA	0–71 months	Good	–	Poor	Excellent	Good	Good	–	Poor	–
TGMD-2	Barnett *et al*[24]	Australia	4–8 years	–	Fair	–	–	–	–	–	–	–
	Farrokhi *et al*[48]	Iran	3–11 years	Fair	Fair	–	Fair	Fair	–	–	–	–
	Houwen *et al*[16]	The Netherlands	6–12 years VI	Fair	Fair	–	–	Fair	–	–	–	–
	Kim *et al*[49]	Korea	8–12 years ID	–	Poor	–	–	–	–	–	–	–
	Kim *et al*[50]	Korea	5–6 years	Poor	Fair	–	–	Poor	–	–	Poor	–
	Logan *et al*[5]	USA	3–6 years	–	–	–	–	–	Fair	–	Fair	–
	Rudd *et al*[13]	Australia	6–12 years	–	–	–	–	Good	–	–	–	–
	Simons *et al*[18]	Belgium	7–10 years ID	Good	Good (inter-rater) Poor (test–retest)	–	Excellent	Good	Good	–	–	–
	Valentini[14]	Brazil	3–10 years	Poor	Fair (test–retest) Good (intrarater, inter-rater)	–	Excellent	Good	–	Fair	Good	–
	Wong and Yin Cheung[17]	China	3–10 years	–	–	–	–	Fair	–	–	–	–
	Ulrich[15]	USA	3–10 years	Good	Fair (test–retest) Poor (inter-rater)	Fair	Poor	Good	–	–	Fair	–

Bayley-III, Bayley Scale of Infant and Toddler Development third edition; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; CP, cerebral palsy; DCD, developmental coordination disorder; ELBW, extremely low birth weight; GDD, global developmental delay; ID, intellectual disability; LI, language impairment; MABC-2, Movement Assessment Battery for Children second edition[29]; MAND, McCarron Assessment of Neuromuscular Development[32]; NSMDA, Neurological Sensory Motor Developmental Assessment[33]; PDMS-2, Peabody Developmental Motor Scales second edition[15 34]; prem, premature; TGMD-II, Test of Gross Motor Development 2nd edition; VI, vision impairment; VLBW, very low birth weight.[31]

Table 4 Content and construct validity of assessment tools

Test	Content	Construct
Bayley-III	Expert opinion for standard and low verbal version.[31 35] Literature reviews. Gross motor score correlated with Motor component 0.70.[31]	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language impairment.[31] H_i (gross motor subset)=0.52–0.97 for children with language impairment and 0.82–0.99 in control group.[35]
BOT-2	Focus groups, product survey, pilot, national tryout and standardisation studies, professional reviews.[12]	Factor analysis, scores increase with age, discriminates between normal and children with DCD (n=50), high-functioning ASD (n=45) and mild-to-moderate ID (n=66).[12]
MABC-2	Expert panel, stakeholder feedback, literature review.[23] Expert panel—clarity (validity content index 71.8–93.9, κ=0.76–0.88) and pertinence (98.5–99.3 and κ=0.83–0.92), p<0.001.[41]	Factor analysis, correlation coefficients.[37] Subtest correlations 0.65–0.76, p<0.001. Discriminates between ASD and control group.[23] Structural equation modelling (for each age group).[40] Expert panel—adequate face validity.[41] Significant difference between TD, DCD and at risk DCD scores (η^2=0.63), p<0.0001.[41] The UK norms not appropriate to use with Dutch/Flemish children as underestimate/overestimate risk of motor impairment.[20] In Chinese population: CFA initially rejected. Acceptable fit achieved after 2 items removed.[19] Age band 2 shows good validity in Japanese population.[38]
MAND	Based on neuropsychological theory. Several rounds of revision/trials of tasks during development.[32]	Factor analysis.[32 43] Scores increase with age, and discriminate between typically developing children and those with head trauma or neurological dysfunction as well as gender.[32 43]
NSMDA	Literature review. Developed by an experienced paediatric physiotherapist.[46]	Factor analysis (up to 2 years of age).[46 47] Stability of test results over time (up to 2 years).[46 47]
PDMS-2	Literature review. Created by experts in the field. Revised with feedback from therapists guided revision. Hierarchical sequence of items.[34]	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with age (r=0.80–0.93).[34]
TGMD-2	Expert panel (3 PE teachers with postgraduate qualifications).[15] Translated version (Brazilian Portuguese) language clarity 0.96, pertinence>0.89. Experts CVI for clarity and pertinence were also strong—α=0.93 clarity and α=0.91 pertinence.[14]	Exploratory and CFA.[13–18] High and significant correlation of increasing age and increasing scores.[48] Age and disability differentiation.[15 18] Subtest correlation 0.41.[15] Galloping, running and leaping not well correlated with locomotion subscale. Object control significant and highly correlated.[50] ANOVA—significant age effect for object control.[18] Moderate correlation between items and subset scores, and between subset scores and total score.[18]

ANOVA, analysis of Variance; ASD, autism spectrum disorder; Bayley-III, Bayley Scale of Infant and Toddler Development third edition; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; CFA, Confirmatory Factor Analysis; H_i, scalability coefficient; ID, intellectual disability; MABC-2, Movement Assessment Battery for Children second edition[29]; MAND, McCarron Assessment of Neuromuscular Development[32]; NDI, Neurodevelopmental Index; NSMDA, Neurological Sensory Motor Developmental Assessment[33]; TD, typically developing; TGMD-II, Test of Gross Motor Development second edition[15]; WISC-R, Wechsler Preschool and Primary Scale of Intelligence-R; WPPSI, Wechsler Preschool and Primary Scale of Intelligence[31].

practice to assess children's development as a whole, rather than assessing individual domains in isolation. A gross motor assessment embedded within a developmental assessment, such as that of the Bayley-III may be more appropriate than an isolated gross motor assessment for children where there is suspicion of multiple impairments.

A review by Slater *et al*[8] reported that the TGMD-2 and the MABC (first edition) were recommended for assessing gross motor skills in children with developmental coordination disorder, but found that the MABC needed further evidence of validity. Cools *et al*[27] also published a detailed review of the clinical utility of gross motor assessment tools for children, but did not address the validity, reliability or responsiveness to change of these measures. This review adds to the literature by including updated information on the psychometric properties of the measures and a thorough methodological assessment using the COSMIN checklist, which allows the reader to interpret these results with confidence. We have identified 10 additional publications to support the content, construct and criterion validity of the MABC-2 and have demonstrated an overall higher methodological quality of the papers assessing the MABC-2 when compared with the TGMD-2. Papers that received lower methodological scores on the COSMIN can be attributed to inadequate reporting statistical methods, small sample sizes and non-independent assessors. Further research in this area should consider addressing these limitations in their study design to reduce potential error and increase confidence when interpreting results.

Table 5 Criterion and predictive validity of assessment tools

Test	Criterion	Predictive
Bayley-III	Given but mean age<22 months. Not relevant to study population.[31]	Motor impairment at 4 years: Bayley-III at 2 years<1 SD=sensitivity 0.32–0.037, specificity 0.97<2 SD sensitivity 0.18–0.21 specificity 1.00. CP at 4 years: Bayley-III at 2 years<1 SD sensitivity 0.83 specificity 0.94. <2 SD sensitivity 0.67 specificity 1.0.[4]
BOT-2	MABC-2 p=0.92, PDMS-2 p=0.88 (n=38).[21] PDMS-2 total motor composite r=0.77.[12]	–
MABC-2	PDMS-2 ρ=0.631–0.84.[19 21] TGMD-2 ρ=0.45.[5] TGMD-2 standard scores (r=0.3, p<0.02).[41] BOT-2 ρ=0.90–0.92.[21]	Classification groups (DCD, at risk and TD) remained same over time (6 months) χ^2=0.67, p=0.72.[41] Predictive of motor impairment over 6–12 months (n=41) ICC 0.88 p<0.007.[41] Scores at 4 years predictive of motor impairment at 8 years in children born<30 weeks gestation (PPV 79, sensitivity 79%, specificity 93%).[39]
MAND	Gross motor subscore: low-to-moderate correlation with manual dexterity (–0.46 to 0.35), reaction time (–0.31 to –0.58), intelligence measures (WISC-R, Metropolitan Achievement Test) (0.30 to 0.39) and visual motor test (–0.33 to 0.39).[32]	–
NSMDA	NSMDA at 2 years (n=148) predictive of medical diagnosis χ2=0.08, p=NS.[47]	Motor outcome at 11–13 years: NSMDA at 2 years—sensitivity 48.8%, specificity 82.4%, NSMDA at 4 years sensitivity 64.5%, and specificity 80%. PPV at 2 years 83%, at 4 years 87%.[44] If classified 'severe' at 24 months, approximately 50% chance walking at 4 years (moderate=80%, mild=93%, minimal=100%).[45]
PDMS-2	MABC-2 ρ=0.63–0.84;[19 21] MABC-2 gross motor composite ρ=0.743.[19] BOT-2 ρ=0.88.[21] Mullen Scales of Early Learning GMQ=0.86, FMQ=0.80.[34]	–
TGMD-2	MABC-2 total r=0.49, p<0.01.[5] 'Teacher report' r=0.34–0.45. Physical fitness r=–0.47–0.55.[50] (n=41) Basic motor generalisations subtest of the CSSA r=0.63. Locomotor 0.63 object control 0.41.[15]	–

Bayley-III, Bayley Scale of Infant and Toddler Development third edition; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; CP, cerebral palsy; CSSA, Comprehensive Scales of Student Abilities; ICC, intraclass correlation coefficient; MABC-2, Movement Assessment Battery for Children second edition[29]; MAND, McCarron Assessment of Neuromuscular Development[32]; NDI, Neurodevelopmental Index; NS, not specified; NSMDA, Neurological Sensory Motor Developmental Assessment[33]; PDMS-2, Peabody Developmental Motor Scales second edition[34]; TD, typically developing; [31]; TGMD-II, Test of Gross Motor Development second edition.[15]

Content validity has been established for five of the included assessment tools; however, further research into the content validity for the MAND and NSMDA is required. The NSMDA's ability to predict a diagnosis of CP and motor outcomes over time does support its content validity; however, the methodology scored as poor to fair on the COSMIN and as such content validity cannot be fully established. The use of expert panels, focus groups and/or stakeholder feedback for the BOT-2, MABC-2, TGMD-2 and PDMS-2 demonstrate thorough consideration of the relevance and comprehensiveness of the each test's assessment items during development.

The TGMD-2 is the only assessment tool considered to have well-established construct validity, with several papers reporting factor analysis. The NSMDA has undergone factor analysis for children up to, but not beyond 2 years of age and as such further research is needed to support its validity in older children. All other included assessment tools have undergone factor analysis assessment of their construct validity in one paper and are supported by the ability to discriminate between medical diagnosis or age, and as such are considered to have adequate construct validity. The criterion validity indicates that the TGMD-2 may be measuring a slightly different construct to the other assessment tools included in this study as it has poor agreement with the MABC-2, which in turn has good agreement with the PDMS-2 and the BOT-2. This difference may be related to the inclusion of the

Table 6 Reliability of assessment tools

Test	Internal consistency	Test–retest	Intrarater	Inter-rater	Minimal detectable change	Minimal clinical important difference
Bayley-III	GM α=0.87–0.93 MC: α 0.90–0.96 (24–42 months)[31]	Gross motor subtest (n=47) r=0.79 Motor component r=0.80[31]	–	–	SEM gross motor subtest 0.85–1.08 of motor component=3.00–4.74 (24–42 months)[31]	–
BOT-2	(n=100) α=0.92;[36] (n=141) α=0.86;[21] 4–7 years (n=620) α=0.95, 8–11 years (n=450) α=0.95[12]	(n=100) ICC=0.99;[36] (n=141) ICC=0.97;[21] 4–7 years (n=43) r=0.81, 8–12 years (n=44) r=0.80[12]	–	Total motor composite 4–21 years (n=47) r=0.98[12]	4.18 (sensitivity 55.10%, specificity 72.55%);[36] 7.43 (sensitivity 42.49%, specificity 65.72%)[21]	6.53 (sensitivity 48.98%, specificity 76.47%);[36] 6.55 (sensitivity 49.99%, specificity 58.78%)[21]
MABC-2 (AB 1)	(n=60) M.D α=0.51, A&C α=0.70, Bal α=0.66;[37] (n=1823) α=0.502;[19] (n=50) α=0.81–0.87[28]	(n=60) ICC=0.85;[37] Item ICCs 0.830–0.985;[19] ICC test–retest=0.83;[28] Inter-rater test–retest ICC=0.79[28]	(n=28) κ=0.71[28]	Item ICCs range 0.892–0.998;[19] (n=22) κ=0.60[28]	(n=28) Intrarater MDC=3.43; (n=22) Inter-tester MDC=3.81[28]	-
MABC-2 (AB 2)	Translated version (Japanese) (n=132) α=0.602[38]	–	ICC=0.64[23]	ICC 0.63[23]	Intra-rater SDC TTS: ±11.7 TSS±3.3; Inter-rater SDC TTS±16.0 TSS±3.8[23]	–
MABC-2	Subscales α=0.78 (M.D=0.77, BS=0.52, Bal=0.77);[41] α=0.88;[42] (n=141) α=0.88[21]	n=60 (all three age bands) r=0.80;[29] r=0.74 p<0.0001 (standard score). ICC standard score=0.85;[41] ICC 0.96;[42] n=141 ICC=0.96[21]	ICC 0.88[41]	ICC 0.96–0.99[41]	SEM 1.34 (95% CI)=3;[29] 1.83 (95% CI);[42] 1.83 (sensitivity 69.69% specificity 52.10%)[21]	1.39 (sensitivity 72.47%, specificity 46.18%)[21 42]
MAND	–	–	–	–	–	–
NSMDA	Cross-correlation matrix Item scoring (12+24 months) 0.73 p<0.001, Functional grade (12+24 months) 0.87 p<0.001[46]	–	–	–	–	–
PDMS-2	(n=141) α=0.89;[21] 24–35 m α=0.97, 36–47 m α=0.95, 48–59 m α=0.97, 60–71 m α=0.98. For subgroups* α=0.99[34]	n=141 ICC=0.97[21]	Unable to extract data for ≥24 months[34]	Unable to extract data for ≥24 months[34]	7.76 (sensitivity 60.65%, specificity 74.13%);[21] SEM 24–59 months=3, 60–71 m=2[34]	8.39 (sensitivity 61.65%, specificity 71.34%)[21]
TGMD-2	(n=1438) α=0.80;[43] n=75 Locomotor subset α=0.71, object control α=0.72;[21] n=120 α=0.72;[45] n=99 α=0.90;[23] n=1208 Cronbach's α=0.91 (gross motor quotient). Locomotor 0.85 and object control 0.88. Note SEM GMQ=4–5 SEM subsets=1[15]	n=63 ICC=0.81 95% CI;[48] n=23 ICC=0.92 total 95% CI;[16] n=99 r=0.98.[18] Locomotor test r=0.90 p<0.0001, object control test r=0.91 p<0.001.[14] n=75 r=0.96 overall (3–5 years r=0.91), 6–8 years r=0.95), (9–10 years r=0.94)[15]	n=32 ICC=0.97 95% CI;[48] n=25 ICC=0.95 95% CI;[16] ICC=0.78[49] ICC=0.92–0.99[14]	Obj ICC=0.93;[24] (n=50) ICC=0.89;[16] ICC=0.75;[49] n=8 r=1.00.[18] LS ICC=0.88, Obj ICC=0.89;[14] n=30 r=0.98[15]	–	–

*Gender, ethnicity, speech/language or physical disorder.[31]
A&C, aiming and catching; BAL, balance; Bayley-III, Bayley Scale of Infant and Toddler Development third edition; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; BS, Ball Skills; GM, gross motor subset; LS, locomotion subset; MABC-2, Movement Assessment Battery for Children second edition[29]; MAND, McCarron Assessment of Neuromuscular Development[32]; MC, motor component; MD, manual dexterity; NSMDA, Neurological Sensory Motor Developmental Assessment[33]; Obj, object control subset; PDMS-2, Peabody Developmental Motor Scales second edition[34]; SDC, Smallest Detectable Change; TGMD-II, Test of Gross Motor Development second edition[15]; TSS, total standard score; TTS, total test score.

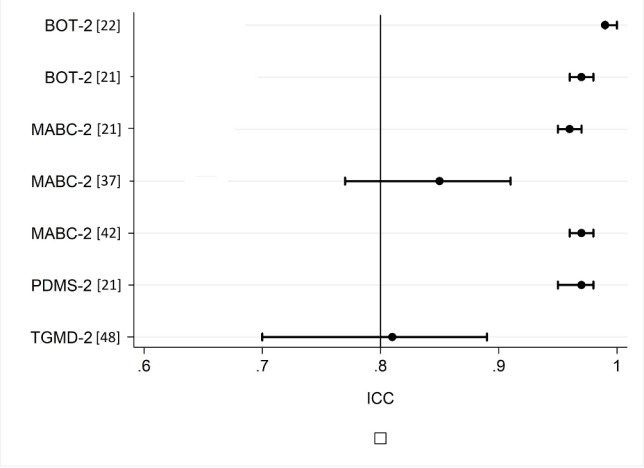

Figure 2 Test–retest reliability of gross motor assessment tools. BOT-2, Bruininks-Oseretsky Test of Motor Proficiency second edition[12]; ICC, intraclass correlation coefficient; MABC-2, Movement Assessment Battery for Children second edition[29]; PDMS-2, Peabody Developmental Motor Scales second edition[34]; TGMD-II, Test of Gross Motor Development second edition.[15]

assessment of quality of movement in the TGMD-2, or the inclusion of balance and/or fine motor tasks on the other assessments. There is scope to investigate the criterion validity of the MAND and the gross motor subsections of the Bayley-III and the NSMDA with the other assessment tools in this study in the future.

The BOT-2 was the only assessment tool to have its reliability assessed with excellent methodology. In conjunction with its reported results, it can be considered to have the strongest evidence for internal consistency and test–retest reliability out of the included assessment tools. The PDMS-2 and the MABC-2 can be considered to have the next best established test–retest reliability with good methodological quality. The reported test–retest

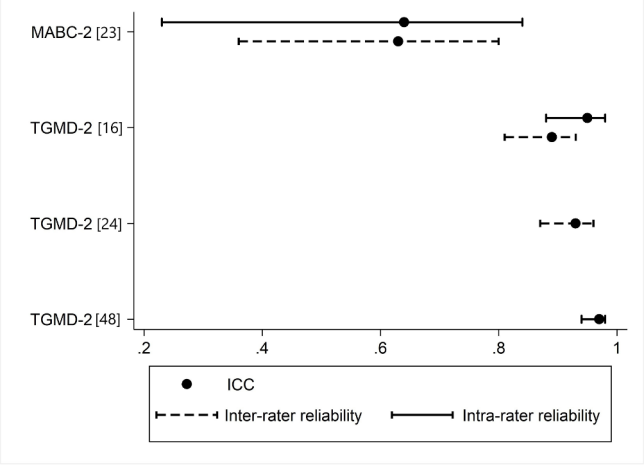

Figure 3 Inter-rater and intrarater reliability of gross motor assessment tools. ICC, intraclass correlation coefficient; MABC-2, Movement Assessment Battery for Children second edition[29]; TGMD-II, Test of Gross Motor Development second edition.[15]

reliability values for the TGMD-2 are impacted by the poor to fair methodological quality, and further high-quality research needs to be done to support its body of evidence. Test–rest, inter-rater or intrarater reliability has not been assessed in the MAND and NSMDA. In the clinical context, gross motor assessments are often repeated over time or between therapists and as such these measures of reliability should be established. The Bayley-III would also benefit from further research into its reliability, with no published inter-rater or intrarater reliability measures, and with only one, fair quality report of good test–retest reliability.

As yet, there is little evidence to support the use of these assessments as outcome measures. The inclusion in some of the articles of minimal detectable change (MDC) and minimal clinically important difference (MCID) is valuable for clinicians.[7] The difference between MDC and MCID is also of importance, as a change in score does not necessarily relate to a meaningful change for the child or their family. Only the Bayley-III, BOT-2, MABC-2 and PDMS-2 have a reported MCID with satisfactory sensitivity and specificity; however, due to the fair methodological quality used to obtain these values they cannot be used with a high level of confidence until further studies have been performed. The TGMD-2 was created in part to be used as an outcome measure; however, there are no articles to date investigating its responsiveness to change.[15] It should also be noted that all of the included assessment tools measure impairment and activity limitations, but do not specifically address the other elements of the International Classification of Functioning, Disability and Health domains of participation, personal factors and environment.[2] Clinicians should use appropriate assessments or questionnaires to ensure that these domains of health are also addressed in line with the WHO guidelines.[2]

When considering a test's reliability all three elements of test error should be taken into account— these can be described as time sampling (assessed with test–retest reliability), content sampling (assessed as internal consistency) and interscorer difference (or inter-rater reliability).[15] This is one of the reasons that clinicians should consider repeating assessments and/or completing a second alternative assessment. All assessments should be interpreted in conjunction with clinical reasoning and observation. Included assessment tools are not intended to be diagnostic on their own; results need to be combined with other assessments and expert opinion to arrive at a clinical diagnosis.

The clinical utility varied across all of the included assessment tools, with the primary differences being in cost and time to administer the assessments. Clinicians and researches should select their assessment tool with consideration of psychometric properties (inclusive of the methodological rigour behind them), clinical utility and for the population, situation and age group in question.

A potential limitation of this study was that one author screened the titles and abstracts, which may have led to a sampling bias. While care was taken to include all

potentially relevant papers and assessment tools until the second round of assessment with two authors, the potential for exclusion of papers relevant to this review remains. The process of excluding both papers and assessment tools in this single step may also be seen as a limitation, as the total number of assessment tools (or different versions of tools) was not reported. This process does, however comply with the COSMIN and PRISMA guidelines. A second limitation was the restriction of included papers and manuals to those published in English. Unfortunately, this resulted in the exclusion of three assessment tools that have been reported as commonly used in Europe: The Motoriktest für Vier- bis Sechjärige Kinder 4–6, the Körperkoordinationtest für Kinder and the Maastrichtse Motoriek Test.[27] The authors also note the third edition of the TGMD is soon to be published and will need to be subjected to a similar level of assessment of psychometric properties in the future.

Clinicians and parents who need guidance to set realistic therapy goals and to understand future intervention requirements benefit from understanding a test's predictive ability. The NSMDA and the MABC-2 are the only tools that have demonstrated long-term (≥4 years follow-up) predictive validity, while the Bayley-III has good predictive validity at 2 years for future movement difficulties and for the diagnosis of cerebral palsy at 4 years. However, further research into the long-term predictive validity of all included gross motor assessment tools is warranted.

While validity and reliability should guide selection of assessment tools, clinical utility must also be taken into consideration. Most tests have ongoing costs associated with forms and equipment replacement, which may be prohibitive to some users. The NSMDA requires the therapist to handle the child for several items, which should be considered in relation to manual handling policies of institutions. Assessment burden for children and families should also be taken into consideration when selecting an assessment tool. Younger children are more likely to be distracted and may not understand test items as well, which may also increase assessment times.[28]

When a new edition of an assessment tool is released resulting in a change in age groups, scoring or tasks, it is insufficient to rely on the psychometric assessments that were performed on the original test. The MABC-2 manual provides justification for the inclusion of reliability and validity assessment of the original MABC[29]; however, owing to the significant changes in age groups and tasks between editions these were not included for the analysis of the MABC-2 in this review. Two studies quoted in the MABC-2 manual to support the validity and reliability are both unpublished works and as such are also unable to be included in this systematic review. This could indicate a publication for the MABC-2.

The thorough methodological assessment of the included articles using the COSMIN checklist should be seen as a strength of this paper, as should the range of assessment tools included in this review. While it has previously been argued that the 'worst score counts' criteria in the COSMIN creates a floor effect,[30] the COSMIN authors argue that only 'fatal flaws' contribute to an overall score of poor.[10] There are few tools available to assess the psychometric properties of assessment tools and arguably none so robustly validated as the COSMIN.

There are many appropriate gross motor assessment tools available for use in research and clinical settings today. Most of the available tools demonstrate adequate validity and reliability in children aged 2–12 years and as such the authors do not believe that new assessment tools need to be developed for use. There is scope however to improve the evidence of inter-rater and intra-rater reliability and predictive validity should be ascertained over a longer period of time and with greater methodological rigour. Tools also need clearer assessment of their responsiveness to change to assist clinicians and researchers with outcome measure selection. Researchers should be mindful of the methods they use to assess validity and reliability. Clarity of reporting, statistical methods and sample sizes should be carefully considered to ensure the highest quality of evidence.

CONCLUSION

Currently available gross motor assessment tools for children have good to excellent content and construct validity. The BOT-2, MABC-2, PDMS-2 and TGMD-2 are the most reliable assessments in this age group. The Bayley-III has the best predictive validity at 2 years of age, and the NSMDA and the MABC-2 both have good predictive validity at 4 years of age. There is scope for further research into the predictive validity, reliability and responsiveness of gross motor assessment tools in preschool and school-aged children. In practice, clinicians should choose assessments with consideration of their psychometric properties in the context of the child that they are assessing.

Contributors All individuals listed as authors meet the appropriate authorship criteria and have approved the acknowledgement of their contributions. The primary author, AG, was responsible for the drafting of the paper and liaising with the coauthors on findings and conclusions. RT contributed to the paper through interpretation of data, completing methodological assessments and revising manuscript content throughout its development. PEM and AJS both contributed to the paper through assisting with the development of research design, interpretation of data and revising manuscript content through its development.

Funding This study was part-funded by grants from the National Health and Medical Research Council Career Development Fellowship (AJS) 1053767 and Centre of Research Excellence in Newborn Medicine 1060733 (AJS and AG) and the Victorian Government's Operational Infrastructure Support Programme.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement This paper includes data obtained from reviewing papers of published manuscripts. Data can be accessed by contacting the primary author.

REFERENCES

1. Piek JP, Baynam GB, Barrett NC. The relationship between fine and gross motor ability, self-perceptions and self-worth in children and adolescents. *Hum Mov Sci* 2006;25:65–75.
2. World Health Organization. *International classification of functioning, disability and health: ICF*. Geneva: World Health Organization, 2001.
3. Magalhães LC, Cardoso AA, Missiuna C. Activities and participation in children with developmental coordination disorder: a systematic review. *Res Dev Disabil* 2011;32:1309–16.
4. Spittle AJ, Spencer-Smith MM, Eeles AL, *et al*. Does the Bayley-III Motor Scale at 2 years predict motor outcome at 4 years in very preterm children? *Dev Med Child Neurol* 2013;55:448–52.
5. Logan SW, Robinson LE, Getchell N. The comparison of performances of preschool children on two motor assessments. *Percept Mot Skills* 2011;113:715–23.
6. Venetsanou F, Kambas A. Environmental factors affecting preschoolers' motor development. *Early Child Educ J* 2010;37:319–27.
7. Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
8. Slater LM, Hillier SL, Civetta LR. The clinimetric properties of performance-based gross motor tests used for children with developmental coordination disorder: a systematic review. *Pediatr Phys Ther* 2010;22:170–9.
9. Brown T, Lalor A. The Movement Assessment Battery for Children–Second Edition (MABC-2): a review and critique. *Phys Occup Ther Pediatr* 2009;29:86–103.
10. Terwee CB, Mokkink LB, Knol DL, *et al*. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21:651–7.
11. Law M. *Outcome measures rating form*. Ontario, Canada: CanChild Centre for Disability Research, 2004.
12. Bruininks R, Bruininks B. *Bruininks-Oseretsky Test of Motor Proficiency–2nd Edition (BOT-2): Manual*. Circle Pines: MN: AGS Publishing, 2005.
13. Rudd J, Butson ML, Barnett L, *et al*. A holistic measurement model of movement competency in children. *J Sports Sci* 2016;34:477–85.
14. Valentini NC. Validity and reliability of the TGMD-2 for Brazilian children. *J Mot Behav* 2012;44:275–80.
15. Ulrich DA. *Test of gross motor development-2*. Austin: Prod-Ed, 2000.
16. Houwen S, Hartman E, Jonker L, *et al*. Reliability and validity of the TGMD-2 in primary-school-age children with visual impairments. *Adapt Phys Activ Q* 2010;27:143–59.
17. Wong KYA, Yin Cheung S. Confirmatory factor analysis of the test of gross motor development-2. *Meas Phys Educ Exerc Sci* 2010;14:202–9.
18. Simons J, Daly D, Theodorou F, *et al*. Validity and reliability of the TGMD-2 in 7-10-year-old Flemish children with intellectual disability. *Adapt Phys Activ Q* 2008;25:71–82.
19. Hua J, Gu G, Meng W, *et al*. Age band 1 of the movement assessment battery for children-second edition: exploring its usefulness in mainland China. *Res Dev Disabil* 2013;34:801–8.
20. Niemeijer AS, van Waelvelde H, Smits-Engelsman BC. Crossing the North Sea seems to make DCD disappear: cross-validation of Movement Assessment Battery for Children-2 norms. *Hum Mov Sci* 2015;39:177–88.
21. Wuang YP, Su CY, Huang MH. Psychometric comparisons of three measures for assessing motor functions in preschoolers with intellectual disabilities. *J Intellect Disabil Res* 2012;56:567–78.
22. Wuang YP, Lin YH, Su CY. Rasch analysis of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in intellectual disabilities. *Res Dev Disabil* 2009;30:1132–44.
23. Holm I, Tveter AT, Aulie VS, *et al*. High intra- and inter-rater chance variation of the movement assessment battery for children 2, ageband 2. *Res Dev Disabil* 2013;34:795–800.
24. Barnett LM, Minto C, Lander N, *et al*. Interrater reliability assessment using the Test of Gross Motor Development-2. *J Sci Med Sport* 2014;17:667–70.
25. Spittle AJ, Doyle LW, Boyd RN. A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life. *Dev Med Child Neurol* 2008;50:254–66.
26. McDowell I. *Measuring health: a guide to rating scales and questionnaires*. Oxford: Oxford university press, 2006.
27. Cools W, Martelaer KD, Samaey C, *et al*. Movement skill assessment of typically developing preschool children: a review of seven movement skill assessment tools. *J Sports Sci Med* 2009;8:154.
28. Smits-Engelsman BC, Niemeijer AS, van Waelvelde H. Is the movement assessment battery for Children-2nd edition a reliable instrument to measure motor performance in 3 year old children? *Res Dev Disabil* 2011;32:1370–7.
29. Henderson SE, Sugden DA, Barnett AL. *Movement assessment battery for children-2: Movement ABC-2: Examiner's manual*: Pearson, 2007.
30. Adair B, Said CM, Rodda J, *et al*. Psychometric properties of functional mobility tools in hereditary spastic paraplegia and other childhood neurological conditions. *Dev Med Child Neurol* 2012;54:596–605.
31. Bayley N. *Bayley scales of infant development and toddler development: technical manual*: The PsychCorp, 2006.
32. McCarron LT. *MAND: McCarron assessment of neuromuscular development, fine and gross motor abilities*: McCarron-Dial Systems, Incorporated, 1997.
33. Burns YR. *N.S.M.D.A Physiotherapy Assessment for Infants and Young Children*. 2nd edn. Brisbane, Queensland: CopyRight Publishing Company, 2014.
34. Folio M, Fewell R. Peabody Developmental Motor Scales. *Examiner's Manual*. 2nd Edn. Austin, Texas: Pro-Ed, 2000.
35. Visser L, Ruiter SA, Van der Meulen BF, *et al*. Low verbal assessment with the Bayley-III. *Res Dev Disabil* 2015;36C:230–43.
36. Wuang YP, Su CY. Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in children with intellectual disability. *Res Dev Disabil* 2009;30:847–55.
37. Ellinoudis T, Evaggelinou C, Kourtessis T, *et al*. Reliability and validity of age band 1 of the Movement Assessment Battery for Children–second edition. *Res Dev Disabil* 2011;32:1046–51.
38. Kita Y, Suzuki K, Hirata S, *et al*. Applicability of the movement assessment battery for children-second edition to Japanese children: a study of the age band 2. *Brain Dev* 2016;38:706–13.
39. Griffiths A, Morgan P, Anderson PJ, *et al*. Predictive value of the movement assessment battery for children - second edition at 4 years, for motor impairment at 8 years in children born preterm. *Dev Med Child Neurol* 2017;59:490–6.
40. Schulz J, Henderson SE, Sugden DA, *et al*. Structural validity of the Movement ABC-2 test: factor structure comparisons across three age groups. *Res Dev Disabil* 2011;32:1361–9.
41. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation, reliability, and validity for Brazilian children. *Res Dev Disabil* 2014;35:733–40.
42. Wuang YP, Su JH, Su CY. Reliability and responsiveness of the movement assessment battery for Children-Second Edition Test in children with developmental coordination disorder. *Dev Med Child Neurol* 2012;54:160–5.
43. Hands B, Larkin D, Rose E. The psychometric properties of the McCarron assessment of neuromuscular development as a longitudinal measure with Australian youth. *Hum Mov Sci* 2013;32:485–97.
44. Danks M, Maideen MF, Burns YR, *et al*. The long-term predictive validity of early motor development in "apparently normal" ELBW survivors. *Early Hum Dev* 2012;88:637–41.
45. MacDonald J, Burns Y. Performance on the NSMDA During the First and Second Year of Life to Predict Functional Ability at the Age Of 4 in Children with Cerebral Palsy. *Hong Kong Physiotherapy Journal* 2005;23:40–5.
46. Burns YR, Ensbey RM, Norrie MA. The neuro-sensory motor developmental assessment part 1: development and administration of the test. *Aust J Physiother* 1989;35:141–9.
47. Burns YR, Ensbey RM, Norrie MA. The neuro-sensory motor developmental assessment part II: predictive and concurrent validity. *Aust J Physiother* 1989;35:151–7.
48. Farrokhi A, Zareh Zadeh M, Karimi Alvar L, *et al*. Reliability and validity of test of gross motor development-2 (Ulrich, 2000) among 3-10 aged children of Tehran City. *Journal of Physical Education and Sports Management* 2014;5:18–28.
49. Kim Y, Park I, Kang M. Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapt Phys Activ Q* 2012;29:346–65.
50. Kim CI, Han DW, Park IH. Reliability and validity of the test of gross motor development-II in Korean preschool children: applying AHP. *Res Dev Disabil* 2014;35:800–7.