BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-021734
Article Type:	Research
Date Submitted by the Author:	15-Jan-2018
Complete List of Authors:	Griffiths, Alison; Monash University, Physiotherapy; The Royal Children's Hospital, Physiotherapy Toovey, Rachel; Murdoch Childrens Research Institute, Developmental Disability and Rehabilitation Research; University of Melbourne, Physiotherapy Morgan, Prue; Monash University, Physiotherapy Spittle, Alicia; University of Melbourne, Physiotherapy; Murdoch Childrens Research Institute, Victorian Infant Brain Studies
Keywords:	PAEDIATRICS, REHABILITATION MEDICINE, validity, gross motor assessment, reliability

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]

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

Corresponding Author*:*

Name: A/Prof Alicia J. Spittle

Address: Department of Physiotherapy, The University of Melbourne, Level 7 Alan Gilbert Building, 161 Barry Street, Parkville, Vic 3010, Australia

Email: aspittle@unimelb.edu.au

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 Program.

Financial disclosure: The authors have declared that they have no financial relationships relevant to this article.

Conflict of interest: The authors have no conflict of interest.

Keywords: paediatrics, reliability, validity, rehabilitation medicine, gross motor assessment

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 2-12 years.

Method:

A systematic search of MEDLINE, Embase, CINAHL and AMED was performed. Methodological quality was assessed with the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) 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 re-test reliability is excellent in the BOT-2 (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 interrater (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 COSMIN 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 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.

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 to 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]. Further, assessment tools are essential to monitor the effects of intervention [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].

Health 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 [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 Movement Assessment Battery for Children (MABC) in age range, age bands, materials and tasks, that the MABC-2 requires independent reliability and validity assessment. Over the past eight years there has 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 2-12 years. The secondary aim of this review is to identify the methodological quality of the included studies and 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). 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 required regarding availability of manuals authors were contacted.

Assessment tools were included if they were 1. Discriminative, predictive or evaluative of gross motor skills, 2. Assessed $\geq$ two gross motor (e.g. balance, jumping etc.) items, 3. Able to extract a meaningful gross motor sub-score, 4. Applicable to children 2-12 years of age, 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 (e.g. 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. 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 agreement by 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 nine measurement properties: content, construct, cross cultural and criterion validity, hypothesis testing, internal consistency, reliability, measurement error and responsiveness [7] (Supplementary Table 1).

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]. Clinical utility includes the cost of manuals, kits, training requirements, time to administer the assessment and the ease of scoring. All reported values for reliability were collected with Intraclass Correlation Coefficients (ICC) directly compared.

Results

Figure 1 provides details of study selection. Eight 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), 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). The corresponding manuals were then added to the final yield resulting in 30 papers and 7 manuals. Nineteen assessment tools were excluded (Supplementary Table 2).

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

Most of the tools assess at least two domains of function: gross motor and fine motor skills, although the Bayley III and the NSMDA assess six to seven different domains of development. The TGMD-2 is the only tool that only assesses gross motor skills.

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 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.

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 [12]	Gross motor, fine motor, cognitive, communication, social/emotional, adaptive	Static postures, dynamic movement, balance	1 mth – 3 yrs	Developmental delay: <25th centile or below 2SD. *	Discriminative	Predictive, Evaluative, Research tool	Norm	1700 children from the USA (2000)
BOT-2 [13]	Gross motor, fine motor	Coordination, balance, running speed and agility, strength	4 – 21 yrs	*	Discriminative Evaluative	Research tool	Norm	1520 children from the USA (2005)
MABC-2 [14]	Gross motor, fine motor, balance	Aiming and catching, static and dynamic balance	3 – 16 yrs	Traffic light system: Green = normal, amber = 'at risk' and red = definite motor impairment (<15%). *	Discriminative Evaluative	Intervention planning, Research tool	Norm	1172 children from United Kingdom (2006)
MAND [15]	Gross and fine motor	Coordination, jumping, static and dynamic balance	3 yrs – 25 yrs	NDI 70-85 = mild 55-69 = moderate <55 = severe disability *	Evaluative	Research tool	Norm	2000 3-35 yrs from the USA (1970's)
NSMDA [16]	Gross Motor, Fine Motor, Neurological, Postural Development, Infant Patterns of Movement, Sensory Motor. †	Sitting, kneeling, walking, balance, running, hopping, jumping, catching, motor planning	1 mth – 6 yrs	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	N/A
PDMS-2 [17]	Gross motor, fine motor	Stationary (standing balance, sit-ups, push-ups), locomotion (walking, running, jumping, hopping, etc.), object manipulation (kick, throw, hit, catch)	Birth – 5 yrs	*	Discriminative Evaluative	Predictive, Research tool	Norm	2003 USA and Canada (1997-8)

TGMD-2 [19]	Gross Motor	Locomotion (run, gallop, hop leap, jump, slide) and Object control (batting, dribbling, catch, kick, throw, roll)	3 – 10 yrs	*	Discriminative Evaluative	Outcome measure, research tool, intervention planning	Norm	1208 USA children (1997-1998)

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[19] NDI, Neurodevelopmental Index; SD, Standard Deviation; mth, month; yrs, years *, Advisable to use clinical reasoning; †, requires some manual handling; USA, United States of America

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 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 Supplementary Table 3. The shortest administration time is for the TGMD-2 which has 10 items; whilst most manuals report 20-60 minutes 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 post assessment (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, having 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 Supplementary Table 4. Most assessments were developed by or with input from experts in the field, with most also performing literature reviews. Bruininks and Bruininks [13] 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 MABC-2 and the TGMD-2 have the most evidence for construct validity, with the MABC-2 requiring some changes to remain valid in the Chinese and Dutch speaking populations [20 21]. The BOT-2, MABC-2 and TGMD-2 all provide evidence of discriminant validity in particular age or diagnosis groups. 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 (Suppl. Table 5). Concurrent validity between the MABC-2, PDMS-2 and BOT-2 is moderate to high, whilst the TGMD-2 is only weakly correlated with the MABC-2 [5] (Suppl. Table 5). The PDMS-2, TMGD-2 and NSMDA report correlations with other criteria such as paediatrician diagnosis, physical fitness or psychomotor/intelligence tests.

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 [12]	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 [13]	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 [14]	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 [15]	15-20	Therapist administers items in standardised order.	Professionals e.g. education, neurology, OT, PT, psychology etc.	Formal training not required.	Manual and test kit: £1366 includes equipment
NSMDA [16]	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 [17]	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 [19]	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 3rd edition [12]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; MAND, McCarron Assessment of Neuromuscular Development [15]; NSMDA, Neurological Sensory Motor Developmental Assessment [16]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17]; TGMD-II, Test of Gross Motor Development 2nd edition [19]; GM, Gross motor; OT, Occupational Therapy; PT, Physiotherapy; PE, Physical Education

Table 3. Methodological quality of included articles

Test	First author, Year	Country	Population (Age, Diagnosis)	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsive-ness
BAYLEY III	Bayley [12]	USA	1-42 mths	Fair	Fair	Good	Excellent	Good	Good	-	Good	-
	Spittle, et al. [4]	Australia	2,4 yrs, Ex prem	-	-	-	-	-	-	-	Good	-
	Visser, et al. [23]	Netherlands	2.2-10.8 yrs, GDD, L.I.	-	-	-	Excellent	Poor	-	-	-	-
BOT-2	Wuang and Su [24]	Taiwan	4-12 yrs ID	Excellent	Excellent	Excellent	-	-	-	-	-	Fair
	Wuang, et al. [25]	Taiwan	3-6 yrs ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Bruininks and Bruininks [13]	USA	4-21 yrs	Good	Fair (interrater) Fair (test-retest)	Good	Excellent	Good	-	-	Good	-
MABC-2 (AB 1)	Ellinoudis, et al. [26]	Greece	3-5.5 yrs	Excellent	Good	-	-	-	-	-	-	-
	Hua, et al. [20]	China	3-6 yrs	Excellent	Good	-	Excellent	Excellent	-	Poor	Excellent	-
	Logan, et al. [5]	USA	3-6 yrs	-	-	-	-	-	Fair	-	Fair	-
	Smits-Engelsman, et al. [27]	Belgium	3-4 yrs	Poor	Poor	Poor	-	-	-	-	-	-
MABC-2 (AB 2)	Holm, et al. [28]	Norway	7-9 yrs	-	Fair (interrater) Poor (intrarater)	Poor	-	-	-	-	-	-
	Kita, et al. [29]	Japan	7-10 yrs	Excellent	-	-	-	-	-	Poor	-	-
MABC-2	Griffiths, et al. [30]	Australia	4-8 yrs	-	-	-	-	-	-	-	Good	-
	Henderson, et al. [14]	UK	3-16 yrs	-	Fair	Good	Excellent	-	-	-	-	-
	Niemeijer, et al. [21]	Netherlands + Belgium	-	-	-	-	-	-	-	Poor	-	-
	Schulz, et al. [31]	U.K	3-16 yrs	-	-	-	Excellent	Good	-	-	-	-
	Valentini, et al. [32]	Brazil	3-13 yrs	Fair	Fair	-	Fair	Poor	-	Poor	Poor	-

	Wuang, et al. [25]	Taiwan	3-6 yrs, ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Wuang, et al. [33]	Taiwan	6-12 yrs DCD	Poor	Fair	Good	-	-	-	-	-	Fair
MAND	Hands, et al. [34]	Australia	10-17 yrs	-	-	-	-	Excellent	-	-	-	-
	McCarron [15]	USA	7yrs	-	-	-	Fair	Poor	-	-	Poor	-
NSMDA	Danks, et al. [35]	Australia	2 + 4 yrs ELBW	-	-	-	-	-	-	-	Fair	-
	MacDonald and Burns [36]	Australia	2 + 4 yrs CP	-	-	-	-	Fair	-	-	Poor	-
	Burns, et al. [37]	Australia	1-24 mths VLBW	Poor	-	-	Poor	-	-	-	-	-
	Burns, et al. [38]	Australia	1-mnths VLBW	-	-	-	-	Poor	-	-	Fair	-
PDMS-2	Hua, et al. [20]	China	3-6 yrs.	Excellent	Good	-	Excellent	Excellent	-	Poor	Excellent	-
	Wuang, et al. [25]	Taiwan	3-6 yrs ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Folio and Fewell [17]	USA	0-71 mnths	Good	-	Poor	Excellent	Good	Good	-	Poor	-
TGMD-2	Barnett, et al. [39]	Australia	4-8 yrs	-	Fair	-	-	-	-	-	-	-
	Farrokhi, et al. [40]	Iran	3-11 yrs	Fair	Fair	-	Fair	Fair	-	-	-	-
	Houwen, et al. [41]	Netherlands	6-12 yrs VI	Fair	Fair	-	-	Fair	-	-	-	-
	Kim, et al. [42]	Korea	8-12 yrs ID	-	Poor	-	-	-	-	-	-	-
	Kim, et al. [43]	Korea	5-6 yrs	Poor	Fair	-	-	Poor	-	-	Poor	-
	Logan, et al. [5]	USA	3-6 yrs	-	-	-	-	-	Fair	-	Fair	-
	Rudd, et al. [44]	Australia	6-12 yrs	-	-	-	-	Good	-	-	-	-
	Simons, et al. [45]	Belgium	7-10 yrs ID	Good	Good (interrater) Poor (test-retest)	-	Excellent	Good	Good	-	-	-
	Valentini [46]	Brazil	3-10 yrs	Poor	Fair (test-retest) Good (intra,	-	Excellent	Good	-	Fair	Good	-

				interrater)							
Wong and Yin Cheung [47]	China	3-10 yrs	-	-	-	-	Fair	-	-	-	-
Ulrich [19]	USA	3-10 yrs	Good	Fair (test-retest) Poor (interrater)	Fair	Poor	Good	-	-	Fair	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17]; TGMD-II, Test of Gross Motor Development 2nd edition;[19] Mths, Months; yrs, years; DCD, Developmental Coordination Disorder; VI, Vision Impairment; ID, Intellectual Disability; GDD, global developmental delay; L.I, Language Impairment; ELBW, Extremely Low Birth Weight; VLBW, Very Low Birth Weight; CP, Cerebral Palsy; prem, premature; USA, United States of America

Reliability

Internal consistency of assessments are summarised in supplementary table 6. The BOT-2's high internal consistency is well supported, including for children with an intellectual disability [25 48]. The MABC-2 appears to have lower internal consistency than the BOT-2, which may be related 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 MAND does not currently have published internal consistency data in this age group. 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 reliability findings are summarised in Supplementary Table 6 and in Figures 2 and 3. Test-retest reliability was excellent in the Bayley-III (Supplementary Table 6), BOT-2 and PDMS-2; and was good to excellent in the MABC-2 and TGMD-2 (Figure 2). Intra-rater 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) [28 39]. Inter-rater reliability is also supported in the BOT-2 with Pearson 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 [49 50]. Only studies reporting ICC's are visually represented in Figures 2 (test-retest) and 3 (inter and intra-rater). The TGMD-2 test-retest reliability results from Houwen, et al. [41] were believed to contain an error as the reported ICC was outside of the reported confidence intervals (ICC 0.92, 0.82-0.91). 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 standard error of measurement (SEM) [25]. There have been no studies to date on the responsiveness of the TGMD-2, NSMDA or MAND.

Discussion

This review identified eight gross motor assessment tools appropriate for use in clinical or research settings, each with their own strengths and limitations. Interestingly, only one of the eight

assessments measured gross motor skills in isolation. This is likely a reflection on current practice to assess children's development as a whole, rather than assessing individual domains in isolation.

The current review adds to the literature by including a thorough methodological assessment using the COSMIN checklist. Our finding are consistent with an earlier review by Slater, et al. [8] who reported that the psychometric properties of the TGMD-2 and the BOT-2 were robust in children with developmental coordination disorder. The MABC-2 and the PDMS-2 were also identified as well supported assessment tools in this review. All assessment tools were found to have merits and limitations and should be chosen with consideration for their psychometric properties, clinical utility and for the population and age group in question.

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 tools. Younger children are more likely to be distracted and may not understand test items as well, which may also increase assessment times[27].

When a new edition of an assessment tools 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 [14], 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 bias for the MABC-2.

As yet there is little evidence to support the use of these assessments as outcome measures. 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 [19]. The inclusion in some of the articles of minimal detectable change (MDC) and minimal clinically important difference (MCID) is valuable for clinicians. The difference between the two values is also of importance, as a change in score does not necessarily relate to a meaningful change for the child or their family. 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 (ICF) domains of participation, personal factors and environment [2]. Clinicians should utilise appropriate assessments or questionnaires to ensure that these domains of health are also addressed in line with World Health Organisation 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 inter-scorer difference (or interrater reliability) [19]. 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.

In this review 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.

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 [51], 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. The available tools demonstrate adequate validity and reliability 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 and intrarater 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 motor assessment tools 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.

Author Contributions

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

Figures

Figure 1. PRISMA flow diagram detailing study selection

Figure 2. Test re-test reliability of gross motor assessment tools

Figure 2 legend: BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17]; TGMD-II, Test of Gross Motor Development 2nd edition [19].

Figure 3. Inter and interrater reliability of gross motor assessment tools

Figure 3 legend: MABC-2, Movement Assessment Battery for Children 2nd edition [14]; TGMD-II, Test of Gross Motor Development 2nd edition [19]

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(1):65-75. doi: http://dx.doi.org/10.1016/j.humov.2005.10.011
2. World Health Organization. International Classification of Functioning, Disability and Health: ICF: World Health Organization 2001.
3. Magalhaes LC, Cardoso AA, Missiuna C. Activities and participation in children with developmental coordination disorder: a systematic review. *Res Dev Disabil* 2011;32(4):1309-16. doi: 10.1016/j.ridd.2011.01.029 [published Online First: 2011/02/19]
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? *Developmental Medicine And Child Neurology* 2013;55(5):448-52. doi: 10.1111/dmcn.12049
5. Logan SW, Robinson LE, Getchell N. The Comparison of Performances of Preschool Children on Two Motor Assessments. *Perceptual and Motor Skills* 2011;113(3):715-23.
6. Venetsanou F, Kambas A. Environmental factors affecting preschoolers' motor development. *Early Childhood Education Journal* 2010;37(4):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(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006 [published Online First: 2010/05/25]
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* 2010;22(2):170.
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(1):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(4):651-7. doi: 10.1007/s11136-011-9960-1 [published Online First: 2011/07/07]
11. Law M. Outcome measures rating form. Ontario, Canada: CanChild Centre for Disability Research, 2004.
12. Bayley N. Bayley Scales of Infant Development and Toddler Development: Technical Manual: The PsychCorp 2006.
13. Bruininks R, Bruininks B. Bruininks-Oseretsky Test of Motor Proficiency–2nd Edition (BOT-2): Manual. Circle Pines: MN: AGS Publishing 2005.
14. Henderson SE, Sugden DA, Barnett AL. Movement assessment battery for children-2: Movement ABC-2: Examiner's manual: Pearson 2007.

15. McCarron LT. MAND: McCarron assessment of neuromuscular development, fine and gross motor abilities: McCarron-Dial Systems, Incorporated 1997.
16. Burns YR. N.S.M.D.A Physiotherapy Assessment for Infants and Young Children Second Edition. Brisbane, Queensland: CopyRight Publishing Company 2014.
17. Folio M, Fewell R. Peabody Developmental Motor Scales. Examiner's Manual . 2nd Edition. Austin, Texas.: Pro-Ed. 2000.
18. Khan NZ, Muslima H, El Arifeen S, et al. Validation of a rapid neurodevelopmental assessment tool for 5 to 9 year-old children in Bangladesh. *J Pediatr* 2014;164(5):1165-70.e6. doi: http://dx.doi.org/10.1016/j.jpeds.2013.12.037
19. Ulrich DA. Test of gross motor development-2. *Austin: Prod-Ed* 2000
20. 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(2):801-8. doi: http://dx.doi.org/10.1016/j.ridd.2012.10.012
21. 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. doi: 10.1016/j.humov.2014.11.004
22. Khan NZ, Muslima H, Shilpi AB, et al. Validation of rapid neurodevelopmental assessment for 2- to 5-year-old children in Bangladesh. *Pediatrics* 2013;131(2):e486-94. doi: http://dx.doi.org/10.1542/peds.2011-2421
23. Visser L, Ruiter SAJ, Van der Meulen BF, et al. Low verbal assessment with the Bayley-III. *Res Dev Disabil* 2015;36:230-43.
24. 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(5):847-55. doi: http://dx.doi.org/10.1016/j.ridd.2008.12.002
25. 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(6):567-78. doi: http://dx.doi.org/10.1111/j.1365-2788.2011.01491.x
26. 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(3):1046-51. doi: http://dx.doi.org/10.1016/j.ridd.2011.01.035
27. Smits-Engelsman BCM, 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(4):1370-77. doi: http://dx.doi.org/10.1016/j.ridd.2011.01.031
28. 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(2):795-800. doi: http://dx.doi.org/10.1016/j.ridd.2012.11.002
29. 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(8):706-13.

30. 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(5):490-96. doi: 10.1111/dmcn.13367 [published Online First: 2017/01/10]

31. 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(4):1361-9. doi: http://dx.doi.org/10.1016/j.ridd.2011.01.032

32. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation, reliability, and validity for Brazilian children. *Res Dev Disabil* 2014;35(3):733-40. doi: http://dx.doi.org/10.1016/j.ridd.2013.10.028

33. 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(2):160-5. doi: http://dx.doi.org/10.1111/j.1469-8749.2011.04177.x

34. 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(3):485-97. doi: http://dx.doi.org/10.1016/j.humov.2013.02.007

35. 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(8):637-41.

36. 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(1):40-45. doi: 10.1016/S1013-7025(09)70058-2

37. 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(3):141-49.

38. Burns YR, Ensbey RM, Norrie MA. The neuro-sensory motor developmental assessment part II: predictive and concurrent validity. *Aust J Physiother* 1989;35(3):151-57.

39. Barnett LM, Minto C, Lander N, et al. Interrater reliability assessment using the Test of Gross Motor Development-2. *Journal of Science and Medicine in Sport* 2014;17(6):667-70. doi: http://dx.doi.org/10.1016/j.jsams.2013.09.013

40. 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(2):18-28. doi: 10.5897/JPESM12.003

41. 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 Act Q* 2010;27(2):143-59.

42. Kim Y, Park I, Kang M. Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapt Phys Act Q* 2012;29(4):346-65.

43. 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(4):800-7. doi: http://dx.doi.org/10.1016/j.ridd.2014.01.019

44. Rudd J, Butson ML, Barnett L, et al. A holistic measurement model of movement competency in children. *J Sports Sci* 2016;34(5):477-85.

45. 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. *Adapted physical activity quarterly : APAQ* 2008;25(1):71-82.

46. Valentini NC. Validity and reliability of the TGMD-2 for Brazilian children. *J Mot Behav* 2012;44(4):275-80. doi: http://dx.doi.org/10.1080/00222895.2012.700967

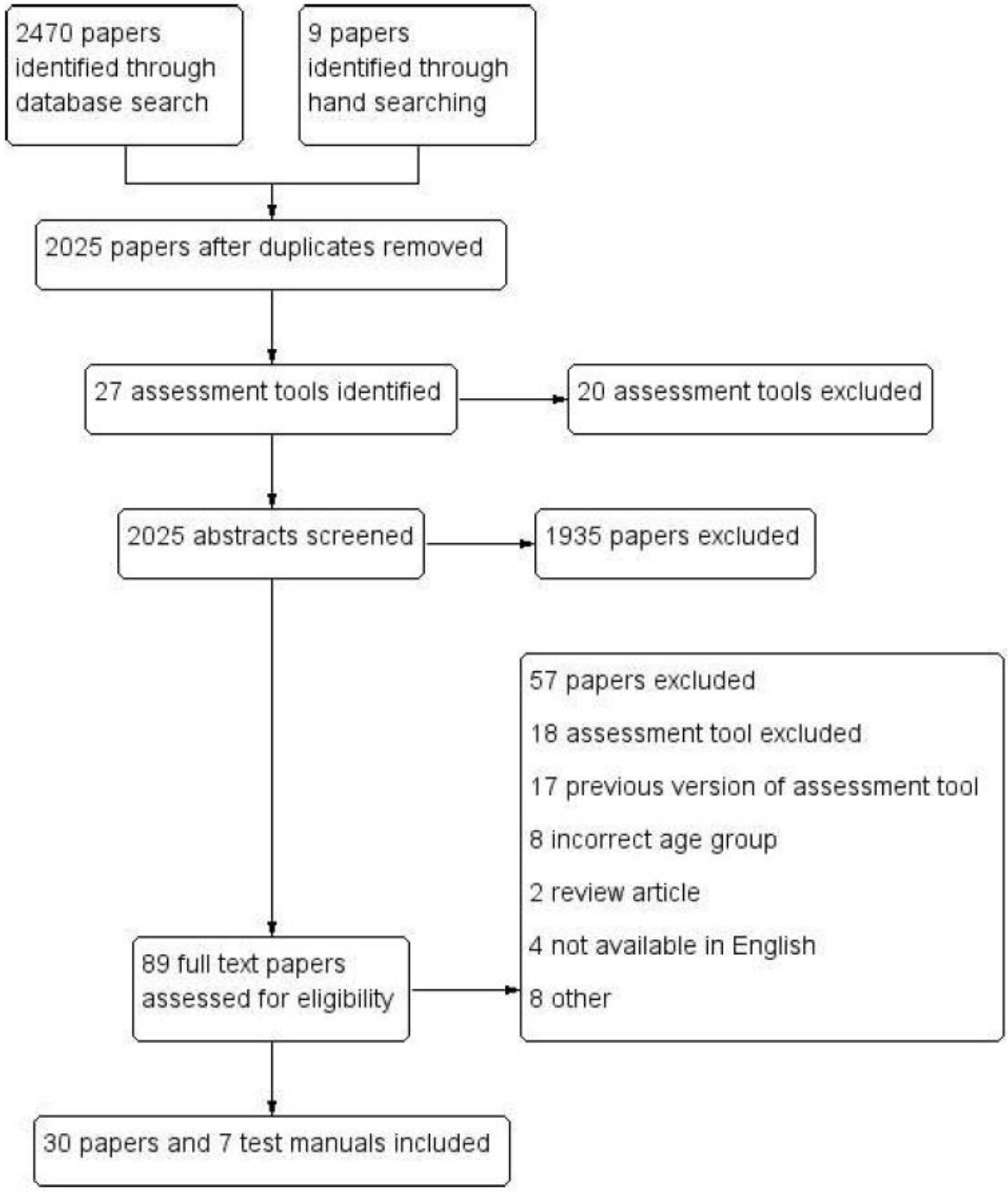
47. Wong KYA, Yin Cheung S. Confirmatory factor analysis of the Test of Gross Motor Development-2. *Measurement in Physical Education & Exercise Science* 2010;14(3):202-09. doi: 10.1080/10913671003726968

48. 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(6):1132-44. doi: http://dx.doi.org/10.1016/j.ridd.2009.03.003

49. 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(4):254-66. doi: 10.1111/j.1469-8749.2008.02025.x [published Online First: 2008/01/15]

50. McDowell I. Measuring health: a guide to rating scales and questionnaires: Oxford university press 2006.

51. 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(7):596-605.


2470 papers identified through database search
9 papers identified through hand searching
2025 papers after duplicates removed
27 assessment tools identified
20 assessment tools excluded
2025 abstracts screened
1935 papers excluded
57 papers excluded
18 assessment tool excluded
17 previous version of assessment tool
8 incorrect age group
2 review article
4 not available in English
8 other
89 full text papers assessed for eligibility
30 papers and 7 test manuals included

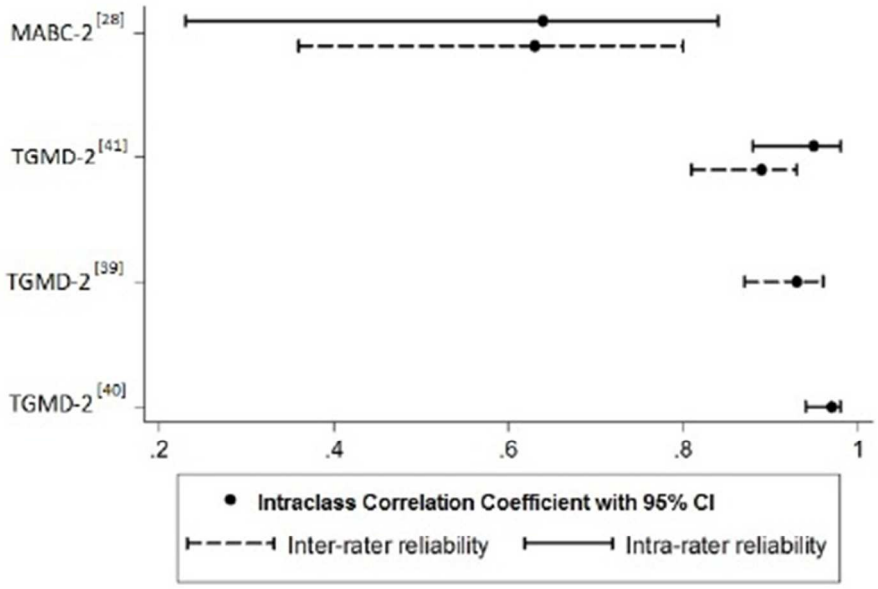


Figure 2. Test re-test reliability of gross motor assessment tools
Figure 2 legend: BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition 13; MABC-2, Movement Assessment Battery for Children 2nd edition 14; PDMS-2, Peabody Developmental Motor Scales 2nd edition 17; TGMD-II, Test of Gross Motor Development 2nd edition 19.

41x28mm (300 x 300 DPI)

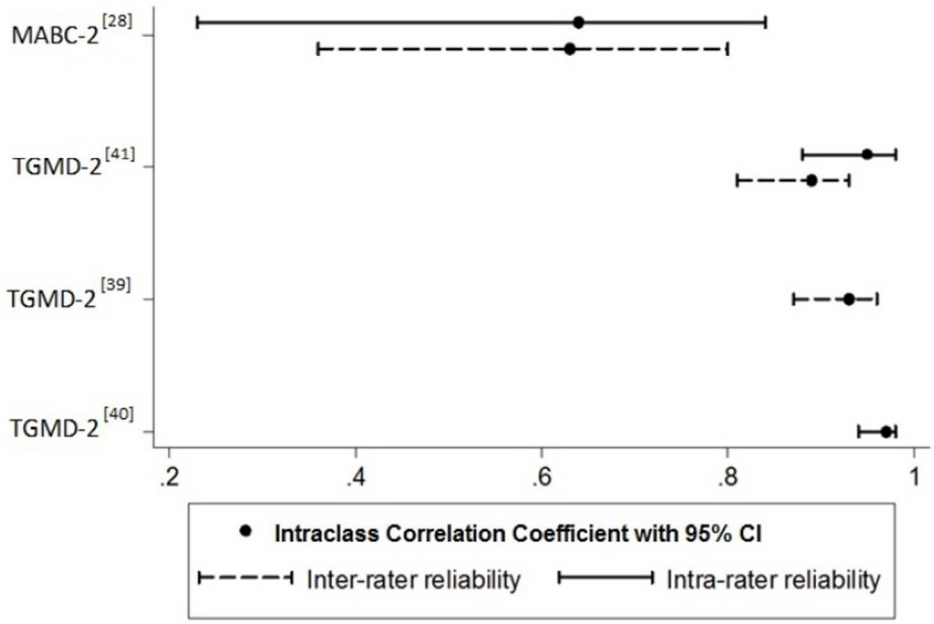


Figure 3. Inter and interrater reliability of gross motor assessment tools
Figure 3 legend: MABC-2, Movement Assessment Battery for Children 2nd edition 14; TGMD-II, Test of Gross Motor Development 2nd edition 19

67x48mm (300 x 300 DPI)

Supplementary table 1: Definition of terms

		Definition	Example/explanation
Validity	Content	The degree to which an assessment tool's content measures the construct that it intends to measure [7]	Concerned with the relevance and comprehensiveness of the items included in the assessment tool
	Construct	Measures the degree to which the scores obtained from the test are an adequate reflection of the construct to be measured [7]	Examples include structural validity, hypothesis testing and cross-cultural validity
	Criterion	Assesses whether or not the test scores reflect a 'gold standard' assessment [7]	As there is no gold standard of assessment for gross motor function in children this is often assessed with correlations of scores obtained from two or three other frequently used tools.
Reliability		Refers to the consistency of a test score regardless of the time between assessments (test-retest) or the person administrating (intra and inter-rater) [27]	Usually measured with intraclass correlation coefficient (ICC), but can be measured using Cohen's kappa coefficient. Percentage agreement and Pearson's correlation coefficient do not incorporate error into the calculations and as such is not a true measure of agreement [27]. Scores > 0.80 are considered excellent, 0.60-0.79 adequate and <0.59 poor [12]
	Internal consistency	The degree of interrelatedness of an assessment tool's items [7]	Usually measured using Cronbach's alpha (α) [7]. scores > 0.70 demonstrates high relationship, 0.5 to 0.69 a moderate relationship, 0.26 to 0.49 a low relationship and < 0.26

			little relationship [27].
	Measurement Error	Refers to the error obtained between measurements that cannot be attributed to the patients true change [7]	May be systematic or random error [7]
Responsiveness		An assessment tool's ability to detect change over time in the construct to be measured [7]	This is central to a tools capacity to be used as an outcome measure.

Supplementary table 2: Excluded Assessment Tools

Reason	Assessments
Manual not available in English	Maastricht's Motor Test (MMT)
	The Motor-Proficiency-Test for children between 4 and 6 years of age (MOT 4-6)
	Zuk Assessment
	Körperkoordinationtest für Kinder (KTK)
Cannot extract meaningful gross motor score	Early Intervention Developmental Profile (EIDP)
	Neurological Developmental Exam
	Preschooler Gross Motor Quality Scale (PGMQ)
	The Malawi Developmental Assessment Tool (MDAT)
	Dutch table tennis motor skills assessment
Screening Tool	Brief Assessment of Motor Function (BAMF)
	The Motor Performance Checklist
	Motor skill checklist (MSC)
Diagnosis specific/requires a diagnosis	Assessment Battery for the Atypical Handicapped Child (VAB)
	Video-based documentation and rating system of the motor behaviour of handicapped children
Only assesses one motor domain (e.g. gait)	Standardized Walking Obstacle Course (SWOC)
	Timed floor to stand test
Manual not published/commercially available	Rapid Neurodevelopmental Assessment (RNDA)
	Tufts Assessment of Motor Performance (TAMP)
	Zurich Neuromotor Assessment (ZNA)

Supplementary table 3: Scoring and administration of assessment tools

Assessment Tool	Scoring	Interpretation of scores	Other
Bayley-III [28]	Motor score - gross (varying items) and fine motor (varying items) subscales. Binary score with reverse/discontinue rules	Raw scores Composite scores Centile ranks Age equivalents Growth scores	Lends itself to multidisciplinary team testing.
BOT-2 [13]	Fine manual (15 items) manual coordination (12 items) body coordination (16 items) strength and agility (10 items) subscales. Scoring differs for subtests	Raw scores Age adjusted standard scores Composite scores Centile ranks Age equivalents Descriptive categories. Complex conversions	Administration Easel includes instructions, diagrams and photos of test procedure
MABC-2[24]	Manual dexterity (3 items), aiming & catching (2 items) and balance (3 items) subscales.	Raw scores component scores centile ranks total test score traffic light system. Simple conversion	Also Available: MABC-2 Checklist (screening tool) and intervention manual
MAND [29]	Fine motor (5 items) Gross motor (5 items)	Raw scores Scaled scores converted to an NDI. Factor scores. Complex conversions	Case studies included in manual for hyperactivity, encephalitis, mild head trauma, CP and muscular dystrophy
NSMDA [30]	Functional grade given for each subscale, which is combined to create an overall score.	Indicates: normal range, minimal dysfunction, mild problems, moderate, severe or profound disability	Sections for comment on strengths, behavioural state during testing, musculoskeletal system and recommendations.

PDMS-2 [31]	GM: Stationary (30 items), locomotion (89 items), object manipulation (24 item). FM: grasping(26 items) , visual-motor integration (72 items)	Raw scores, Age equivalent, centile rank. Standard scores (subtests) Composite quotient. Complex conversions.	Motor activities program (intervention ideas)
TGMD-2 [25]	Locomotion (6 items) and Object Control (6 items). Separate male/female norms for object control subset	Raw scores, standard scores, percentile rank, age equivalent, Gross Motor Quotient. Simple conversion.	Simple to administer

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition [28]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [24]; MAND, McCarron Assessment of Neuromuscular Development [29]; NSMDA, Neurological Sensory Motor Developmental Assessment [30]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [31];; TGMD-II, Test of Gross Motor Development 2nd edition [25]; GM, Gross Motor; FM, Fine Motor; NDI, Neurodevelopmental Index

Supplementary table 4: Content and construct validity of assessment tools

Test	Content	Construct
BAYLEY III	Expert opinion for standard and low verbal version [28, 34]. Literature reviews. Gross motor score correlated with Motor component 0.70 [28]	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language impairment [28]. H_i (gross motor subset) = 0.52-0.97 for children with language impairment and 0.82-0.99 in control group [34]
BOT-2	Focus groups, product survey, pilot, national tryout and standardisation studies, professional reviews[13]	Factor analysis, scores increase with age, discriminates between normal and children with DCD (*N*=50), high-functioning ASD (*N* = 45) and mild-moderate ID (*N* = 66) [13]
MABC-2	Expert Panel, Stakeholder feedback, Literature review [18] Expert panel - clarity (validity content index 71.8-93.9, Kappa 0.76-0.88) and pertinence (98.5-99.3 and kappa 0.83-0.92) $p<0.001$ [40]	Factor analysis, correlation coefficients [36] Subtest correlations 0.65-0.76 $p<0.001$. Discriminates between ASD and control group [18]. Structural equation modelling (for each age group) [39]. Expert panel - adequate face validity [40]. Significant difference between TD, DCD and at risk DCD scores ($\eta 2$ = 0.63) $p< 0.0001$ [40]. UK norms not appropriate to use with Dutch/Flemish children as under/over-estimate risk of motor impairment [15]. In Chinese population: CFA initially rejected. Acceptable fit achieved after 2 items removed [14]. Age band 2 shows good validity in Japanese population [37].
MAND	Based on neuropsychological theory. Several rounds of revision/trials of tasks during development [29]	Factor analysis [29] [42]. Scores increase with age, and discriminate between typically developing children and those with head trauma or neurological dysfunction as well as gender [29] [42]
NSMDA	Literature review. Developed by an experienced paediatric physiotherapist [45]	Factor analysis (up to 2 years of age) [45] [46]. Stability of test results over time (up to 2 years) [45] [46].
PDMS-2	Literature review. Created by experts in the field. Revised with feedback from therapists guided revision. Hierarchical sequence	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with age (r=0.80-0.93) [31]

	of items [31]	
TGMD-2	Expert Panel (3 PE teachers with post-grad qualifications) [25]. 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 [52]	Exploratory and confirmatory factor analysis [50] [52] [25] [22] [53] [51] High and significant correlation of increasing age and increasing scores [47]. Age and disability differentiation [25] [51] Subtest correlation 0.41 [25] Galloping, running and leaping not well correlated with locomotion subscale. Object control significant & highly correlated [49]. ANOVA - significant age effect for object control [51] Moderate correlation between items and subset scores, and between subset scores and total score [51]

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[28] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[24] MAND, McCarron Assessment of Neuromuscular Development;[29] NSMDA, Neurological Sensory Motor Developmental Assessment;[30] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[31] TGMD-II, Test of Gross Motor Development 2nd edition;[25]; H_i, scalability coefficient; CFA, Confirmatory Factor Analysis; TD, Typically Developing; ASD, Autism Spectrum Disorder, ID, Intellectual Disability; WPPSI, *Wechsler Preschool and Primary Scale of Intelligence*; WISC-R, Wechsler Preschool and Primary Scale of Intelligence-R; NDI, Neurodevelopmental Index; ANOVA, Analysis of Variance

Supplementary table 5: Criterion and predictive validity of assessment tools

Test	Concurrent/criterion	Predictive
BAYLEY III	Given but mean age <22 months. Not relevant to study population. [28]	Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.32-.037 specificity 0.97 <2SD sensitivity 0.18-0.21 specificity 1.00.

		CP at 4 years: Bayley III at 2 years <1SD sensitivity 0.83 specificity 0.94. <2SD sensitivity 0.67 specificity 1.0 [4]
BOT-2	MABC-2 p = 0.92 PDMS-2 p = 0.88 (N = 38) [17]. PDMS-2 Total motor composite r = 0.77 [13].	-
MABC-2	PDMS-2 ρ = 0.631 – 0.84 [17] [14]. TGMD-2 ρ = 0.45 [5]. TGMD-2 standard scores (r = 0.3, p < 0.02) [40]. BOT-2 ρ =0.90 - 0.92 [17].	Classification groups (DCD, at risk and TD) remained same over time (6 months) $\chi 2$ = 0.67 p = 0.72 [40]. Predictive of motor impairment over 6-12 months (N=41) ICC 0.88 p < 0.007 [40]. Scores at 4 years predictive of motor impairment at 8 years in children born <30 weeks gestation (PPV 79, sensitivity 79%, specificity 93%)) [38]
MAND	Gross motor subscore: Low-moderate correlation with manual dexterity (-0.46 to 0.35), reaction time (-0.31 to -0.58), intelligence measures (WISC-R, Metropolitan Achievement Test) (0.30-0.39) and visual motor test (-0.33 to 0.39) [29]	-
NSMDA	NSMDA at 2 years (N = 148) predictive of medical diagnosis $\chi 2$ = 0.08 p = NS [46]	Motor outcome at 11-13 yrs. NSMDA at 2years - 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% [43]. If classified 'severe' at 24 months - approximately 50% chance walking at 4 years (moderate = 80%, mild = 93% minimal = 100%) [44]
PDMS-2	MABC-2 ρ = 0.63- 0.84, [14, 17] MABC-2 gross motor composite ρ = 0.743 [14] BOT-2 ρ = 0.88 [17]. Mullen Scales of Early Learning GMQ = 0.86 FMQ = 0.80 [31]	-

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 [49] (N=41) Basic Motor Generalizations subtest of the CSSA r = 0.63. Locomotor 0.63 object control 0.41 [25]	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[28] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[24] MAND, McCarron Assessment of Neuromuscular Development;[29] NSMDA, Neurological Sensory Motor Developmental Assessment;[30] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[31] TGMD-II, Test of Gross Motor Development 2nd edition;[25] NS, Not Specified; SD, Standard Deviation; CP, Cerebral Palsy; TD, Typically Developing; ICC, Intraclass Correlation Coefficient; χ2, Chi Squared; NDI, Neurodevelopmental Index; CSSA, Comprehensive Scales of Student Abilities

Supplementary table 6: Reliability of assessment tools

Test	Internal Consistency	Test-Retest	Intra-rater	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) [28]	Gross Motor subtest (N=47) r=0.79 Motor component r=0.80 [28]	-	-	SEM Gross motor subtest 0.85-1.08. of Motor component = 3.00-4.74 (24-	-

					42 months) [28]	
BOT-2	(*N* = 100) α = 0.92 [35] (*N* = 141) α = 0.86 [17] 4-7 yrs (*N*= 620) α = 0.95 8-11 yrs (*N*= 450) α = 0.95 [13]	(*N* = 100) ICC = 0.99 [35] (*N* = 141) ICC = 0.97 [17] 4-7 yrs (*N* = 43) *r* = 0.81 (8-12 yrs (*N*= 44) *r* = 0.80 [13]	-	Total motor composite 4-21 yrs (*N* = 47) *r* = 0.98 [13]	4.18 (sensitivity 55.10% specificity 72.55%) [35] 7.43 (sensitivity 42.49% specificity 65.72%) [17]	6.53 (sensitivity 48.98% specificity 76.47%) [35] 6.55 (sensitivity 49.99% specificity 58.78%) [17]
MABC-2 (AB 1)	(*N* = 60) M.D α = 0.51, A&C α = 0.70, Bal α = 0.66 [36] (*N* = 1823) α = 0.502 [14] (*N*=50) α = 0.81-0.87 [23]	(*N*=60) ICC = 0.85 [36] Item ICC's 0.830-0.985 [14] ICC test-retest = 0.83 [23] Inter-rater test-retest ICC = 0.79 [23]	(*N*=28) κ = 0.71 [23]	Item ICC's range 0.892-0.998 [14] (*N*=22) κ = 0.60 [23]	(*N*=28) Intrarater MDC = 3.43 (*N*=22) Inter-tester MDC = 3.81 [23]	-
MABC-2 (AB 2)	Translated version (Japanese) (N=132) α = 0.602 [37]	-	ICC = 0.64 [18]	ICC 0.63 [18]	Intra-rater SDC TTS: +/- 11.7 TSS +/- 3.3. Inter-rater SDC TTS +/-16.0 TSS +/- 3.8 [18]	-
MABC-2	Subscales α = 0.78 (M.D = 0.77, BS = 0.52, Bal = 0.77) [40] α = 0.88 [41] (*N* = 141) α = 0.88 [17]	*N*=60 (all 3 age bands) *r*=0.80 [24] *r*=0.74 p<0.0001 (standard score). ICC standard score = 0.85 [40] ICC 0.96 [41] *N* = 141 ICC =0.96 [17]	ICC 0.88 [40]	ICC 0.96-0.99 [40]	SEM 1.34 (95%CI) = 3 [24] 1.83 (95%CI) [41] 1.83 (sensitivity 69.69% specificity 52.10%) [17]	1.39 (sensitivity 72.47% specificity 46.18%) [17, 41]

MAND	-	-	-	-	-	-
NSMDA	Cross correlation matrix Item scoring (12+24months) 0.73 $p<0.001$, Functional grade (12+24months) 0.87 $p<0.001$ [45]	-	-	-	-	-
PDMS-2	(N=141) α=0.89 [17] 24-35m α=0.97, 36-47m α=0.95, 48-59m α=0.97, 60-71m α= 0.98. For subgroups† α=0.99 [31]	N=141 ICC= 0.97 [17]	unable to extract data for ≥24months [31]	unable to extract data for ≥24months [31]	7.76 (sensitivity 60.65% specificity 74.13%) [17] SEM 24-59 months = 3, 60-71m = 2 [31]	8.39 (sensitivity 61.65% specificity 71.34%) [17]
TGMD-2	(N=1438) α=0.80 [47] N=75 Locomotor subset α=0.71 object control α=0.72 [22] N=120 α = 0.72 [49] N= 99 α = 0.90 [51] 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 [25]	N=63 ICC=0.81 95% CI [47] N=23 ICC=0.92 total 95% CI [22] N=99 r=0.98 [51] Locomotor test r = 0.90 p <0.0001 object control test r = 0.91 p <0.001 [52] N = 75 r=0.96 overall (3-5 yrs r = 0.91), 6-8 years r = 0.95), (9-10 years r = 0.94) [25]	N=32 ICC=0.97 95% CI [47] N=25 ICC=0.95 95% CI [22] ICC = 0.78 [48] ICC=0.92-0.99 [52]	Obj ICC=0.93 [19] (N=50) ICC=0.89 [22] ICC=0.75 [48] N=8 r= 1.00 [51] L.S ICC=0.88 Obj ICC=0.89 [52] N = 30 r=0.98 [25]	-	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[28] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[24] MAND, McCarron Assessment of Neuromuscular Development;[29] NSMDA, Neurological Sensory Motor Developmental Assessment;[30] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[31] TGMD-II, Test of Gross Motor Development 2nd edition;[25] GM, Gross Motor Subset; MC, Motor Component; K, Kappa Coefficient; M.D, Manual Dexterity; BS, Ball Skills; BAL, Balance; A&C, Aiming and catching; SDC, Smallest Detectable Change; TTS, Total Test Score; TSS, Total Standard Score; †, gender, ethnicity, speech/language or physical disorder; Obj, Object Control Subset; L.S, Locomotion Subset

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1 Title page
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	-
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5-6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5-6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7

PRISMA 2009 Checklist

Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	-

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	-
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	-
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Figure 1 + page 7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1 – page 8 + Suppl table 3
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	10 + Table 3 – page 13-14
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	10-11, 16 + Table 2 – page 12 + Figures 2 – 3 + Suppl tables 4-6
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	-
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	-
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	-

PRISMA 2009 Checklist

DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	16-18
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	19
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-021734.R1
Article Type:	Research
Date Submitted by the Author:	19-Jun-2018
Complete List of Authors:	Griffiths, Alison; Monash University, Physiotherapy; The Royal Children's Hospital, Physiotherapy Toovey, Rachel; Murdoch Childrens Research Institute, Developmental Disability and Rehabilitation Research; University of Melbourne, Physiotherapy Morgan, Prue; Monash University, Physiotherapy Spittle, Alicia; University of Melbourne, Physiotherapy; Murdoch Childrens Research Institute, Victorian Infant Brain Studies
Primary Subject Heading:	Paediatrics
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	PAEDIATRICS, REHABILITATION MEDICINE, validity, gross motor assessment, reliability

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]

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

Corresponding Author*:*

Name: A/Prof Alicia J. Spittle

Address: Department of Physiotherapy, The University of Melbourne, Level 7 Alan Gilbert Building, 161 Barry Street, Parkville, Vic 3010, Australia

Email: aspittle@unimelb.edu.au

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 Program.

Financial disclosure: The authors have declared that they have no financial relationships relevant to this article.

Conflict of interest: The authors have no conflict of interest.

Keywords: paediatrics, reliability, validity, rehabilitation medicine, gross motor assessment

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 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 (COSMIN) 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 re-test reliability is excellent in the BOT-2 (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 interrater (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 COSMIN 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 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 to 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]. Further, 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].

Health 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 [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 Movement Assessment Battery for Children (MABC) in age range, age bands, materials and tasks, that the MABC-2 requires independent reliability and validity assessment. Over the past eight 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 2-12 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) (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 required regarding availability of manuals authors were contacted.

Assessment tools were included if they were 1. Discriminative, predictive or evaluative of gross motor skills, 2. Assessed ≥ two gross motor (e.g. balance, jumping etc.) items, 3. Able to extract a meaningful gross motor sub-score, 4. Applicable to children 2-12 years of age, 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 (e.g. 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 agreement by 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] (Supplementary Table 5). Cross-cultural validity, structural validity and hypothesis testing are all considered to be a component of construct validity[7]. Whilst predictive validity is considered to be a component of content validity, it is reported on 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 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.

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), 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). The corresponding manuals were then added to the final yield resulting in thirty papers and seven manuals. Twenty assessment tools were excluded (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 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.

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 [12]	Gross motor, fine motor, cognitive, communication, social/emotional, adaptive	Static postures, dynamic movement, balance	1 mth – 3 yrs	Developmental delay: <25th centile or below 2SD. *	Discriminative	Predictive, Evaluative, Research tool	Norm	1700 children from the USA (2000)
BOT-2 [13]	Gross motor, fine motor	Coordination, balance, running speed and agility, strength	4 – 21 yrs	*	Discriminative Evaluative	Research tool	Norm	1520 children from the USA (2005)
MABC-2 [14]	Gross motor, fine motor, balance	Aiming and catching, static and dynamic balance	3 – 16 yrs	Traffic light system: Green = normal, amber = 'at risk' and red = definite motor impairment (<15%). *	Discriminative Evaluative	Intervention planning, Research tool	Norm	1172 children from United Kingdom (2006)
MAND [15]	Gross and fine motor	Coordination, jumping, static and dynamic balance	3 yrs – 25 yrs	NDI 70-85 = mild 55-69 = moderate <55 = severe disability *	Evaluative	Research tool	Norm	2000 3-35 yrs from the USA (1970's)
NSMDA [16]	Gross Motor, Fine Motor, Neurological, Postural Development, Infant Patterns of Movement, Sensory Motor. †	Sitting, kneeling, walking, balance, running, hopping, jumping, catching, motor planning	1 mth – 6 yrs	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	N/A
PDMS-2 [17]	Gross motor, fine motor	Stationary (standing balance, sit-ups, push-ups), locomotion (walking, running, jumping, hopping, etc.), object manipulation (kick, throw, hit, catch)	Birth – 5 yrs	*	Discriminative Evaluative	Predictive, Research tool	Norm	2003 USA and Canada (1997-8)

TGMD-2 [18]	Gross Motor	Locomotion (run, gallop, hop, leap, jump, slide) and Object control (batting, dribbling, catch, kick, throw, roll)	3 – 10 yrs	*	Discriminative Evaluative	Outcome measure, research tool, intervention planning	Norm	1208 USA children (1997-1998)

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18] NDI, Neurodevelopmental Index; SD, Standard Deviation; mth, month; yrs, years *, Advisable to use clinical reasoning; †, requires some manual handling; USA, United States of America

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 Supplementary Table 7. The shortest administration time is 15-20 minutes for the TGMD-2 and the MAND; whilst most manuals report 20-60 minutes 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 post assessment (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 [13] 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 [19 20 18 21 22 23]. 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 and Dutch speaking populations [24 25]. 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, whilst 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.

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 [12]	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 [13]	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 [14]	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 [15]	15-20	Therapist administers items in standardised order.	Professionals e.g. education, neurology, OT, PT, psychology etc.	Formal training not required.	Manual and test kit: £1366 includes equipment
NSMDA [16]	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 [17]	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 [18]	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 3rd edition [12]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; MAND, McCarron Assessment of Neuromuscular Development [15]; NSMDA, Neurological Sensory Motor Developmental Assessment [16]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17]; TGMD-II, Test of Gross Motor Development 2nd edition [18]; GM, Gross motor; OT, Occupational Therapy; PT, Physiotherapy; PE, Physical Education

Table 3. Methodological quality of included articles

Test	First author, Year	Country	Population (Age, Diagnosis)	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsive - ness
BAYLEY III	Bayley [12]	USA	1-42 mths	Fair	Fair	Good	Excellent	Good	Good	-	Good	-
	Spittle, et al. [4]	Australia	2,4 yrs, Ex prem	-	-	-	-	-	-	-	Good	-
	Visser, et al. [26]	Netherlands	2.2-10.8 yrs, GDD, L.I.	-	-	-	Excellent	Poor	-	-	-	-
BOT-2	Wuang and Su [27]	Taiwan	4-12 yrs ID	Excellent	Excellent	Excellent	-	-	-	-	-	Fair
	Wuang, et al. [28]	Taiwan	3-6 yrs ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Bruininks and Bruininks [13]	USA	4-21 yrs	Good	Fair (interrater) Fair (test-retest)	Good	Excellent	Good	-	-	Good	-
MABC-2 (AB 1)	Ellinoudis, et al. [29]	Greece	3-5.5 yrs	Excellent	Good	-	-	-	-	-	-	-
	Hua, et al. [24]	China	3-6 yrs	Excellent	Good	-	Excellent	Excellent	-	Poor	Excellent	-
	Logan, et al. [5]	USA	3-6 yrs	-	-	-	-	-	Fair	-	Fair	-
	Smits-Engelsman, et al. [30]	Belgium	3-4 yrs	Poor	Poor	Poor	-	-	-	-	-	-
MABC-2 (AB 2)	Holm, et al. [31]	Norway	7-9 yrs	-	Fair (interrater) Poor (intrarater)	Poor	-	-	-	-	-	-
	Kita, et al. [32]	Japan	7-10 yrs	Excellent	-	-	-	-	-	Poor	-	-
MABC-2	Griffiths, et al. [33]	Australia	4-8 yrs	-	-	-	-	-	-	-	Good	-
	Henderson, et al. [14]	UK	3-16 yrs	-	Fair	Good	Excellent	-	-	-	-	-
	Niemeijer, et al. [25]	Netherlands + Belgium	-	-	-	-	-	-	-	Poor	-	-
	Schulz, et al. [34]	U.K	3-16 yrs	-	-	-	Excellent	Good	-	-	-	-
	Valentini, et al. [35]	Brazil	3-13 yrs	Fair	Fair	-	Fair	Poor	-	Poor	Poor	-

	Wuang, et al. [28]	Taiwan	3-6 yrs, ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Wuang, et al. [36]	Taiwan	6-12 yrs DCD	Poor	Fair	Good	-	-	-	-	-	Fair
MAND	Hands, et al. [37]	Australia	10-17 yrs	-	-	-	-	Excellent	-	-	-	-
	McCarron [15]	USA	7yrs	-	-	-	Fair	Poor	-	-	Poor	-
NSMDA	Danks, et al. [38]	Australia	2 + 4 yrs ELBW	-	-	-	-	-	-	-	Fair	-
	MacDonald and Burns [39]	Australia	2 + 4 yrs CP	-	-	-	-	Fair	-	-	Poor	-
	Burns, et al. [40]	Australia	1-24 mths VLBW	Poor	-	-	Poor	-	-	-	-	-
	Burns, et al. [41]	Australia	1-mnths VLBW	-	-	-	-	Poor	-	-	Fair	-
PDMS-2	Hua, et al. [24]	China	3-6 yrs.	Excellent	Good	-	Excellent	Excellent	-	Poor	Excellent	-
	Wuang, et al. [28]	Taiwan	3-6 yrs ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Folio and Fewell [17]	USA	0-71 mths	Good	-	Poor	Excellent	Good	Good	-	Poor	-
TGMD-2	Barnett, et al. [42]	Australia	4-8 yrs	-	Fair	-	-	-	-	-	-	-
	Farrokhi, et al. [43]	Iran	3-11 yrs	Fair	Fair	-	Fair	Fair	-	-	-	-
	Houwen, et al. [21]	Netherlands	6-12 yrs VI	Fair	Fair	-	-	Fair	-	-	-	-
	Kim, et al. [44]	Korea	8-12 yrs ID	-	Poor	-	-	-	-	-	-	-
	Kim, et al. [45]	Korea	5-6 yrs	Poor	Fair	-	-	Poor	-	-	Poor	-
	Logan, et al. [5]	USA	3-6 yrs	-	-	-	-	-	Fair	-	Fair	-
	Rudd, et al. [19]	Australia	6-12 yrs	-	-	-	-	Good	-	-	-	-
	Simons, et al. [23]	Belgium	7-10 yrs ID	Good	Good (interrater) Poor (test-retest)	-	Excellent	Good	Good	-	-	-
	Valentini [20]	Brazil	3-10 yrs	Poor	Fair (test-retest) Good (intra,	-	Excellent	Good	-	Fair	Good	-

				interrater)							
Wong and Yin Cheung [22]	China	3-10 yrs	-	-	-	-	Fair	-	-	-	-
Ulrich [18]	USA	3-10 yrs	Good	Fair (test-retest) Poor (interrater)	Fair	Poor	Good	-	-	Fair	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17]; TGMD-II, Test of Gross Motor Development 2nd edition;[18] Mths, Months; yrs, years; DCD, Developmental Coordination Disorder; VI, Vision Impairment; ID, Intellectual Disability; GDD, global developmental delay; L.I, Language Impairment; ELBW, Extremely Low Birth Weight; VLBW, Very Low Birth Weight; CP, Cerebral Palsy; prem, premature; USA, United States of America

Reliability

Internal consistency of assessments are summarised in Table 6. The BOT-2's high internal consistency is well supported, including for children with an intellectual disability [28 46]. 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.

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). Intra-rater 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) [31 42]. Inter-rater reliability is also supported in the BOT-2 with Pearson 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 [47 48]. Only studies reporting ICC's are visually represented in Figures 2 (test-retest) and 3 (inter and intra-rater). The TGMD-2 test-retest reliability results from Houwen, et al. [21] were believed to contain an error as the reported ICC was outside of the reported confidence intervals (ICC 0.92, 0.82-0.91). 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 standard error of measurement (SEM) [28]. Sensitivity and specificity for detecting change was shown to be satisfactory in the MABC-2, PDMS-2 and MABC-2 [28] (Table 6). There have been no studies to date on the responsiveness of the TGMD-2, NSMDA or MAND.

Table 4: Content and construct validity of assessment tools

Test	Content	Construct
BAYLEY III	Expert opinion for standard and low verbal version [12 26]. Literature reviews. Gross motor score correlated with Motor component 0.70 [12]	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language impairment [12]. H_i (gross motor subset) = 0.52-0.97 for children with language impairment and 0.82-0.99 in control group [26]
BOT-2	Focus groups, product survey, pilot, national tryout and standardisation studies, professional reviews[13]	Factor analysis, scores increase with age, discriminates between normal and children with DCD (*N*=50), high-functioning ASD (*N* = 45) and mild-moderate ID (*N* = 66) [13]
MABC-2	Expert Panel, Stakeholder feedback, Literature review [31] Expert panel - clarity (validity content index 71.8-93.9, Kappa 0.76-0.88) and pertinence (98.5-99.3 and kappa 0.83-0.92) $p<0.001$ [35]	Factor analysis, correlation coefficients [29] Subtest correlations 0.65-0.76 $p<0.001$. Discriminates between ASD and control group [31]. Structural equation modelling (for each age group) [34]. Expert panel - adequate face validity [35]. Significant difference between TD, DCD and at risk DCD scores ($\eta 2 = 0.63$) $p< 0.0001$ [35]. UK norms not appropriate to use with Dutch/Flemish children as under/over-estimate risk of motor impairment [25]. In Chinese population: CFA initially rejected. Acceptable fit achieved after 2 items removed [24]. Age band 2 shows good validity in Japanese population [32].
MAND	Based on neuropsychological theory. Several rounds of revision/trials of tasks during development [15]	Factor analysis [15 37]. Scores increase with age, and discriminate between typically developing children and those with head trauma or neurological dysfunction as well as gender [15 37]
NSMDA	Literature review. Developed by an experienced paediatric physiotherapist [40]	Factor analysis (up to 2 years of age) [40 41]. Stability of test results over time (up to 2 years) [40 41].
PDMS-2	Literature review. Created by experts in the field. Revised with feedback from therapists guided revision. Hierarchical sequence of items [17]	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with age (r=0.80-0.93) [17]
TGMD-2	Expert Panel (3 PE teachers with post-grad qualifications) [18]. 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 [20]	Exploratory and confirmatory factor analysis [19 20 18 21 22 23] High and significant correlation of increasing age and increasing scores [43]. Age and disability differentiation [18 23] Subtest correlation 0.41 [18] Galloping, running and leaping not well correlated with locomotion subscale. Object control significant & highly correlated [45]. ANOVA - significant age effect for object control [23] Moderate correlation between items and subset scores, and between subset scores and total score [23]

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18]; H_i, scalability coefficient; CFA, Confirmatory Factor Analysis; TD, Typically Developing; ASD, Autism Spectrum Disorder, ID, Intellectual Disability; WPPSI, *Wechsler Preschool and Primary Scale of Intelligence*; WISC-R, Wechsler Preschool and Primary Scale of Intelligence-R; NDI, Neurodevelopmental Index; ANOVA, Analysis of Variance

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. [12]	Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.32-.037 specificity 0.97 <2SD sensitivity 0.18-0.21 specificity 1.00. CP at 4 years: Bayley III at 2 years <1SD sensitivity 0.83 specificity 0.94. <2SD sensitivity 0.67 specificity 1.0 [4]
BOT-2	MABC-2 p = 0.92 PDMS-2 p = 0.88 (N = 38) [28]. PDMS-2 Total motor composite r = 0.77 [13].	-
MABC-2	PDMS-2 ρ = 0.631 – 0.84 [28 24]. TGMD-2 ρ = 0.45 [5]. TGMD-2 standard scores (r = 0.3, p < 0.02) [35]. BOT-2 ρ =0.90 - 0.92 [28].	Classification groups (DCD, at risk and TD) remained same over time (6 months) $\chi 2$ = 0.67 p = 0.72 [35]. Predictive of motor impairment over 6-12 months (N=41) ICC 0.88 p < 0.007 [35]. Scores at 4 years predictive of motor impairment at 8 years in children born <30 weeks gestation (PPV 79, sensitivity 79%, specificity 93%)) [33]
MAND	Gross motor subscore: Low-moderate correlation with manual dexterity (-0.46 to 0.35), reaction time (-0.31 to -0.58), intelligence measures (WISC-R, Metropolitan Achievement Test) (0.30-0.39) and visual motor test (-0.33 to 0.39) [15]	-
NSMDA	NSMDA at 2 years (N = 148) predictive of medical diagnosis $\chi 2$ = 0.08 p = NS [41]	Motor outcome at 11-13 yrs. NSMDA at 2years - 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% [38]. If classified 'severe' at 24 months - approximately 50% chance walking at 4 years (moderate = 80%, mild = 93% minimal = 100%) [39]
PDMS-2	MABC-2 ρ = 0.63- 0.84, [24 28] MABC-2 gross motor composite ρ = 0.743 [24] BOT-2 ρ = 0.88 [28]. Mullen Scales of Early Learning GMQ = 0.86 FMQ = 0.80 [17]	-
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 [45] (N=41) Basic Motor Generalizations subtest of the CSSA r = 0.63. Locomotor 0.63 object control 0.41 [18]	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18] NS, Not Specified; SD, Standard Deviation; CP, Cerebral Palsy; TD, Typically Developing; ICC, Intraclass Correlation Coefficient; χ2, Chi Squared; NDI, Neurodevelopmental Index; CSSA, Comprehensive Scales of Student Abilities

Table 6: Reliability of assessment tools

Test	Internal Consistency	Test-Retest	Intra-rater	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) [12]	Gross Motor subtest (N=47) *r*=0.79 Motor component *r*=0.80 [12]	-	-	SEM Gross motor subtest 0.85-1.08. of Motor component = 3.00-4.74 (24-42 months) [12]	-
BOT-2	(*N* = 100) α = 0.92 [27] (*N* = 141) α = 0.86 [28] 4-7 yrs (*N*= 620) α = 0.95 8-11 yrs (*N*= 450) α = 0.95 [13]	(*N* = 100) ICC = 0.99 [27] (*N* = 141) ICC = 0.97 [28] 4-7 yrs (*N* = 43) *r* = 0.81 (8-12 yrs (*N*= 44) *r* = 0.80 [13]	-	Total motor composite 4-21 yrs (*N* = 47) *r* = 0.98 [13]	4.18 (sensitivity 55.10% specificity 72.55%) [27] 7.43 (sensitivity 42.49% specificity 65.72%) [28]	6.53 (sensitivity 48.98% specificity 76.47%) [27] 6.55 (sensitivity 49.99% specificity 58.78%) [28]
MABC-2 (AB 1)	(*N* = 60) M.D α = 0.51, A&C α = 0.70, Bal α = 0.66 [29] (*N* = 1823) α = 0.502 [24] (*N*=50) α = 0.81-0.87 [30]	(*N*=60) ICC = 0.85 [29] Item ICC's 0.830-0.985 [24] ICC test-retest = 0.83 [30] Inter-rater test-retest ICC = 0.79 [30]	(*N*=28) κ = 0.71 [30]	Item ICC's range 0.892-0.998 [24] (*N*=22) κ = 0.60 [30]	(*N*=28) Intrarater MDC = 3.43 (*N*=22) Inter-tester MDC = 3.81 [30]	-
MABC-2 (AB 2)	Translated version (Japanese) (N=132) α = 0.602 [32]	-	ICC = 0.64 [31]	ICC 0.63 [31]	Intra-rater SDC TTS: +/- 11.7 TSS +/- 3.3. Inter-rater SDC TTS +/-16.0 TSS +/- 3.8 [31]	-
MABC-2	Subscales α = 0.78 (M.D = 0.77, BS = 0.52, Bal = 0.77) [35] α = 0.88 [36] (*N* = 141) α = 0.88 [28]	*N*=60 (all 3 age bands) *r*=0.80 [14] *r*=0.74 p<0.0001 (standard score). ICC standard score = 0.85 [35] ICC 0.96 [36] *N* = 141 ICC =0.96 [28]	ICC 0.88 [35]	ICC 0.96-0.99 [35]	SEM 1.34 (95%CI) = 3 [14] 1.83 (95%CI) [36] 1.83 (sensitivity 69.69% specificity 52.10%) [28]	1.39 (sensitivity 72.47% specificity 46.18%) [28 36]
MAND	-	-	-	-	-	-
NSMDA	Cross correlation matrix Item scoring (12+24months) 0.73 *p*<0.001, Functional grade (12+24months) 0.87 *p*<0.001 [40]	-	-	-	-	-
PDMS-2	(*N*=141) α=0.89 [28] 24-35m α=0.97, 36-47m α=0.95, 48-59m α=0.97, 60-71m α=	*N*=141 ICC= 0.97 [28]	unable to extract data for ≥24months	unable to extract data for ≥24months [17]	7.76 (sensitivity 60.65% specificity 74.13%) [28] SEM 24-	8.39 (sensitivity 61.65% specificity 71.34%) [28]

	0.98. For subgroups† α=0.99 [17]		[17]		59 months = 3, 60-71m = 2 [17]	
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 [18]	*N*=63 ICC=0.81 95% CI [43] *N*=23 ICC=0.92 total 95% CI [21] *N*=99 *r*=0.98 [23] Locomotor test *r* = 0.90 *p* <0.0001 object control test *r* = 0.91 *p* <0.001 [20] *N* = 75 *r*=0.96 overall (3-5 yrs *r* = 0.91), 6-8 years *r* = 0.95), (9-10 years *r* = 0.94) [18]	*N*=32 ICC=0.97 95% CI [43] *N*=25 ICC=0.95 95% CI [21] ICC = 0.78 [44] ICC=0.92-0.99 [20]	Obj ICC=0.93 [42] (*N*=50) ICC=0.89 [21] ICC=0.75 [44] *N*=8 *r*= 1.00 [23] L.S ICC=0.88 Obj ICC=0.89 [20] *N* = 30 *r*=0.98 [18]	-	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18] GM, Gross Motor Subset; MC, Motor Component; *K*, Kappa Coefficient; M.D, Manual Dexterity; BS, Ball Skills; BAL, Balance; A&C, Aiming and catching; SDC, Smallest Detectable Change; TTS, Total Test Score; TSS, Total Standard Score; †, gender, ethnicity, speech/language or physical disorder; Obj, Object Control Subset; L.S, Locomotion Subset

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 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. [49] 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 ten additional publications to support the content, construct and criterion validity of the MABC-2 and have demonstrates an overall higher methodological quality of the papers assessing the MABC-2 when compared with the TGMD-2. Papers that had were given 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.

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 two 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 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 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 or intra-rater 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 or intra-rater 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, 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 utilised 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 [18]. 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 (ICF) domains of participation, personal factors and environment [2]. Clinicians should utilise appropriate assessments or questionnaires to ensure that these domains of health are also addressed in line with World Health Organisation 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 inter-scorer difference (or interrater reliability) [18]. 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.

All of the included assessment tools were found to have merits and limitations in their clinical utility the body of evidence to support their use. 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. Whilst 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. 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 (MOT 4-6), the Körperkoordinationtest für Kinder (KTK) and the Maastrichtse Motoriek Test (MMT)[49]. 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 tools. Younger children are more likely to be distracted and may not understand test items as well, which may also increase assessment times[30].

When a new edition of an assessment tools 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 [14], 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 [50], 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 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 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.

Author Contributions

All individuals listed as authors meet the appropriate authorship criteria and have approved the acknowledgement of their contributions. AG was responsible for the drafting of the paper and liaising with the co-authors on findings and conclusions. RT contributed to the paper through interpretation of data, completing methodological assessments and revising manuscript content throughout its development. A/Profs 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.

Data Sharing Statement

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

Figures

Figure 1. PRISMA flow diagram detailing study selection

Figure 2. Test re-test reliability of gross motor assessment tools

Figure 2 legend: BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17]; TGMD-II, Test of Gross Motor Development 2nd edition [18].

Figure 3. Inter and interrater reliability of gross motor assessment tools

Figure 3 legend: MABC-2, Movement Assessment Battery for Children 2nd edition [14]; TGMD-II, Test of Gross Motor Development 2nd edition [18]

1

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(1):65-75. doi: http://dx.doi.org/10.1016/j.humov.2005.10.011
2. World Health Organization. International Classification of Functioning, Disability and Health: ICF: World Health Organization 2001.
3. Magalhaes LC, Cardoso AA, Missiuna C. Activities and participation in children with developmental coordination disorder: a systematic review. *Res Dev Disabil* 2011;32(4):1309-16. doi: 10.1016/j.ridd.2011.01.029 [published Online First: 2011/02/19]
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? *Developmental Medicine And Child Neurology* 2013;55(5):448-52. doi: 10.1111/dmcn.12049
5. Logan SW, Robinson LE, Getchell N. The Comparison of Performances of Preschool Children on Two Motor Assessments. *Perceptual and Motor Skills* 2011;113(3):715-23.
6. Venetsanou F, Kambas A. Environmental factors affecting preschoolers' motor development. *Early Childhood Education Journal* 2010;37(4):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(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006 [published Online First: 2010/05/25]
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. *PEDIATRIC PHYSICAL THERAPY* 2010;22(2):170.
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(1):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(4):651-7. doi: 10.1007/s11136-011-9960-1 [published Online First: 2011/07/07]
11. Law M. Outcome measures rating form. Ontario, Canada: CanChild Centre for Disability Research, 2004.
12. Bayley N. Bayley Scales of Infant Development and Toddler Development: Technical Manual: The PsychCorp 2006.
13. Bruininks R, Bruininks B. Bruininks-Oseretsky Test of Motor Proficiency–2nd Edition (BOT-2): Manual. Circle Pines: MN: AGS Publishing 2005.
14. Henderson SE, Sugden DA, Barnett AL. Movement assessment battery for children-2: Movement ABC-2: Examiner's manual: Pearson 2007.
15. McCarron LT. MAND: McCarron assessment of neuromuscular development, fine and gross motor abilities: McCarron-Dial Systems, Incorporated 1997.

16. Burns YR. N.S.M.D.A Physiotherapy Assessment for Infants and Young Children Second Edition. Brisbane, Queensland: CopyRight Publishing Company 2014.
17. Folio M, Fewell R. Peabody Developmental Motor Scales. Examiner's Manual . 2nd Edition. Austin, Texas.: Pro-Ed. 2000.
18. Ulrich DA. Test of gross motor development-2. *Austin: Prod-Ed* 2000
19. Rudd J, Butson ML, Barnett L, et al. A holistic measurement model of movement competency in children. *Journal of Sports Sciences* 2016;34(5):477-85.
20. Valentini NC. Validity and reliability of the TGMD-2 for Brazilian children. *Journal of Motor Behavior* 2012;44(4):275-80.
21. Houwen S, Hartman E, Jonker L, et al. Reliability and validity of the TGMD-2 in primary-school-age children with visual impairments. *Adapted Physical Activity Quarterly* 2010;27(2):143-59.
22. Wong KYA, Yin Cheung S. Confirmatory factor analysis of the Test of Gross Motor Development-2. *Measurement in Physical Education & Exercise Science* 2010;14(3):202-09. doi: 10.1080/10913671003726968
23. 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. *Adapted physical activity quarterly : APAQ* 2008;25(1):71-82.
24. 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. *Research in Developmental Disabilities* 2013;34(2):801-8.
25. 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. doi: 10.1016/j.humov.2014.11.004
26. Visser L, Ruiter SAJ, Van der Meulen BF, et al. Low verbal assessment with the Bayley-III. *Research in Developmental Disabilities* 2015;36:230-43.
27. Wuang YP, Su CY. Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in children with intellectual disability. *Research in Developmental Disabilities* 2009;30(5):847-55.
28. Wuang YP, Su CY, Huang MH. Psychometric comparisons of three measures for assessing motor functions in preschoolers with intellectual disabilities. *Journal of Intellectual Disability Research* 2012;56(6):567-78.
29. 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(3):1046-51. doi: http://dx.doi.org/10.1016/j.ridd.2011.01.035
30. Smits-Engelsman BCM, 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? *Research in Developmental Disabilities* 2011;32(4):1370-77.
31. 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. *Research in Developmental Disabilities* 2013;34(2):795-800.

32. 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 & Development* 2016;38(8):706-13.
33. 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(5):490-96. doi: 10.1111/dmcn.13367 [published Online First: 2017/01/10]
34. Schulz J, Henderson SE, Sugden DA, et al. Structural validity of the Movement ABC-2 test: factor structure comparisons across three age groups. *Research in Developmental Disabilities* 2011;32(4):1361-9.
35. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation, reliability, and validity for Brazilian children. *Res Dev Disabil* 2014;35(3):733-40. doi: http://dx.doi.org/10.1016/j.ridd.2013.10.028
36. 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. *Developmental Medicine & Child Neurology* 2012;54(2):160-5.
37. Hands B, Larkin D, Rose E. The psychometric properties of the McCarron Assessment of Neuromuscular Development as a longitudinal measure with Australian youth. *Human Movement Science* 2013;32(3):485-97.
38. Danks M, Maideen MF, Burns YR, et al. The long-term predictive validity of early motor development in "apparently normal" ELBW survivors. *Early Human Development* 2012;88(8):637-41.
39. 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(1):40-45. doi: 10.1016/S1013-7025(09)70058-2
40. 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(3):141-49.
41. Burns YR, Ensbey RM, Norrie MA. The neuro-sensory motor developmental assessment part II: predictive and concurrent validity. *Aust J Physiother* 1989;35(3):151-57.
42. Barnett LM, Minto C, Lander N, et al. Interrater reliability assessment using the Test of Gross Motor Development-2. *Journal of Science and Medicine in Sport* 2014;17(6):667-70. doi: http://dx.doi.org/10.1016/j.jsams.2013.09.013
43. 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(2):18-28. doi: 10.5897/JPESM12.003
44. Kim Y, Park I, Kang M. Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly* 2012;29(4):346-65.
45. Kim CI, Han DW, Park IH. Reliability and validity of the test of gross motor development-II in Korean preschool children: applying AHP. *Research in Developmental Disabilities* 2014;35(4):800-7.
46. Wuang YP, Lin YH, Su CY. Rasch analysis of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in intellectual disabilities. *Research in Developmental Disabilities* 2009;30(6):1132-44.

47. 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(4):254-66. doi: 10.1111/j.1469-8749.2008.02025.x [published Online First: 2008/01/15]

48. McDowell I. Measuring health: a guide to rating scales and questionnaires: Oxford university press 2006.

49. Cools W, De Martelaer K, Samaey C, et al. Movement skill assessment of typically developing preschool children: a review of seven movement skill assessment tools.(Report). *Journal of Sports Science and Medicine* 2009;8(2):154.

50. 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(7):596-605.

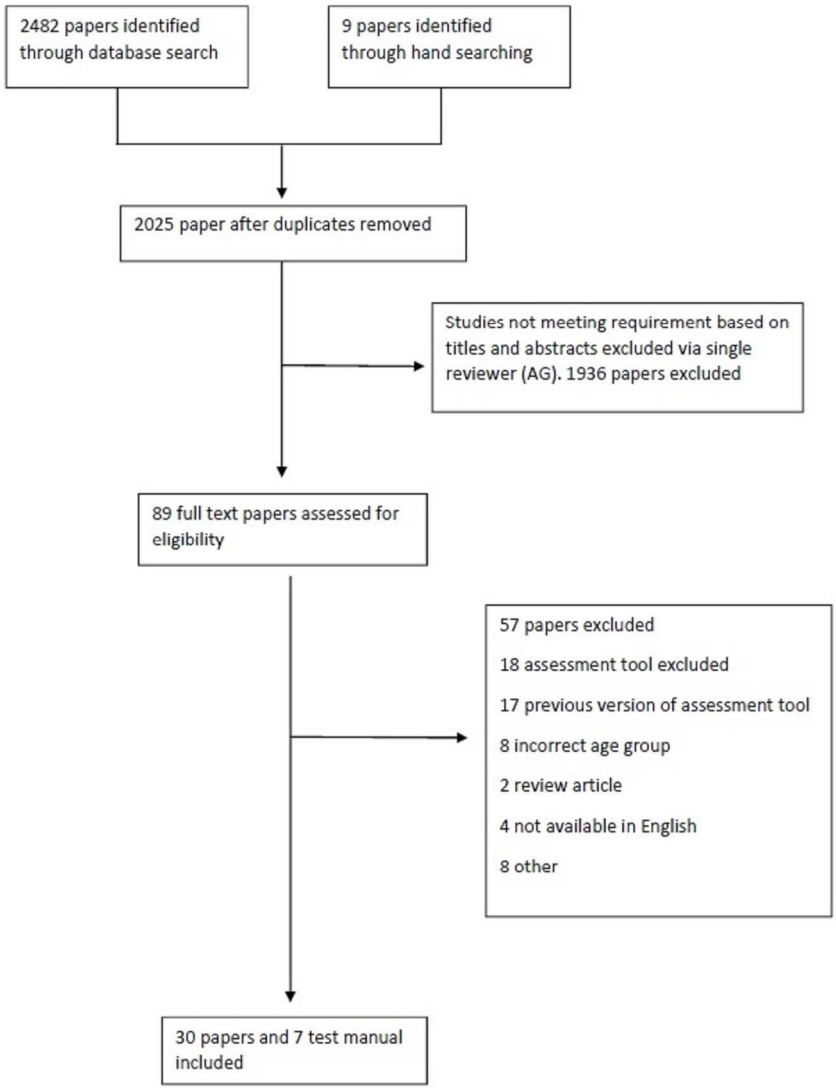


Figure 1. PRISMA flow diagram detailing study selection

195x256mm (300 x 300 DPI)

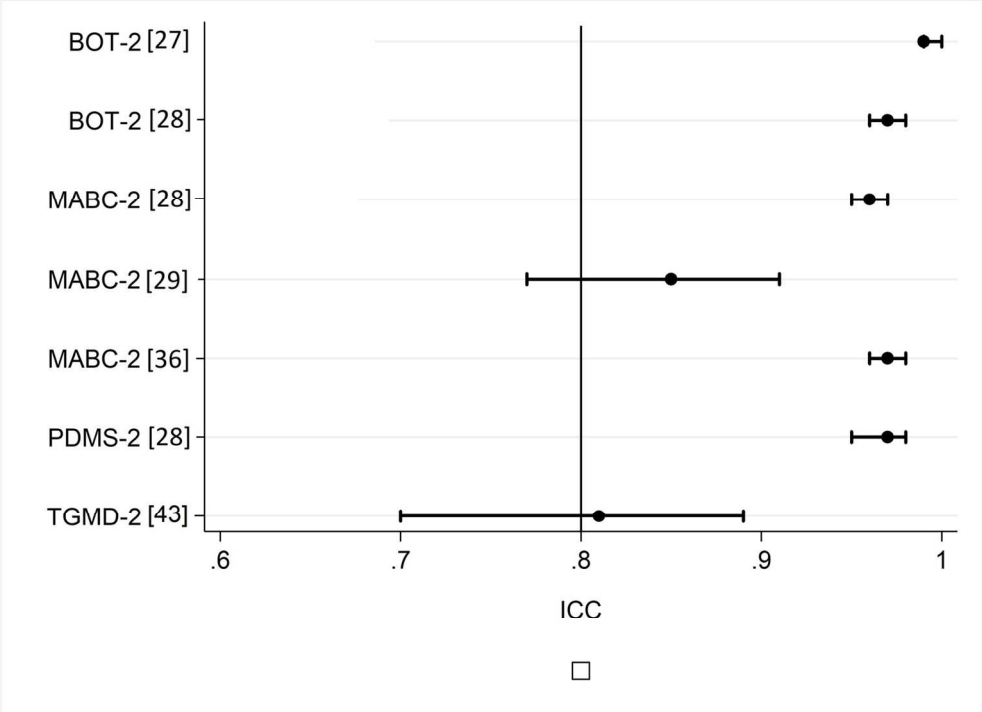


Figure 2. Test re-test reliability of gross motor assessment tools
Figure 2 legend: BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition 13; MABC-2, Movement Assessment Battery for Children 2nd edition 14; PDMS-2, Peabody Developmental Motor Scales 2nd edition 17; TGMD-II, Test of Gross Motor Development 2nd edition 18.

139x102mm (300 x 300 DPI)

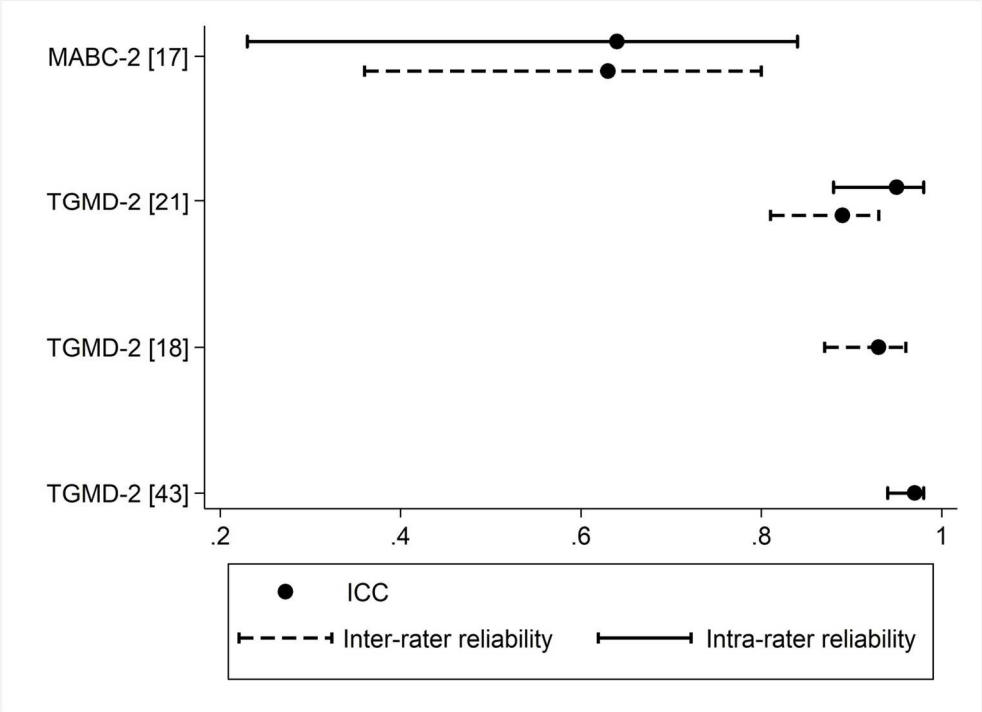


Figure 3. Inter and interrater reliability of gross motor assessment tools
Figure 3 legend: MABC-2, Movement Assessment Battery for Children 2nd edition 14; TGMD-II, Test of Gross Motor Development 2nd edition 18

139x102mm (300 x 300 DPI)

Supplementary Table 1. OVID Medline database search (1996 to present)

Search No	Search	Yield 5/5/17
1	Child/	811722
2	Child, Preschool/	457484
3	paediatric*.mp.	45528
4	Motor Skills/	12726
5	Motor Activity/	64838
6	gross motor.mp.	3821
7	Psychomotor Disorders/	2609
8	Motor Skills Disorders/	2580
9	Developmental Disabilities/	13484
10	developmental coordination disorder.mp.	845
11	Movement/ph (physiology)	22342
12	Questionnaires/	336296
13	"Outcome Assessment (Health Care)"/	57491
14	scale*.mp.	608566
15	instrument*.mp.	197131
16	outcome*.mp.	1813266
17	measure*.mp.	2255187
18	evaluat*.mp.	2552240
19	assess*.mp.	2273012
20	“Task Performance and Analysis”/	22017 (or 5969)
21	Reproducibility of Results”/	319899
22	1 or 2 or 3	936097
23	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	116200
24	12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20	6579985
25	21 and 22 and 23 and 24	1152

Supplementary table 2: CINAHL plus database search

Search Number	Search	Yield 2/7/17
S1	(MH "Child")	338732
S2	(MH "Child, Preschool")	156847
S3	"paediatric"	14351
S4	(MH "Motor Skills")	7420
S5	(MH "Motor Activity")	9664
S6	(MH “Psychomotor Performance”)	9457
S7	(MH "Motor Skills Disorders")	1515
S8	(MH "Developmental Disabilities")	7114
S9	(MH “Child Development Disorders”)	1708
S10	"gross motor"	2234
S11	(MH "Clinical Assessment Tools")	110291
S12	(MH "Outcome Assessment")	29335
S13	(MH Physical Therapy Assessment”)	2027
S14	“scale”	290924
S15	instrument*	113984
S16	outcome*	602132
S17	measure*	534465
S18	evaluat*	760447
S19	assess*	710533
S20	(MH Reliability and Validity”)	11043
S21	S1 OR S2 OR S3	387809
S22	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10	35562
S23	S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19	1763829
S24	S20 AND S21 AND S22 AND S23	144

Supplementary table 3: EMBASE database search (1974-present)

Search No	Search	Search Yield 5/5/17
1	Child/	1346641
2	Preschool child/	457981
3	paediatric*.mp.	86036
4	Motor performance/	57571
5	Motor Activity/	39751
6	psychomotor performance	19515
7	Motor development/	4906
8	Motor dysfunction/	53155
9	Developmental disorder/	30473
10	Gross motor.mp.	6840
11	Outcome Assessment/	358121
12	Outcome measure.mp.	60507
13	Questionnaire/	513199
14	Task performance/	125167
15	Functional assessment/	55415
16	Clinical assessment tool/	19865
17	evaluat*.mp.	3874341
18	instrument*.mp.	515930
19	outcome*.mp.	2425627
20	Assess*.mp.	3815907
21	Scale*.mp.	903216
22	Measure*.mp.	3444366
23	Measurement accuracy/	18209
24	Measurement repeatability/	2849
25	Reproducibility/	173988
26	Validity/	40192
27	Reliability/	114002
28	1 or 2 or 3	1535605
29	4 or 5 or 6 or 7 or 8 or 9 or 10	195237
30	11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22	10121631
31	23 or 24 or 25 or 26 or 27	324779
32	28 and 29 and 30 and 31	1105

Supplementary table 4. Allied and Complementary Medicine Database (AMED) database search: (1985-present)

Search Number	Search	Yield 2/7/17
1	Child/	15192
2	Child preschool/	1223
3	Adolescent/	3979
4	paediatric*.mp.	812
5	1 or 2 or 3 or 4	18429
6	Motor skills/	1220
7	Motor activity/	1468
8	Gross motor*.mp.	599
9	Psychomotor disorders/	1067
10	Developmental disabilities/ or motor skills disorders/	947
11	Developmental coordination disorder*.mp.	219
12	DCD.mp.	113
13	6 or 7 or 8 or 9 or 10 or 11 or 12	4982
14	5 and 13	1510
15	Clinical assessment scales/	4318
16	Questionnaires/	4123
17	Disability evaluation/	7023
18	Outcome measure*.mp/	9845
19	Outcome*.mp.	38379
20	Assess*.mp.	43680
21	Scale*.mp.	17562
22	Evaluat*.mp.	40621
23	15 or 16 or 17 or 18 or 19 or 20 or 21 or 22	93570
24	14 and 23	865
25	Measurement/	1629
26	Reproducibility of results.mp/	2241
27	“Consistency and reliability”/	1898
28	Statistics/	1075
29	Specificity.mp.	1241
30	Sensitivity.mp.	2860
31	“Predictive value of tests”/	839
32	25 or 26 or 27 or 28 or 29 or 30 or 31	10256
33	24 and 32	81

Supplementary table 5: Definition of terms

	Measurement Property	Definition	Example/explanation
Validity	Content	The degree to which an assessment tool's content measures the construct that it intends to measure [7]	Concerned with the relevance and comprehensiveness of the items included in the assessment tool
	Construct	Measures the degree to which the scores obtained from the test are an adequate reflection of the construct to be measured [7]	Examples include structural validity (whether scores reflect the dimensionality of the construct), hypothesis testing (item construct validity) and cross-cultural validity (whether translated or culturally adapted assessments adequately reflect the original version) [7]
	Criterion	Assesses whether or not the test scores reflect a 'gold standard' assessment [7]	As there is no gold standard of assessment for gross motor function in children this is often assessed with correlations of scores obtained from two or three other frequently used tools.
Reliability	Reliability	Refers to the consistency of a test score regardless of the time between assessments (test-retest) or the person administrating (intra and inter-rater) [50]	Usually measured with intraclass correlation coefficient (ICC), but can be measured using Cohen's kappa coefficient. Percentage agreement and Pearson's correlation coefficient do not incorporate error into the calculations and as such is not a true measure of agreement [50]. Scores > 0.80 are considered excellent, 0.60-0.79 adequate and <0.59 poor [11]
	Internal consistency	The degree of interrelatedness of an assessment tool's items [7]	Usually measured using Cronbach's alpha (α) [7]. scores > 0.70 demonstrates high relationship, 0.5 to 0.69 a moderate relationship, 0.26 to 0.49 a low relationship and < 0.26 little relationship [50].
	Measurement Error	Refers to the error obtained between measurements that cannot be attributed to the patients true change [7]	May be systematic or random error [7]
Responsiveness	Responsiveness	An assessment tool's ability to detect change over time in the construct it purports to measure [7]	This is central to a tools capacity to be used as an outcome measure.

Supplementary table 6: Excluded Assessment Tools

Reason	Assessments
Manual not available in English	Maastricht's Motor Test (MMT)
	The Motor-Proficiency-Test for children between 4 and 6 years of age (MOT 4-6)
	Zuk Assessment
	Körperkoordinationtest für Kinder (KTK)
Cannot extract meaningful gross motor score	Early Intervention Developmental Profile (EIDP)
	Neurological Developmental Exam
	Preschooler Gross Motor Quality Scale (PGMQ)
	The Malawi Developmental Assessment Tool (MDAT)
	Dutch table tennis motor skills assessment
Screening Tool	Brief Assessment of Motor Function (BAMF)
	The Motor Performance Checklist
	Motor skill checklist (MSC)
Diagnosis specific/requires a diagnosis	Assessment Battery for the Atypical Handicapped Child (VAB)
	Video-based documentation and rating system of the motor behaviour of handicapped children
Only assesses one motor domain (e.g. gait)	Standardized Walking Obstacle Course (SWOC)
	Timed floor to stand test
Manual not published/commercially available	Rapid Neurodevelopmental Assessment (RNDA)
	Tufts Assessment of Motor Performance (TAMP)
	Zurich Neuromotor Assessment (ZNA)

Supplementary table 7: Scoring and administration of assessment tools

Assessment Tool	Scoring	Interpretation of scores	Other
Bayley-III [12]	Motor score - gross (varying items) and fine motor (varying items) subscales. Binary score with reverse/discontinue rules	Raw scores Composite scores Centile ranks Age equivalents Growth scores	Lends itself to multidisciplinary team testing.
BOT-2 [13]	Fine manual (15 items) manual coordination (12 items) body coordination (16 items) strength and agility (10 items) subscales. Scoring differs for subtests	Raw scores Age adjusted standard scores Composite scores Centile ranks Age equivalents Descriptive categories. Complex conversions	Administration Easel includes instructions, diagrams and photos of test procedure
MABC-2[14]	Manual dexterity (3 items), aiming & catching (2 items) and balance (3 items) subscales.	Raw scores component scores centile ranks total test score traffic light system. Simple conversion	Also Available: MABC-2 Checklist (screening tool) and intervention manual
MAND [15]	Fine motor (5 items) Gross motor (5 items)	Raw scores Scaled scores converted to an NDI. Factor scores. Complex conversions	Case studies included in manual for hyperactivity, encephalitis, mild head trauma, CP and muscular dystrophy
NSMDA [16]	Functional grade given for each subscale, which is combined to create an overall score.	Indicates: normal range, minimal dysfunction, mild problems, moderate, severe or profound disability	Sections for comment on strengths, behavioural state during testing, musculoskeletal system and recommendations.
PDMS-2 [17]	GM: Stationary (30 items), locomotion (89 items), object manipulation (24 item). FM: grasping(26 items) , visual-motor integration (72 items)	Raw scores, Age equivalent, centile rank. Standard scores (subtests) Composite quotient. Complex conversions.	Motor activities program (intervention ideas)
TGMD-2 [18]	Locomotion (6 items) and Object Control (6 items). Separate male/female norms for object control subset	Raw scores, standard scores, percentile rank, age equivalent, Gross Motor Quotient. Simple conversion.	Simple to administer

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition [12]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; MAND, McCarron Assessment of Neuromuscular Development [15]; NSMDA, Neurological Sensory Motor Developmental Assessment [16]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17];; TGMD-II, Test of Gross Motor Development 2nd edition [18]; GM, Gross Motor; FM, Fine Motor; NDI, Neurodevelopmental Index

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1 Title page
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	-
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5-6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5 and supplementary tables 2
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5-6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6-7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7

PRISMA 2009 Checklist

Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	-

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	24, 25
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	-
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Figure 1 + page 7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1 – page 9 + Suppl table 3
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	11 + Table 3 – page 14-15
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	7-8, 11-12, 17 + Table 2 page 13, Table 4 page 18, Table 5 page 19, Table 6 page 20 + Figures 2 & 3
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	-
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	-
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	-

PRISMA 2009 Checklist

DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	22-24
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	24
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	25-26
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-021734.R2
Article Type:	Research
Date Submitted by the Author:	23-Aug-2018
Complete List of Authors:	Griffiths, Alison; Monash University, Physiotherapy; The Royal Children's Hospital, Physiotherapy Toovey, Rachel; Murdoch Childrens Research Institute, Developmental Disability and Rehabilitation Research; University of Melbourne, Physiotherapy Morgan, Prue; Monash University, Physiotherapy Spittle, Alicia; University of Melbourne, Physiotherapy; Murdoch Childrens Research Institute, Victorian Infant Brain Studies
Primary Subject Heading:	Paediatrics
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	PAEDIATRICS, REHABILITATION MEDICINE, validity, gross motor assessment, reliability

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]

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

Corresponding Author*:*

Name: A/Prof Alicia J. Spittle

Address: Department of Physiotherapy, The University of Melbourne, Level 7 Alan Gilbert Building, 161 Barry Street, Parkville, Vic 3010, Australia

Email: aspittle@unimelb.edu.au

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 Program.

Financial disclosure: The authors have declared that they have no financial relationships relevant to this article.

Conflict of interest: The authors have no conflict of interest.

Keywords: paediatrics, reliability, validity, rehabilitation medicine, gross motor assessment

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 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 (COSMIN) 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 re-test reliability is excellent in the BOT-2 (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 interrater (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 COSMIN 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 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 to 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]. Further, 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].

Health 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 [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 Movement Assessment Battery for Children (MABC) in age range, age bands, materials and tasks, that the MABC-2 requires independent reliability and validity assessment. Over the past eight 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 2-12 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) (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 required regarding availability of manuals authors were contacted.

Assessment tools were included if they were 1. Discriminative, predictive or evaluative of gross motor skills, 2. Assessed ≥ two gross motor (e.g. balance, jumping etc.) items, 3. Able to extract a meaningful gross motor sub-score, 4. Applicable to children 2-12 years of age, 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 (e.g. 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 agreement by 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] (Supplementary Table 5). Cross-cultural validity, structural validity and hypothesis testing are all considered to be a component of construct validity[7]. Whilst predictive validity is considered to be a component of content validity, it is reported on 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 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 patients 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), 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). The corresponding manuals were then added to the final yield resulting in thirty papers and seven manuals. Twenty assessment tools were excluded (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 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.

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 [12]	Gross motor, fine motor, cognitive, communication, social/emotional, adaptive	Static postures, dynamic movement, balance	1 mth – 3 yrs	Developmental delay: <25th centile or below 2SD. *	Discriminative	Predictive, Evaluative, Research tool	Norm	1700 children from the USA (2000)
BOT-2 [13]	Gross motor, fine motor	Coordination, balance, running speed and agility, strength	4 – 21 yrs	*	Discriminative Evaluative	Research tool	Norm	1520 children from the USA (2005)
MABC-2 [14]	Gross motor, fine motor, balance	Aiming and catching, static and dynamic balance	3 – 16 yrs	Traffic light system: Green = normal, amber = 'at risk' and red = definite motor impairment (<15%). *	Discriminative Evaluative	Intervention planning, Research tool	Norm	1172 children from United Kingdom (2006)
MAND [15]	Gross and fine motor	Coordination, jumping, static and dynamic balance	3 yrs – 25 yrs	NDI 70-85 = mild 55-69 = moderate <55 = severe disability *	Evaluative	Research tool	Norm	2000 3-35 yrs from the USA (1970's)
NSMDA [16]	Gross Motor, Fine Motor, Neurological, Postural Development, Infant Patterns of Movement, Sensory Motor. †	Sitting, kneeling, walking, balance, running, hopping, jumping, catching, motor planning	1 mth – 6 yrs	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	N/A
PDMS-2 [17]	Gross motor, fine motor	Stationary (standing balance, sit-ups, push-ups), locomotion (walking, running, jumping, hopping, etc.), object manipulation (kick, throw, hit, catch)	Birth – 5 yrs	*	Discriminative Evaluative	Predictive, Research tool	Norm	2003 USA and Canada (1997-8)

TGMD-2 [18]	Gross Motor	Locomotion (run, gallop, hop, leap, jump, slide) and Object control (batting, dribbling, catch, kick, throw, roll)	3 – 10 yrs	*	Discriminative Evaluative	Outcome measure, research tool, intervention planning	Norm	1208 USA children (1997-1998)

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18] NDI, Neurodevelopmental Index; SD, Standard Deviation; mth, month; yrs, years *, Advisable to use clinical reasoning; †, requires some manual handling; USA, United States of America

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 Supplementary Table 7. The shortest administration time is 15-20 minutes for the TGMD-2 and the MAND; whilst most manuals report 20-60 minutes 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 post assessment (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 [13] 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 [19 20 18 21 22 23]. 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 and Dutch speaking populations [24 25]. 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, whilst 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.

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 [12]	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 [13]	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 [14]	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 [15]	15-20	Therapist administers items in standardised order.	Professionals e.g. education, neurology, OT, PT, psychology etc.	Formal training not required.	Manual and test kit: £1366 includes equipment
NSMDA [16]	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 [17]	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 [18]	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 3rd edition [12]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; MAND, McCarron Assessment of Neuromuscular Development [15]; NSMDA, Neurological Sensory Motor Developmental Assessment [16]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17]; TGMD-II, Test of Gross Motor Development 2nd edition [18]; GM, Gross motor; OT, Occupational Therapy; PT, Physiotherapy; PE, Physical Education

Table 3. Methodological quality of included articles

Test	First author, Year	Country	Population (Age, Diagnosis)	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Criterion validity	Responsive - ness
BAYLEY III	Bayley [12]	USA	1-42 mths	Fair	Fair	Good	Excellent	Good	Good	-	Good	-
	Spittle, et al. [4]	Australia	2,4 yrs, Ex prem	-	-	-	-	-	-	-	Good	-
	Visser, et al. [26]	Netherlands	2.2-10.8 yrs, GDD, L.I.	-	-	-	Excellent	Poor	-	-	-	-
BOT-2	Wuang and Su [27]	Taiwan	4-12 yrs ID	Excellent	Excellent	Excellent	-	-	-	-	-	Fair
	Wuang, et al. [28]	Taiwan	3-6 yrs ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Bruininks and Bruininks [13]	USA	4-21 yrs	Good	Fair (interrater) Fair (test-retest)	Good	Excellent	Good	-	-	Good	-
MABC-2 (AB 1)	Ellinoudis, et al. [29]	Greece	3-5.5 yrs	Excellent	Good	-	-	-	-	-	-	-
	Hua, et al. [24]	China	3-6 yrs	Excellent	Good	-	Excellent	Excellent	-	Poor	Excellent	-
	Logan, et al. [5]	USA	3-6 yrs	-	-	-	-	-	Fair	-	Fair	-
	Smits-Engelsman, et al. [30]	Belgium	3-4 yrs	Poor	Poor	Poor	-	-	-	-	-	-
MABC-2 (AB 2)	Holm, et al. [31]	Norway	7-9 yrs	-	Fair (interrater) Poor (intrarater)	Poor	-	-	-	-	-	-
	Kita, et al. [32]	Japan	7-10 yrs	Excellent	-	-	-	-	-	Poor	-	-
MABC-2	Griffiths, et al. [33]	Australia	4-8 yrs	-	-	-	-	-	-	-	Good	-
	Henderson, et al. [14]	UK	3-16 yrs	-	Fair	Good	Excellent	-	-	-	-	-
	Niemeijer, et al. [25]	Netherlands + Belgium	-	-	-	-	-	-	-	Poor	-	-
	Schulz, et al. [34]	U.K	3-16 yrs	-	-	-	Excellent	Good	-	-	-	-
	Valentini, et al. [35]	Brazil	3-13 yrs	Fair	Fair	-	Fair	Poor	-	Poor	Poor	-

	Wuang, et al. [28]	Taiwan	3-6 yrs, ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Wuang, et al. [36]	Taiwan	6-12 yrs DCD	Poor	Fair	Good	-	-	-	-	-	Fair
MAND	Hands, et al. [37]	Australia	10-17 yrs	-	-	-	-	Excellent	-	-	-	-
	McCarron [15]	USA	7yrs	-	-	-	Fair	Poor	-	-	Poor	-
NSMDA	Danks, et al. [38]	Australia	2 + 4 yrs ELBW	-	-	-	-	-	-	-	Fair	-
	MacDonald and Burns [39]	Australia	2 + 4 yrs CP	-	-	-	-	Fair	-	-	Poor	-
	Burns, et al. [40]	Australia	1-24 mths VLBW	Poor	-	-	Poor	-	-	-	-	-
	Burns, et al. [41]	Australia	1-mnths VLBW	-	-	-	-	Poor	-	-	Fair	-
PDMS-2	Hua, et al. [24]	China	3-6 yrs.	Excellent	Good	-	Excellent	Excellent	-	Poor	Excellent	-
	Wuang, et al. [28]	Taiwan	3-6 yrs ID	Fair	Good	Good	-	-	-	-	Good	Fair
	Folio and Fewell [17]	USA	0-71 mths	Good	-	Poor	Excellent	Good	Good	-	Poor	-
TGMD-2	Barnett, et al. [42]	Australia	4-8 yrs	-	Fair	-	-	-	-	-	-	-
	Farrokhi, et al. [43]	Iran	3-11 yrs	Fair	Fair	-	Fair	Fair	-	-	-	-
	Houwen, et al. [21]	Netherlands	6-12 yrs VI	Fair	Fair	-	-	Fair	-	-	-	-
	Kim, et al. [44]	Korea	8-12 yrs ID	-	Poor	-	-	-	-	-	-	-
	Kim, et al. [45]	Korea	5-6 yrs	Poor	Fair	-	-	Poor	-	-	Poor	-
	Logan, et al. [5]	USA	3-6 yrs	-	-	-	-	-	Fair	-	Fair	-
	Rudd, et al. [19]	Australia	6-12 yrs	-	-	-	-	Good	-	-	-	-
	Simons, et al. [23]	Belgium	7-10 yrs ID	Good	Good (interrater) Poor (test-retest)	-	Excellent	Good	Good	-	-	-
	Valentini [20]	Brazil	3-10 yrs	Poor	Fair (test-retest) Good (intra,	-	Excellent	Good	-	Fair	Good	-

				interrater)							
Wong and Yin Cheung [22]	China	3-10 yrs	-	-	-	-	Fair	-	-	-	-
Ulrich [18]	USA	3-10 yrs	Good	Fair (test-retest) Poor (interrater)	Fair	Poor	Good	-	-	Fair	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17]; TGMD-II, Test of Gross Motor Development 2nd edition;[18] Mths, Months; yrs, years; DCD, Developmental Coordination Disorder; VI, Vision Impairment; ID, Intellectual Disability; GDD, global developmental delay; L.I, Language Impairment; ELBW, Extremely Low Birth Weight; VLBW, Very Low Birth Weight; CP, Cerebral Palsy; prem, premature; USA, United States of America

Reliability

Internal consistency of assessments are summarised in Table 6. The BOT-2's high internal consistency is well supported, including for children with an intellectual disability [28 46]. 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.

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). Intra-rater 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) [31 42]. Inter-rater reliability is also supported in the BOT-2 with Pearson 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 [47 48]. Only studies reporting ICC's are visually represented in Figures 2 (test-retest) and 3 (inter and intra-rater). The TGMD-2 test-retest reliability results from Houwen, et al. [21] were believed to contain an error as the reported ICC was outside of the reported confidence intervals (ICC 0.92, 0.82-0.91). 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 standard error of measurement (SEM) [28]. Sensitivity and specificity for detecting change was shown to be satisfactory in the MABC-2, PDMS-2 and MABC-2 [28] (Table 6). There have been no studies to date on the responsiveness of the TGMD-2, NSMDA or MAND.

Table 4: Content and construct validity of assessment tools

Test	Content	Construct
BAYLEY III	Expert opinion for standard and low verbal version [12 26]. Literature reviews. Gross motor score correlated with Motor component 0.70 [12]	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language impairment [12]. H_i (gross motor subset) = 0.52-0.97 for children with language impairment and 0.82-0.99 in control group [26]
BOT-2	Focus groups, product survey, pilot, national tryout and standardisation studies, professional reviews[13]	Factor analysis, scores increase with age, discriminates between normal and children with DCD (*N*=50), high-functioning ASD (*N* = 45) and mild-moderate ID (*N* = 66) [13]
MABC-2	Expert Panel, Stakeholder feedback, Literature review [31] Expert panel - clarity (validity content index 71.8-93.9, Kappa 0.76-0.88) and pertinence (98.5-99.3 and kappa 0.83-0.92) $p<0.001$ [35]	Factor analysis, correlation coefficients [29] Subtest correlations 0.65-0.76 $p<0.001$. Discriminates between ASD and control group [31]. Structural equation modelling (for each age group) [34]. Expert panel - adequate face validity [35]. Significant difference between TD, DCD and at risk DCD scores ($\eta 2 = 0.63$) $p< 0.0001$ [35]. UK norms not appropriate to use with Dutch/Flemish children as under/over-estimate risk of motor impairment [25]. In Chinese population: CFA initially rejected. Acceptable fit achieved after 2 items removed [24]. Age band 2 shows good validity in Japanese population [32].
MAND	Based on neuropsychological theory. Several rounds of revision/trials of tasks during development [15]	Factor analysis [15 37]. Scores increase with age, and discriminate between typically developing children and those with head trauma or neurological dysfunction as well as gender [15 37]
NSMDA	Literature review. Developed by an experienced paediatric physiotherapist [40]	Factor analysis (up to 2 years of age) [40 41]. Stability of test results over time (up to 2 years) [40 41].
PDMS-2	Literature review. Created by experts in the field. Revised with feedback from therapists guided revision. Hierarchical sequence of items [17]	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with age (r=0.80-0.93) [17]
TGMD-2	Expert Panel (3 PE teachers with post-grad qualifications) [18]. 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 [20]	Exploratory and confirmatory factor analysis [19 20 18 21 22 23] High and significant correlation of increasing age and increasing scores [43]. Age and disability differentiation [18 23] Subtest correlation 0.41 [18] Galloping, running and leaping not well correlated with locomotion subscale. Object control significant & highly correlated [45]. ANOVA - significant age effect for object control [23] Moderate correlation between items and subset scores, and between subset scores and total score [23]

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18]; H_i, scalability coefficient; CFA, Confirmatory Factor Analysis; TD, Typically Developing; ASD, Autism Spectrum Disorder, ID, Intellectual Disability; WPPSI, *Wechsler Preschool and Primary Scale of Intelligence*; WISC-R, Wechsler Preschool and Primary Scale of Intelligence-R; NDI, Neurodevelopmental Index; ANOVA, Analysis of Variance

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. [12]	Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.32-.037 specificity 0.97 <2SD sensitivity 0.18-0.21 specificity 1.00. CP at 4 years: Bayley III at 2 years <1SD sensitivity 0.83 specificity 0.94. <2SD sensitivity 0.67 specificity 1.0 [4]
BOT-2	MABC-2 p = 0.92 PDMS-2 p = 0.88 (N = 38) [28]. PDMS-2 Total motor composite r = 0.77 [13].	-
MABC-2	PDMS-2 ρ = 0.631 – 0.84 [28 24]. TGMD-2 ρ = 0.45 [5]. TGMD-2 standard scores (r = 0.3, p < 0.02) [35]. BOT-2 ρ =0.90 - 0.92 [28].	Classification groups (DCD, at risk and TD) remained same over time (6 months) $\chi 2$ = 0.67 p = 0.72 [35]. Predictive of motor impairment over 6-12 months (N=41) ICC 0.88 p < 0.007 [35]. Scores at 4 years predictive of motor impairment at 8 years in children born <30 weeks gestation (PPV 79, sensitivity 79%, specificity 93%)) [33]
MAND	Gross motor subscore: Low-moderate correlation with manual dexterity (-0.46 to 0.35), reaction time (-0.31 to -0.58), intelligence measures (WISC-R, Metropolitan Achievement Test) (0.30-0.39) and visual motor test (-0.33 to 0.39) [15]	-
NSMDA	NSMDA at 2 years (N = 148) predictive of medical diagnosis $\chi 2$ = 0.08 p = NS [41]	Motor outcome at 11-13 yrs. NSMDA at 2years - 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% [38]. If classified 'severe' at 24 months - approximately 50% chance walking at 4 years (moderate = 80%, mild = 93% minimal = 100%) [39]
PDMS-2	MABC-2 ρ = 0.63- 0.84, [24 28] MABC-2 gross motor composite ρ = 0.743 [24] BOT-2 ρ = 0.88 [28]. Mullen Scales of Early Learning GMQ = 0.86 FMQ = 0.80 [17]	-
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 [45] (N=41) Basic Motor Generalizations subtest of the CSSA r = 0.63. Locomotor 0.63 object control 0.41 [18]	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18] NS, Not Specified; SD, Standard Deviation; CP, Cerebral Palsy; TD, Typically Developing; ICC, Intraclass Correlation Coefficient; χ2, Chi Squared; NDI, Neurodevelopmental Index; CSSA, Comprehensive Scales of Student Abilities

Table 6: Reliability of assessment tools

Test	Internal Consistency	Test-Retest	Intra-rater	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) [12]	Gross Motor subtest (N=47) r=0.79 Motor component r=0.80 [12]	-	-	SEM Gross motor subtest 0.85-1.08. of Motor component = 3.00-4.74 (24-42 months) [12]	-
BOT-2	(N = 100) α = 0.92 [27] (N = 141) α = 0.86 [28] 4-7 yrs (N= 620) α = 0.95 8-11 yrs (N= 450) α = 0.95 [13]	(N = 100) ICC = 0.99 [27] (N = 141) ICC = 0.97 [28] 4-7 yrs (N = 43) r = 0.81 (8-12 yrs (N= 44) r = 0.80 [13]	-	Total motor composite 4-21 yrs (N = 47) r = 0.98 [13]	4.18 (sensitivity 55.10% specificity 72.55%) [27] 7.43 (sensitivity 42.49% specificity 65.72%) [28]	6.53 (sensitivity 48.98% specificity 76.47%) [27] 6.55 (sensitivity 49.99% specificity 58.78%) [28]
MABC-2 (AB 1)	(N = 60) M.D α = 0.51, A&C α = 0.70, Bal α = 0.66 [29] (N = 1823) α = 0.502 [24] (N=50) α = 0.81-0.87 [30]	(N=60) ICC = 0.85 [29] Item ICC's 0.830-0.985 [24] ICC test-retest = 0.83 [30] Inter-rater test-retest ICC = 0.79 [30]	(N=28) κ = 0.71 [30]	Item ICC's range 0.892-0.998 [24] (N=22) κ = 0.60 [30]	(N=28) Intrarater MDC = 3.43 (N=22) Inter-tester MDC = 3.81 [30]	-
MABC-2 (AB 2)	Translated version (Japanese) (N=132) α = 0.602 [32]	-	ICC = 0.64 [31]	ICC 0.63 [31]	Intra-rater SDC TTS: +/- 11.7 TSS +/- 3.3. Inter-rater SDC TTS +/-16.0 TSS +/- 3.8 [31]	-
MABC-2	Subscales α = 0.78 (M.D = 0.77, BS = 0.52, Bal = 0.77) [35] α = 0.88 [36] (N = 141) α = 0.88 [28]	N=60 (all 3 age bands) r=0.80 [14] r=0.74 p<0.0001 (standard score). ICC standard score = 0.85 [35] ICC 0.96 [36] N = 141 ICC =0.96 [28]	ICC 0.88 [35]	ICC 0.96-0.99 [35]	SEM 1.34 (95%CI) = 3 [14] 1.83 (95%CI) [36] 1.83 (sensitivity 69.69% specificity 52.10%) [28]	1.39 (sensitivity 72.47% specificity 46.18%) [28 36]
MAND	-	-	-	-	-	-
NSMDA	Cross correlation matrix Item scoring (12+24months) 0.73 p<0.001, Functional grade (12+24months) 0.87 p<0.001 [40]	-	-	-	-	-
PDMS-2	(N=141) α=0.89 [28] 24-35m α=0.97, 36-47m α=0.95, 48-59m α=0.97, 60-71m α=	N=141 ICC= 0.97 [28]	unable to extract data for ≥24months	unable to extract data for ≥24months [17]	7.76 (sensitivity 60.65% specificity 74.13%) [28] SEM 24-	8.39 (sensitivity 61.65% specificity 71.34%) [28]

	0.98. For subgroups† α=0.99 [17]		[17]		59 months = 3, 60-71m = 2 [17]	
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 [18]	*N*=63 ICC=0.81 95% CI [43] *N*=23 ICC=0.92 total 95% CI [21] *N*=99 *r*=0.98 [23] Locomotor test *r* = 0.90 *p* <0.0001 object control test *r* = 0.91 *p* <0.001 [20] *N* = 75 *r*=0.96 overall (3-5 yrs *r* = 0.91), 6-8 years *r* = 0.95), (9-10 years *r* = 0.94) [18]	*N*=32 ICC=0.97 95% CI [43] *N*=25 ICC=0.95 95% CI [21] ICC = 0.78 [44] ICC=0.92-0.99 [20]	Obj ICC=0.93 [42] (*N*=50) ICC=0.89 [21] ICC=0.75 [44] *N*=8 *r*= 1.00 [23] L.S ICC=0.88 Obj ICC=0.89 [20] *N* = 30 *r*=0.98 [18]	-	-

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;[12] BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition;[13] MABC-2, Movement Assessment Battery for Children 2nd edition;[14] MAND, McCarron Assessment of Neuromuscular Development;[15] NSMDA, Neurological Sensory Motor Developmental Assessment;[16] PDMS-2, Peabody Developmental Motor Scales 2nd edition;[17] TGMD-II, Test of Gross Motor Development 2nd edition;[18] GM, Gross Motor Subset; MC, Motor Component; *K*, Kappa Coefficient; M.D, Manual Dexterity; BS, Ball Skills; BAL, Balance; A&C, Aiming and catching; SDC, Smallest Detectable Change; TTS, Total Test Score; TSS, Total Standard Score; †, gender, ethnicity, speech/language or physical disorder; Obj, Object Control Subset; L.S, Locomotion Subset

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 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. [49] 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 ten additional publications to support the content, construct and criterion validity of the MABC-2 and have demonstrates 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.

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 two 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 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 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 or intra-rater 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 or intra-rater 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, 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 utilised 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 [18]. 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 (ICF) domains of participation, personal factors and environment [2]. Clinicians should utilise appropriate assessments or questionnaires to ensure that these domains of health are also addressed in line with World Health Organisation 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 inter-scorer difference (or interrater reliability) [18]. 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.

All of the included assessment tools were found to have merits and limitations in their clinical utility the body of evidence to support their use. 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. Whilst 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 (MOT 4-6), the Körperkoordinationtest für Kinder (KTK) and the Maastrichtse Motoriek Test (MMT)[49]. 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 tools. Younger children are more likely to be distracted and may not understand test items as well, which may also increase assessment times[30].

When a new edition of an assessment tools 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 [14], 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 [50], 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 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 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.

Author Contributions

All individuals listed as authors meet the appropriate authorship criteria and have approved the acknowledgement of their contributions. The primary author, Ms Griffiths was responsible for the drafting of the paper and liaising with the co-authors on findings and conclusions. Ms Toovey contributed to the paper through interpretation of data, completing methodological assessments and revising manuscript content throughout its development. A/Profs Morgan and Spittle both contributed to the paper through assisting with the development of research design, interpretation of data and revising manuscript content through its development.

Data Sharing Statement

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

Figures

Figure 1. PRISMA flow diagram detailing study selection

Figure 2. Test re-test reliability of gross motor assessment tools

Figure 2 legend: BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17]; TGMD-II, Test of Gross Motor Development 2nd edition [18].

Figure 3. Inter and interrater reliability of gross motor assessment tools

Figure 3 legend: MABC-2, Movement Assessment Battery for Children 2nd edition [14]; TGMD-II, Test of Gross Motor Development 2nd edition [18]

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. *Human Movement Science* 2006;25(1):65-75. doi: http://dx.doi.org/10.1016/j.humov.2005.10.011
2. World Health Organization. International Classification of Functioning, Disability and Health: ICF: World Health Organization 2001.
3. Magalhaes LC, Cardoso AA, Missiuna C. Activities and participation in children with developmental coordination disorder: a systematic review. *Research in developmental disabilities* 2011;32(4):1309-16. doi: 10.1016/j.ridd.2011.01.029 [published Online First: 2011/02/19]
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? *Developmental Medicine And Child Neurology* 2013;55(5):448-52. doi: 10.1111/dmcn.12049
5. Logan SW, Robinson LE, Getchell N. The Comparison of Performances of Preschool Children on Two Motor Assessments. *Perceptual and Motor Skills* 2011;113(3):715-23.
6. Venetsanou F, Kambas A. Environmental factors affecting preschoolers' motor development. *Early Childhood Education Journal* 2010;37(4):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. *Journal of clinical epidemiology* 2010;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006 [published Online First: 2010/05/25]
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. *PEDIATRIC PHYSICAL THERAPY* 2010;22(2):170.
9. Brown T, Lalor A. The Movement Assessment Battery for Children—Second Edition (MABC-2): A Review and Critique. *Physical & Occupational Therapy in Pediatrics* 2009;29(1):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. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2012;21(4):651-7. doi: 10.1007/s11136-011-9960-1 [published Online First: 2011/07/07]
11. Law M. Outcome measures rating form. Ontario, Canada: CanChild Centre for Disability Research, 2004.
12. Bayley N. Bayley Scales of Infant Development and Toddler Development: Technical Manual: The PsychCorp 2006.
13. Bruininks R, Bruininks B. Bruininks-Oseretsky Test of Motor Proficiency–2nd Edition (BOT-2): Manual. Circle Pines: MN: AGS Publishing 2005.
14. Henderson SE, Sugden DA, Barnett AL. Movement assessment battery for children-2: Movement ABC-2: Examiner's manual: Pearson 2007.

15. McCarron LT. MAND: McCarron assessment of neuromuscular development, fine and gross motor abilities: McCarron-Dial Systems, Incorporated 1997.
16. Burns YR. N.S.M.D.A Physiotherapy Assessment for Infants and Young Children Second Edition. Brisbane, Queensland: CopyRight Publishing Company 2014.
17. Folio M, Fewell R. Peabody Developmental Motor Scales. Examiner's Manual . 2nd Edition. Austin, Texas.: Pro-Ed. 2000.
18. Ulrich DA. Test of gross motor development-2. *Austin: Prod-Ed* 2000
19. Rudd J, Butson ML, Barnett L, et al. A holistic measurement model of movement competency in children. *Journal of Sports Sciences* 2016;34(5):477-85.
20. Valentini NC. Validity and reliability of the TGMD-2 for Brazilian children. *Journal of Motor Behavior* 2012;44(4):275-80.
21. 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 Act Q* 2010;27(2):143-59.
22. Wong KYA, Yin Cheung S. Confirmatory factor analysis of the Test of Gross Motor Development-2. *Measurement in Physical Education & Exercise Science* 2010;14(3):202-09. doi: 10.1080/10913671003726968
23. 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. *Adapted physical activity quarterly : APAQ* 2008;25(1):71-82.
24. 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. *Research in developmental disabilities* 2013;34(2):801-8.
25. 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. doi: 10.1016/j.humov.2014.11.004
26. Visser L, Ruiter SAJ, Van der Meulen BF, et al. Low verbal assessment with the Bayley-III. *Research in developmental disabilities* 2015;36:230-43.
27. Wuang YP, Su CY. Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in children with intellectual disability. *Research in developmental disabilities* 2009;30(5):847-55.
28. Wuang YP, Su CY, Huang MH. Psychometric comparisons of three measures for assessing motor functions in preschoolers with intellectual disabilities. *Journal of Intellectual Disability Research* 2012;56(6):567-78.
29. Ellinoudis T, Evaggelinou C, Kourtessis T, et al. Reliability and validity of age band 1 of the Movement Assessment Battery for Children--second edition. *Research in developmental disabilities* 2011;32(3):1046-51. doi: http://dx.doi.org/10.1016/j.ridd.2011.01.035
30. Smits-Engelsman BCM, 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? *Research in developmental disabilities* 2011;32(4):1370-77.

31. 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. *Research in developmental disabilities* 2013;34(2):795-800.
32. 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 & Development* 2016;38(8):706-13.
33. 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(5):490-96. doi: 10.1111/dmcn.13367 [published Online First: 2017/01/10]
34. Schulz J, Henderson SE, Sugden DA, et al. Structural validity of the Movement ABC-2 test: factor structure comparisons across three age groups. *Research in developmental disabilities* 2011;32(4):1361-9.
35. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation, reliability, and validity for Brazilian children. *Research in developmental disabilities* 2014;35(3):733-40. doi: http://dx.doi.org/10.1016/j.ridd.2013.10.028
36. 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. *Developmental Medicine & Child Neurology* 2012;54(2):160-5.
37. Hands B, Larkin D, Rose E. The psychometric properties of the McCarron Assessment of Neuromuscular Development as a longitudinal measure with Australian youth. *Human Movement Science* 2013;32(3):485-97.
38. Danks M, Maideen MF, Burns YR, et al. The long-term predictive validity of early motor development in "apparently normal" ELBW survivors. *Early Human Development* 2012;88(8):637-41.
39. 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(1):40-45. doi: 10.1016/S1013-7025(09)70058-2
40. Burns YR, Ensbey RM, Norrie MA. The Neuro-sensory motor developmental assessment part 1: development and administration of the test. *Australian Journal of Physiotherapy* 1989;35(3):141-49.
41. Burns YR, Ensbey RM, Norrie MA. The neuro-sensory motor developmental assessment part II: predictive and concurrent validity. *Australian Journal of Physiotherapy* 1989;35(3):151-57.
42. Barnett LM, Minto C, Lander N, et al. Interrater reliability assessment using the Test of Gross Motor Development-2. *Journal of Science and Medicine in Sport* 2014;17(6):667-70. doi: http://dx.doi.org/10.1016/j.jsams.2013.09.013
43. 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(2):18-28. doi: 10.5897/JPESM12.003
44. Kim Y, Park I, Kang M. Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapt Phys Act Q* 2012;29(4):346-65.
45. Kim CI, Han DW, Park IH. Reliability and validity of the test of gross motor development-II in Korean preschool children: applying AHP. *Research in developmental disabilities* 2014;35(4):800-7.

46. Wuang YP, Lin YH, Su CY. Rasch analysis of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in intellectual disabilities. *Research in developmental disabilities* 2009;30(6):1132-44.

47. 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(4):254-66. doi: 10.1111/j.1469-8749.2008.02025.x [published Online First: 2008/01/15]

48. McDowell I. Measuring health: a guide to rating scales and questionnaires: Oxford university press 2006.

49. Cools W, De Martelaer K, Samaey C, et al. Movement skill assessment of typically developing preschool children: a review of seven movement skill assessment tools.(Report). *Journal of Sports Science and Medicine* 2009;8(2):154.

50. Adair B, Said CM, Rodda J, et al. Psychometric properties of functional mobility tools in hereditary spastic paraplegia and other childhood neurological conditions. *Developmental Medicine & Child Neurology* 2012;54(7):596-605.

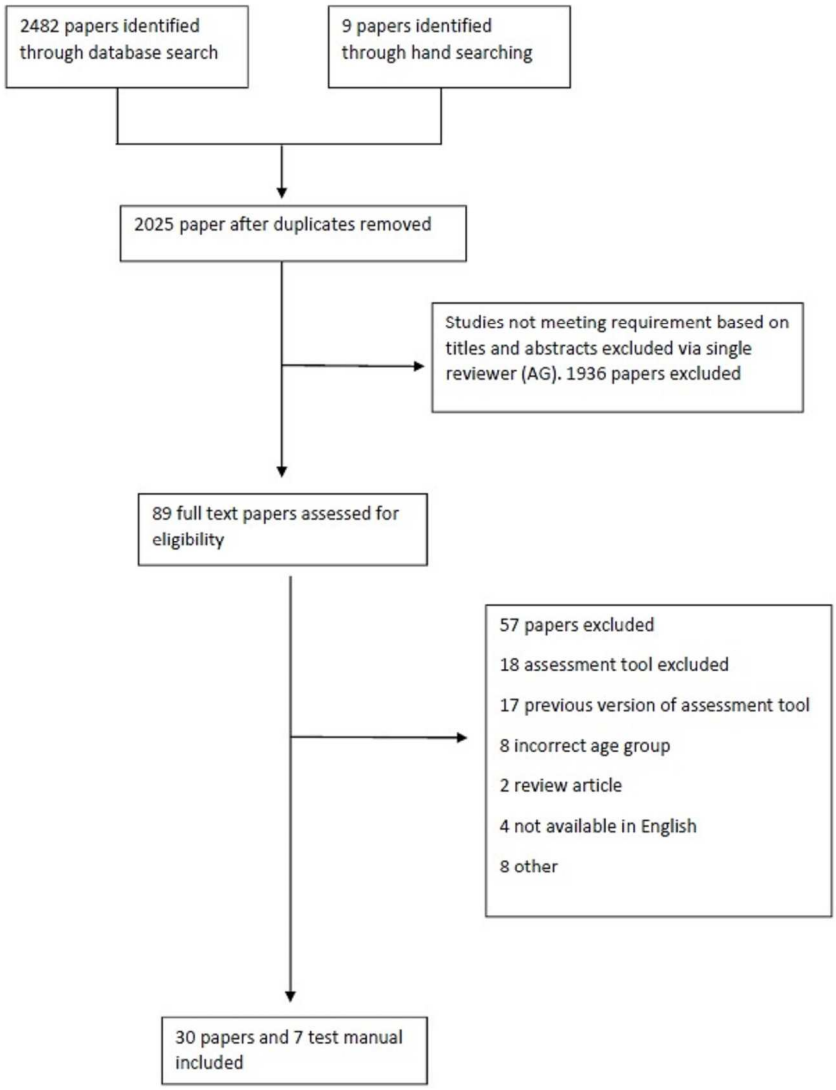


Figure 1. PRISMA flow diagram detailing study selection

195x256mm (300 x 300 DPI)

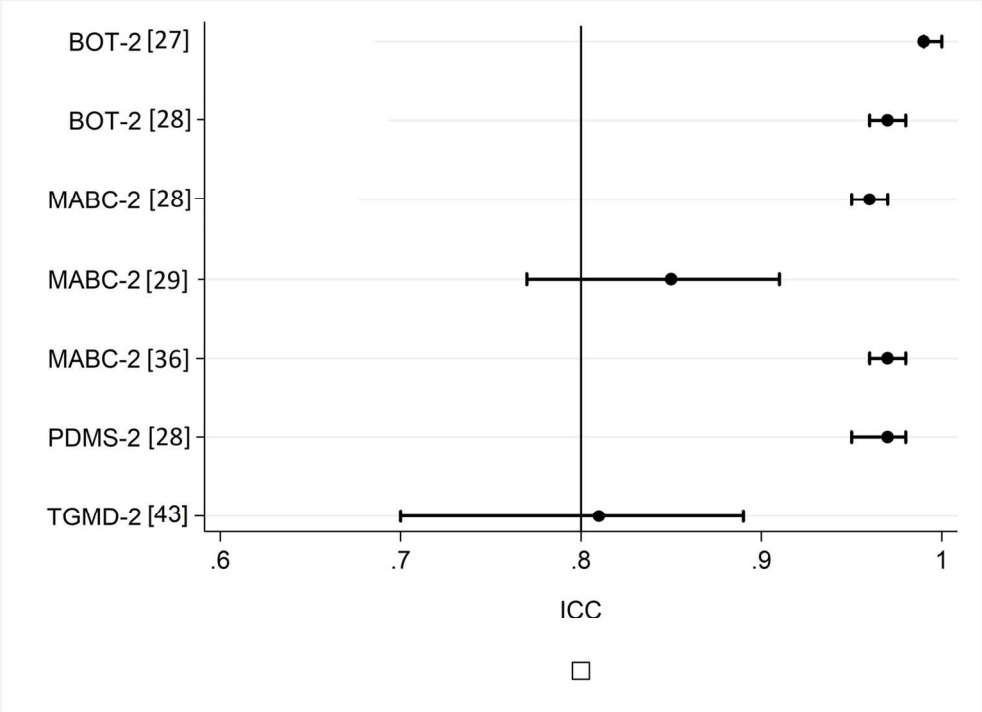


Figure 2. Test re-test reliability of gross motor assessment tools
Figure 2 legend: BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition 13; MABC-2, Movement Assessment Battery for Children 2nd edition 14; PDMS-2, Peabody Developmental Motor Scales 2nd edition 17; TGMD-II, Test of Gross Motor Development 2nd edition 18.

139x102mm (300 x 300 DPI)

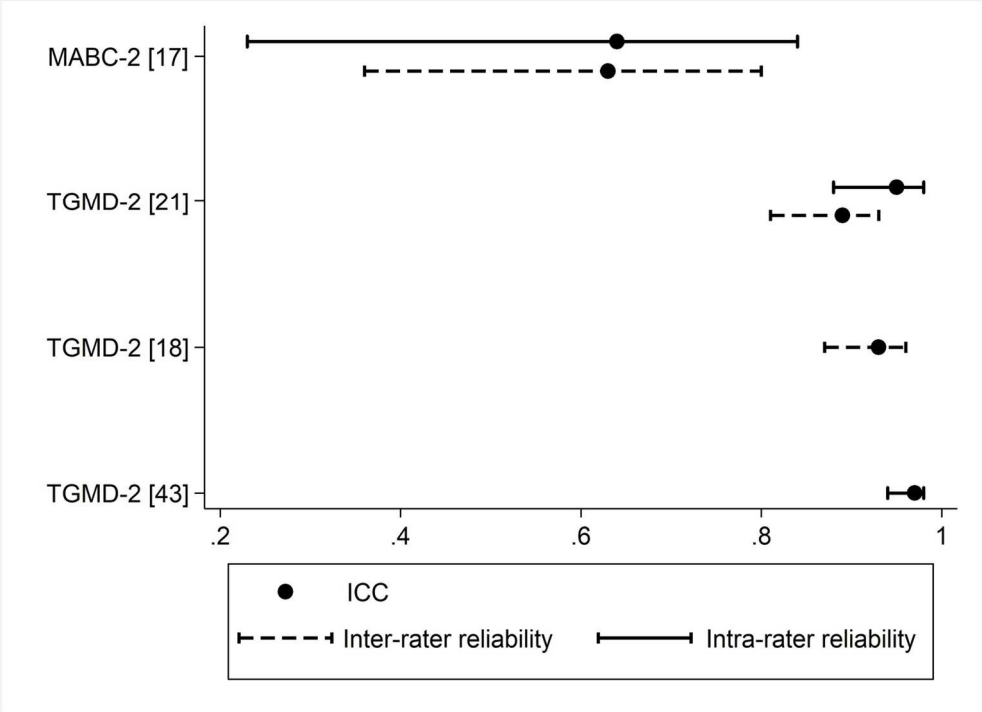


Figure 3. Inter and interrater reliability of gross motor assessment tools
Figure 3 legend: MABC-2, Movement Assessment Battery for Children 2nd edition 14; TGMD-II, Test of Gross Motor Development 2nd edition 18

139x102mm (300 x 300 DPI)

Supplementary Table 1. OVID Medline database search (1996 to present)

Search No	Search	Yield 5/5/17
1	Child/	811722
2	Child, Preschool/	457484
3	paediatric*.mp.	45528
4	Motor Skills/	12726
5	Motor Activity/	64838
6	gross motor.mp.	3821
7	Psychomotor Disorders/	2609
8	Motor Skills Disorders/	2580
9	Developmental Disabilities/	13484
10	developmental coordination disorder.mp.	845
11	Movement/ph (physiology)	22342
12	Questionnaires/	336296
13	"Outcome Assessment (Health Care)"/	57491
14	scale*.mp.	608566
15	instrument*.mp.	197131
16	outcome*.mp.	1813266
17	measure*.mp.	2255187
18	evaluat*.mp.	2552240
19	assess*.mp.	2273012
20	“Task Performance and Analysis”/	22017 (or 5969)
21	Reproducibility of Results”/	319899
22	1 or 2 or 3	936097
23	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	116200
24	12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20	6579985
25	21 and 22 and 23 and 24	1152

Supplementary table 2: CINAHL plus database search

Search Number	Search	Yield 2/7/17
S1	(MH "Child")	338732
S2	(MH "Child, Preschool")	156847
S3	"paediatric"	14351
S4	(MH "Motor Skills")	7420
S5	(MH "Motor Activity")	9664
S6	(MH “Psychomotor Performance”)	9457
S7	(MH "Motor Skills Disorders")	1515
S8	(MH "Developmental Disabilities")	7114
S9	(MH “Child Development Disorders”)	1708
S10	"gross motor"	2234
S11	(MH "Clinical Assessment Tools")	110291
S12	(MH "Outcome Assessment")	29335
S13	(MH Physical Therapy Assessment”)	2027
S14	“scale”	290924
S15	instrument*	113984
S16	outcome*	602132
S17	measure*	534465
S18	evaluat*	760447
S19	assess*	710533
S20	(MH Reliability and Validity”)	11043
S21	S1 OR S2 OR S3	387809
S22	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10	35562
S23	S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19	1763829
S24	S20 AND S21 AND S22 AND S23	144

Supplementary table 3: EMBASE database search (1974-present)

Search No	Search	Search Yield 5/5/17
1	Child/	1346641
2	Preschool child/	457981
3	paediatric*.mp.	86036
4	Motor performance/	57571
5	Motor Activity/	39751
6	psychomotor performance	19515
7	Motor development/	4906
8	Motor dysfunction/	53155
9	Developmental disorder/	30473
10	Gross motor.mp.	6840
11	Outcome Assessment/	358121
12	Outcome measure.mp.	60507
13	Questionnaire/	513199
14	Task performance/	125167
15	Functional assessment/	55415
16	Clinical assessment tool/	19865
17	evaluat*.mp.	3874341
18	instrument*.mp.	515930
19	outcome*.mp.	2425627
20	Assess*.mp.	3815907
21	Scale*.mp.	903216
22	Measure*.mp.	3444366
23	Measurement accuracy/	18209
24	Measurement repeatability/	2849
25	Reproducibility/	173988
26	Validity/	40192
27	Reliability/	114002
28	1 or 2 or 3	1535605
29	4 or 5 or 6 or 7 or 8 or 9 or 10	195237
30	11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22	10121631
31	23 or 24 or 25 or 26 or 27	324779
32	28 and 29 and 30 and 31	1105

Supplementary table 4. Allied and Complementary Medicine Database (AMED) database search: (1985-present)

Search Number	Search	Yield 2/7/17
1	Child/	15192
2	Child preschool/	1223
3	Adolescent/	3979
4	paediatric*.mp.	812
5	1 or 2 or 3 or 4	18429
6	Motor skills/	1220
7	Motor activity/	1468
8	Gross motor*.mp.	599
9	Psychomotor disorders/	1067
10	Developmental disabilities/ or motor skills disorders/	947
11	Developmental coordination disorder*.mp.	219
12	DCD.mp.	113
13	6 or 7 or 8 or 9 or 10 or 11 or 12	4982
14	5 and 13	1510
15	Clinical assessment scales/	4318
16	Questionnaires/	4123
17	Disability evaluation/	7023
18	Outcome measure*.mp/	9845
19	Outcome*.mp.	38379
20	Assess*.mp.	43680
21	Scale*.mp.	17562
22	Evaluat*.mp.	40621
23	15 or 16 or 17 or 18 or 19 or 20 or 21 or 22	93570
24	14 and 23	865
25	Measurement/	1629
26	Reproducibility of results.mp/	2241
27	“Consistency and reliability”/	1898
28	Statistics/	1075
29	Specificity.mp.	1241
30	Sensitivity.mp.	2860
31	“Predictive value of tests”/	839
32	25 or 26 or 27 or 28 or 29 or 30 or 31	10256
33	24 and 32	81

Supplementary table 5: Definition of terms

	Measurement Property	Definition	Example/explanation
Validity	Content	The degree to which an assessment tool's content measures the construct that it intends to measure [7]	Concerned with the relevance and comprehensiveness of the items included in the assessment tool
	Construct	Measures the degree to which the scores obtained from the test are an adequate reflection of the construct to be measured [7]	Examples include structural validity (whether scores reflect the dimensionality of the construct), hypothesis testing (item construct validity) and cross-cultural validity (whether translated or culturally adapted assessments adequately reflect the original version) [7]
	Criterion	Assesses whether or not the test scores reflect a 'gold standard' assessment [7]	As there is no gold standard of assessment for gross motor function in children this is often assessed with correlations of scores obtained from two or three other frequently used tools.
Reliability	Reliability	Refers to the consistency of a test score regardless of the time between assessments (test-retest) or the person administrating (intra and inter-rater) [50]	Usually measured with intraclass correlation coefficient (ICC), but can be measured using Cohen's kappa coefficient. Percentage agreement and Pearson's correlation coefficient do not incorporate error into the calculations and as such is not a true measure of agreement [50]. Scores > 0.80 are considered excellent, 0.60-0.79 adequate and <0.59 poor [11]
	Internal consistency	The degree of interrelatedness of an assessment tool's items [7]	Usually measured using Cronbach's alpha (α) [7]. scores > 0.70 demonstrates high relationship, 0.5 to 0.69 a moderate relationship, 0.26 to 0.49 a low relationship and < 0.26 little relationship [50].
	Measurement Error	Refers to the error obtained between measurements that cannot be attributed to the patients true change [7]	May be systematic or random error [7]
Responsiveness	Responsiveness	An assessment tool's ability to detect change over time in the construct it purports to measure [7]	This is central to a tools capacity to be used as an outcome measure.

Supplementary table 6: Excluded Assessment Tools

Reason	Assessments
Manual not available in English	Maastricht's Motor Test (MMT)
	The Motor-Proficiency-Test for children between 4 and 6 years of age (MOT 4-6)
	Zuk Assessment
	Körperkoordinationtest für Kinder (KTK)
Cannot extract meaningful gross motor score	Early Intervention Developmental Profile (EIDP)
	Neurological Developmental Exam
	Preschooler Gross Motor Quality Scale (PGMQ)
	The Malawi Developmental Assessment Tool (MDAT)
	Dutch table tennis motor skills assessment
Screening Tool	Brief Assessment of Motor Function (BAMF)
	The Motor Performance Checklist
	Motor skill checklist (MSC)
Diagnosis specific/requires a diagnosis	Assessment Battery for the Atypical Handicapped Child (VAB)
	Video-based documentation and rating system of the motor behaviour of handicapped children
Only assesses one motor domain (e.g. gait)	Standardized Walking Obstacle Course (SWOC)
	Timed floor to stand test
Manual not published/commercially available	Rapid Neurodevelopmental Assessment (RNDA)
	Tufts Assessment of Motor Performance (TAMP)
	Zurich Neuromotor Assessment (ZNA)

Supplementary table 7: Scoring and administration of assessment tools

Assessment Tool	Scoring	Interpretation of scores	Other
Bayley-III [12]	Motor score - gross (varying items) and fine motor (varying items) subscales. Binary score with reverse/discontinue rules	Raw scores Composite scores Centile ranks Age equivalents Growth scores	Lends itself to multidisciplinary team testing.
BOT-2 [13]	Fine manual (15 items) manual coordination (12 items) body coordination (16 items) strength and agility (10 items) subscales. Scoring differs for subtests	Raw scores Age adjusted standard scores Composite scores Centile ranks Age equivalents Descriptive categories. Complex conversions	Administration Easel includes instructions, diagrams and photos of test procedure
MABC-2[14]	Manual dexterity (3 items), aiming & catching (2 items) and balance (3 items) subscales.	Raw scores component scores centile ranks total test score traffic light system. Simple conversion	Also Available: MABC-2 Checklist (screening tool) and intervention manual
MAND [15]	Fine motor (5 items) Gross motor (5 items)	Raw scores Scaled scores converted to an NDI. Factor scores. Complex conversions	Case studies included in manual for hyperactivity, encephalitis, mild head trauma, CP and muscular dystrophy
NSMDA [16]	Functional grade given for each subscale, which is combined to create an overall score.	Indicates: normal range, minimal dysfunction, mild problems, moderate, severe or profound disability	Sections for comment on strengths, behavioural state during testing, musculoskeletal system and recommendations.
PDMS-2 [17]	GM: Stationary (30 items), locomotion (89 items), object manipulation (24 item). FM: grasping(26 items) , visual-motor integration (72 items)	Raw scores, Age equivalent, centile rank. Standard scores (subtests) Composite quotient. Complex conversions.	Motor activities program (intervention ideas)
TGMD-2 [18]	Locomotion (6 items) and Object Control (6 items). Separate male/female norms for object control subset	Raw scores, standard scores, percentile rank, age equivalent, Gross Motor Quotient. Simple conversion.	Simple to administer

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition [12]; BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2nd edition [13]; MABC-2, Movement Assessment Battery for Children 2nd edition [14]; MAND, McCarron Assessment of Neuromuscular Development [15]; NSMDA, Neurological Sensory Motor Developmental Assessment [16]; PDMS-2, Peabody Developmental Motor Scales 2nd edition [17];; TGMD-II, Test of Gross Motor Development 2nd edition [18]; GM, Gross Motor; FM, Fine Motor; NDI, Neurodevelopmental Index

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1 Title page
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	-
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5-6
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5 and suppl. tables 1, 2, 3, 4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5-6
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6-7
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6-7
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	-

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	24, 25
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	-
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Figure 1 + page 7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1 – page 9 + Suppl table 7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	11 + Table 3 (page 14-15)
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	7-8, 11-12, 17 + Table 2 page 13, Table 4 page 18, Table 5 page 19, Table 6 page 20 + Figures 2 & 3
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	-
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	-
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	-
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	22-25

PRISMA 2009 Checklist

Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	24
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	26
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.