

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<u>http://bmjopen.bmj.com</u>).

If you have any questions on BMJ Open's open peer review process please email <u>info.bmjopen@bmj.com</u>

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

SCHOLARONE[™] Manuscripts

BMJ Open

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

Alison Griffiths^{1, 2, 3}, Rachel Toovey^{3, 4}, Prue E. Morgan¹, 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

1 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.

7 Method:

8 A systematic search of MEDLINE, Embase, CINAHL and AMED was performed.

9 Methodological quality was assessed with the COnsensus-based Standards for the selection

10 of health status Measurement INstruments (COSMIN) checklist and an outcome measures

11 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. Testretest 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
 - ² motor assessment tools are urgently needed.

³ 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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 Introduction

Motor function promotes cognitive and perceptual development in children and contributes to their ability to participate in their home, school and community environments¹. Motor impairment can negatively affect activity and participation levels of children², which may lead to lower levels of physical activity, fitness and health into adulthood³. 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 peers³. 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 children⁴. Further, assessment tools are essential to monitor the effects of intervention⁴. 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.

BMJ Open

A systematic review in 2010 by Slater⁸ 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 Lalor⁹ 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.

15 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 ≥ 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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

3	
Δ	
5	
ر د	
6	
7	
8	
9	
10	
11	
12	
12	
13	
14	
15	
16	
17	
18	
19	
20	
20 21	
21	
22	
23	
24	
25	
26	
27	
28	
20	
29	
30	
31	
32	
33	
34	
35	
36	
37	
27	
38	
39	
40	
41	
42	
43	
44	
45	
45	
40	
4/	
48	
49	
50	
51	
52	
52	
55	
54 57	
55	
56	
57	
58	
59	
60	

1 2

 $_{\rm 1}$ $\,$ Articles describing use of the assessment tool were included if \geq 90% of the study

2 population were within 2-12 years of age, it was available in English and if validity and/or

3 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.

BMJ Open

2		
3	1	Results
4	-	
6	2	Figure 1 provides details of study selection. Eight assessment tools were identified for
7 8	3	inclusion; Bayley Scale of Infant and Toddler Development III (Bayley-III), Bruininks-
9 10	4	Oseretsky Test of Motor Proficiency 2 (BOT-2), Movement Assessment Battery for Children 2
11	5	(MABC-2), McCarron Assessment of Neuromuscular Development (MAND), Neurological
12	6	Sensory Motor Developmental Assessment (NSMDA), Peabody Developmental Motor Scales
14 15	7	2 (PDMS-2), and Test of Gross Motor Development 2 (TGMD-2). The corresponding manuals
16 17	8	were then added to the final yield resulting in 30 papers and 7 manuals. Nineteen
18 19	9	assessment tools were excluded (Supplementary Table 2).
20 21	10	The majority of assessment tools identified in this review are discriminative and most lend
22 23	11	themselves towards use in a research setting. All norm referenced tools are from western
24 25	12	countries and each identified test covers a different age range as shown in Table 1.
26 27	13	Most of the tools assess at least two domains of function: gross motor and fine motor skills,
28 20	14	although the Bayley III and the NSMDA assess six to seven different domains of
30 31	15	development. The TGMD-2 is the only tool that only assesses gross motor skills.
32 33	16	There is some consistency of items included within the gross motor skill subsets between
34	17	tests. Most include a locomotion task such as walking, running or stair climbing; an object
35 36	18	control or manipulation task such as throwing or catching a ball; and a static or dynamic
37 38	19	balance task such as standing on one leg or hopping. The PDMS-2 and the MAND also
39 40	20	include strength assessments (the PDMS-2 only in some age groups).
41 42	21	The number of gross motor items for assessment vary both within and between the tools
43 44	22	(Table 1). For example, the number of items tested in the Bayley-III and the PDMS-2
45 46	23	depends on the age and ability of the child. Several assessments report criteria for
47	24	describing gross motor delay, although all test manuals warn against diagnosing delay based
48 49 50 51 52	25	on a single assessment.

Table 1. Gross Motor Assessment Tool Characteristics

Assessment	Domains Tested	Gross motor	Age	Diagnostic criteria	Primary	Secondary	Type of test	Normative
Tool		components tested	range		purpose	purpose		sample (year)
Bayley-III ¹²	Gross motor, fine	Static postures, dynamic	1 mth – 3	Developmental delay:	Discriminative	Predictive,	Norm	1700 children
	motor, cognitive,	movement, balance	yrs	<25th centile or below		Evaluative,		from the USA
	communication,			2SD. *		Research tool		(2000)
	social/emotional,							
	adaptive							
BOT-2 ¹³	Gross motor, fine	Coordination, balance,	4 – 21 yrs	*	Discriminative	Research tool	Norm	1520 children
	motor	running speed and			Evaluative			from the USA
		agility, strength						(2005)
MABC-2 ¹⁴	Gross motor, fine	Aiming and catching,	3 – 16 yrs	Traffic light system: Green	Discriminative	Intervention	Norm	1172 children
	motor, balance	static and dynamic		= normal, amber = 'at risk'	Evaluative	planning,		from United
		balance		and red = definite motor		Research tool		Kingdom (2006)
				impairment (<15%). *				
MAND 15	Gross and fine motor	Coordination, jumping,	3 yrs – 25	NDI 70-85 = mild	Evaluative	Research tool	Norm	2000 3-35 yrs
		static and dynamic	yrs	55-69 = moderate				from the USA
		balance		<55 = severe disability *				(1970's)
NSMDA ¹⁶	Gross Motor, Fine	Sitting, kneeling, walking,	1 mth – 6	Total score 6-8 normal, 9-	Evaluative	Predictive,	Criterion	N/A
	Motor, Neurological,	balance, running,	yrs	11 minimal, 12-14 mild, 15-	Discriminative	Research tool		
	Postural	hopping, jumping,		19 moderate, 20-25				
	Development, Infant	catching, motor planning		severe, >25 profound				
	Patterns of			disability *				
	Movement, Sensory							
17	Motor. †							
PDMS-2 1	Gross motor, fine	Stationary (standing	Birth – 5	*	Discriminative	Predictive,	Norm	2003 USA and
	motor	balance, sit-ups, push-	yrs		Evaluative	Research tool		Canada (1997-8)
		ups), locomotion						
		(walking, running,						
		jumping, hopping, etc.),						
		object manipulation						
		(kick, throw, hit, catch)						

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 Sca Children 2 nd edition; Motor Scales 2 nd edi clinical reasoning; †,	ale of Infant and Toddler ¹⁴ MAND, McCarron Asse tion; ¹⁷ TGMD-II, Test of requires some manual h	Development 3 rd edition; ¹² BOT-2, essment of Neuromuscular Develop Gross Motor Development 2 nd edit handling; USA, United States of Ame	Bruininks-Oseretsky ment; ¹⁵ NSMDA, Nei on; ¹⁹ NDI, Neurodev rica	Test of Motor Proficie urological Sensory Me elopmental Index; SD	ency 2 nd edition; ¹³ MABC otor Developmental Asse , Standard Deviation; mt	⁻² , Movement Assess ssment; ¹⁶ PDMS-2, P h, month; yrs, years *	ment Battery fo eabody Develop , Advisable to u	or omental ise
9								
		For peer review only - ht	p://bmjopen.bm	j.com/site/about/	guidelines.xhtml			

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 Bruininks¹³ performed comprehensive surveys, pilot, tryout and standardisation studies before finalising the BOT-2, providing the most comprehensively reported content validity.

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

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.

.rent va s only weakly .report correlations . .ychomotor/intelligence te:

Table 2. Clinical Utility of Gross Motor Assessment Tools

Assessment Tool	Time to	Test Procedure	Target Examiner population	Training	Equipment/Manual
	administer (min)				
Bayley-III ¹²	30-90	Therapist administers in	Paediatric health professionals	Formal training not	Comprehensive manual/kit: £1089
		standardised order	early childhood specialists	required. DVD, webinars	Test kit provides most equipment
				and workshops available	
BOT-2 ¹³	40-60	Therapist administered in	Paediatric health professionals	Formal training not	Comprehensive manual/kit: £961
		standardised order	early childhood specialists	required	Test kit provides most equipment
MABC-2 ¹⁴	20-40	Therapist administers items in	Research psychologists, OT, PT,	Formal training not	Comprehensive manual/ kit: £1191
		standardised order. Some	Paediatricians	required.	Test kit provides most equipment
		flexibility allowed.			
MAND ¹⁵	15-20	Therapist administers items in	Professionals e.g. education,	Formal training not	Manual and test kit: £1366 includes
		standardised order.	neurology, OT, PT, psychology etc.	required.	equipment
NSMDA 16	20-45	Observation followed by	PT, OT	Formal training not	Comprehensive manual: £35.
		therapist administration of test		required (but is available)	Equipment not included
		items.			
PDMS-2 17	45-60 (20-30 for	Standardised procedure.	Paediatric health professionals, PE	Formal training not	Comprehensive manual/kit: £553
	GM only)		teachers, early intervention	required	Includes some but not all equipment
			specialists		required
TGMD-2 ¹⁹	15-20	Standardised procedure.	Teachers, health professionals (OT,	Formal training not	Kit includes manual and record form: £128.
			PT, doctors)	required	Equipment not included
		10		ad 10	

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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 3.	Methodologi	cal quality	of included	l articles
----------	-------------	-------------	-------------	------------

Test	First author, Year	Country	Population	Internal	Reliability	Measurement	Content	Structural	Hypothesis	Cross-	Criterion	Responsive -
			(Age,	consistency		error	validity	validity	testing	cultural	validity	ness
			Diagnosis)							validity		
BAYLEY III	Bayley 12	USA	1-42 mths	Fair	Fair	Good	Excellent	Good	Good	-	Good	-
	Spittle, et al. ⁴	Australia	2,4 yrs, Ex	-	-	-	-	-	-	-	Good	-
			prem									
	Visser, et al. 23	Netherlands	2.2-10.8 yrs,	-	-	-	Excellent	Poor	-	-	-	-
			GDD, L.I.									
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	USA	4-21 yrs	Good	Fair (interrater)	Good	Excellent	Good	-	-	Good	-
	Bruininks ¹³				Fair (test-retest)							
MABC-2	Ellinoudis, et al. 26	Greece	3-5.5 yrs	Excellent	Good	-	-	-	-	-	-	-
(AB 1)	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	Belgium	3-4 yrs	Poor	Poor	Poor	-	-	-	-	-	-
	al. 27											
MABC-2	Holm, et al. ²⁸	Norway	7-9 yrs	-	Fair (interrater)	Poor	-	-	-	-	-	-
(AB 2)					Poor (intrarater)							
	Kita, et al. ²⁹	Japan	7-10 yrs	Excellent	-	-	-	-/1	-	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	-	-	-	-	-	-	-	Poor	-	-
		+ Belgium										
	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	-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	Wuang, et al. 23	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	-
ISMDA	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	-	9/2	-	-	Poor	-	-	Fair	-
DMS-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	Fai
	Folio and Fewell ¹⁷	USA	0-71 mnths	Good	-	Poor	Excellent	Good	Good	-	Poor	-
GMD-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	-	-	<u> </u>	-	-	-	-
	Kim, et al. ⁴³	Korea	5-6 yrs	Poor	Fair	-	-	Poor	-	-	Poor	-
	Logan, et al. ⁵	USA	3-6 yrs	-	-	-	-	5	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	-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 15 of 40

BMJ Open

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

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;¹² BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2rd edition;¹³ MABC-2, Movement Assessment Battery for Children 2rd edition;¹⁷ TGMD-11, Test of Gross Motor Development 2rd edition;¹⁹ Mths, Month; yrs, years; DCD, Developmental Coordination Disorder; VI, Vision Impairment; ID, Intellectual Disability; GDD, global developmental delay; LI, Language Impairment; ELBW, Extremely Low Birth Weight; VLBW, Very Low Birth Weight; CP, Cerebral Palsy; prem, premature; USA, United States of America

1 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.

28 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
 16

Page 17 of 40

BMJ Open

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 times²⁷.

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 MABC¹⁴, 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 environment². Clinicians should utilise appropriate assessments or questionnaires to ensure that these domains of health are also addressed in line with World Health Organisation guidelines².

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 poor¹⁰. 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

Page 19 of 40

BMJ Open

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

3 clearer assessment of their responsiveness to change to assist clinicians and researchers with

4 outcome measure selection. Researchers should be mindful of the methods they use to assess

5 validity and reliability. Clarity of reporting, statistical methods and sample sizes should be carefully

6 considered to ensure the highest quality of evidence.

7 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.

16 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

BMJ Open

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.
- 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]
- 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.
- 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 to9 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

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-yearold children in Bangladesh. *Pediatrics* 2013;131(2):e486-94. doi: <u>http://dx.doi.org/10.1542/peds.2011-2421</u>
- 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: <u>http://dx.doi.org/10.1016/j.ridd.2011.01.031</u>
- 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.

Page 23 of 40	BMJ Open
	30. Griffiths A, Morgan P, Anderson PJ, et al. Predictive value of the Movement Assessment Battery for
1 2	Children - Second Edition at 4 years, for motor impairment at 8 years in children born preterm. Dev Med
3	Child Neurol 2017;59(5):490-96. doi: 10.1111/dmcn.13367 [published Online First: 2017/01/10]
4 5	31. Schulz J, Henderson SE, Sugden DA, et al. Structural validity of the Movement ABC-2 test: factor structure
6 7	comparisons across three age groups. Res Dev Disabil 2011;32(4):1361-9. doi:
8	http://dx.doi.org/10.1016/j.ridd.2011.01.032
9 10	32. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation,
11	reliability, and validity for Brazilian children. Res Dev Disabil 2014;35(3):733-40. doi:
12 13	http://dx.doi.org/10.1016/j.ridd.2013.10.028
14	33. Wuang YP, Su JH, Su CY. Reliability and responsiveness of the Movement Assessment Battery for Children-
16	Second Edition Test in children with developmental coordination disorder. Dev Med Child Neurol
17 19	2012;54(2):160-5. doi: http://dx.doi.org/10.1111/j.1469-8749.2011.04177.x
19	34. Hands B, Larkin D, Rose E. The psychometric properties of the McCarron Assessment of Neuromuscular
20 21	Development as a longitudinal measure with Australian youth. Hum Mov Sci 2013;32(3):485-97. doi:
22	http://dx.doi.org/10.1016/j.humov.2013.02.007
23 24	35. Danks M, Maideen MF, Burns YR, et al. The long-term predictive validity of early motor development in
25	"apparently normal" ELBW survivors. <i>Early Hum Dev</i> 2012;88(8):637-41.
20	36. MacDonald J, Burns Y. Performance on the NSMDA During the First and Second Year of Life to Predict
28 29	Functional Ability at the Age Of 4 in Children with Cerebral Palsy. Hong Kong Physiotherapy Journal
30	2005;23(1):40-45. doi: 10.1016/S1013-7025(09)70058-2
31 32	37. Burns YR, Ensbey RM, Norrie MA. The Neuro-sensory motor developmental assessment part 1:
33	development and administration of the test. Aust J Physiother 1989;35(3):141-49.
34 35	38. Burns YR, Ensbey RM, Norrie MA. The neuro-sensory motor developmental assessment part II: predictive
36	and concurrent validity. Aust J Physiother 1989;35(3):151-57.
38	39. Barnett LM, Minto C, Lander N, et al. Interrater reliability assessment using the Test of Gross Motor
39 40	Development-2. Journal of Science and Medicine in Sport 2014;17(6):667-70. doi:
41	http://dx.doi.org/10.1016/j.jsams.2013.09.013
42 43	40. Farrokhi A, Zareh Zadeh M, Karimi Alvar L, et al. Reliability and validity of test of gross motor development-
44	2 (Ulrich, 2000) among 3-10 aged children of Tehran City. Journal of Physical Education and Sports
45 46	Management 2014;5(2):18-28. doi: 10.5897/JPESM12.003
47	41. Houwen S, Hartman E, Jonker L, et al. Reliability and validity of the TGMD-2 in primary-school-age children
48	with visual impairments. Adapt Phys Act Q 2010;27(2):143-59.
50 51	42. Kim Y, Park I, Kang M. Examining rater effects of the TGMD-2 on children with intellectual disability. Adapt
52	<i>Phys Act Q</i> 2012;29(4):346-65.
53 54	43. Kim CI, Han DW, Park IH. Reliability and validity of the test of gross motor development-II in Korean
55	preschool children: applying AHP. Res Dev Disabil 2014;35(4):800-7. doi:
50 57	http://dx.doi.org/10.1016/j.ridd.2014.01.019
58 ₂₃	
60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

- 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: <u>http://dx.doi.org/10.1080/00222895.2012.700967</u>
 - 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.







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)



2	
3	
4	
5	
6	
7	
8	
9	
10	
10	
11	
12	
15	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
30	
40	
40 //1	
41	
42 //2	
75 77	
77 15	
45 46	
40	
47 10	
40	
49 50	
50	
51	
52	
53	
54	
55	
56	
57	
58	
59	

1

1

Supplementary table 1: Definition of terms

		Definition	Example/explanation
Validity	Content	The degree to which an assessment	Concerned with the relevance
		tool's content measures the construct	and comprehensiveness of the
		that it intends to measure [7]	items included in the assessment
			tool
	Construct	Measures the degree to which the	Examples include structural
		scores obtained from the test are an	validity, hypothesis testing and
		adequate reflection of the construct to	cross-cultural validity
		be measured [7]	
	Criterion	Assesses whether or not the test scores	As there is no gold standard of
		reflect a 'gold standard' assessment [7]	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	Usually measured with intraclass
		regardless of the time between	correlation coefficient (ICC), but
		assessments (test-retest) or the person	can be measured using Cohen's
		administrating (intra and inter-rater)	kappa coefficient. Percentage
		[27]	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	The degree of interrelatedness of an	Usually measured using
	consistency	assessment tool's items [7]	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	Refers to th	ne error obtained between	May be systematic or random
		Error	moscurom	ants that cannot be	orror [7]
		LITOI	measureme		
			attributed t	to the patients true change	
			[7]		
Resi	onsiveness		An assessm	ent tool's ability to detect	This is central to a tools capacity
			cnange ove	r time in the construct to be	to be used as an outcome
			measured [7]	measure.
Sur	plementar	v table 2: Excl	luded Asse	ssment Tools	
Rea	son	,		Assessments	
NCu.		his in Eastich			
Ivian	iual not availa	able in English		Maastricht's Motor Test (Mi	
				The Motor-Proficiency-Test	for children between 4 and 6 year
				of age (MOT 4-6)	
				Zuk Assessment	
				Körperkoordinationtest für H	Kinder (KTK)
Can	not extract m	eaningful gross r	notor score	Early Intervention Developm	nental Profile (EIDP)
				Neurological Developmental	l Exam
				Preschooler Gross Motor Ou	uality Scale (PGMO)
				The Malawi Developmental	Assessment Tool (MDAT)
				Dutch table tennis motor ski	ills assessment
Scro	oning Tool			Brief Assessment of Motor F	Sunction (RAME)
Scie					
				The Motor Performance Che	ecklist
				Motor skill checklist (MSC)	
Diag	nosis specific	c/requires a diagr	nosis	Assessment Battery for the	Atypical Handicapped Child (VAB)
				Video-based documentation	and rating system of the motor
				behaviour of handicapped cl	hildren
Only	assesses on	e motor domain ((e.g. gait)	Standardized Walking Obsta	cle Course (SWOC)
				Timed floor to stand test	
Mar	ual not publi	shed/commercia	llv available	Rapid Neurodevelopmental	Assessment (RNDA)
iiiai			ing available	Tufts Assassment of Motor I	Porformanco (TAMP)
				TUILS ASSESSMENT OF WOLDER	
				7 1 1 1 1 1	(7)(4)
				Zurich Neuromotor Assessm	ent (ZNA)

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

PDMS-2 [31]	GM: Stationary (30 items), locomotion (89 items),	Raw scores, Age equivalent, centile rank. Standard	Motor activities program (intervention ideas
	object manipulation (24 item). FM: grasping(26	scores (subtests) Composite quotient. Complex	
	items), visual-motor integration (72 items)	conversions.	
TGMD-2 [25]	Locomotion (6 items) and Object Control (6 items).	Raw scores, standard scores, percentile rank, age	Simple to administer
	Separate male/female norms for object control	equivalent, Gross Motor Quotient. Simple	
	subset	conversion.	
Bayley-III, Bayley	Scale of Infant and Toddler Development 3 rd edition [28]; BO)T-2, Bruininks-Oseretsky Test of Motor Proficiency 2 nd editio	n [13]; MABC-2, Movement Assessment Battery for
Children 2 nd editi	on [24]; MAND, McCarron Assessment of Neuromuscular De	velopment [29]; NSMDA, Neurological Sensory Motor Develo	pmental Assessment [30]; PDMS-2, Peabody
Developmental M	Notor Scales 2 nd edition [31];; TGMD-II, Test of Gross Motor D	Development 2 nd edition [25]; GM, Gross Motor; FM, Fine Mo	tor; NDI, Neurodevelopmental Index
4			

Supplementary table 4: Content and construct validity of assessment tools

Test	Content	Construct
BAYLEY	Expert opinion for standard and low verbal version [28, 34].	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language
ш	Literature reviews. Gross motor score correlated with Motor	impairment [28]. H_i (gross motor subset) = 0.52-0.97 for children with language impairment and 0.82-
	component 0.70 [28]	0.99 in control group [34]
BOT-2	Focus groups, product survey, pilot, national tryout and	Factor analysis, scores increase with age, discriminates between normal and children with DCD (N=50),
	standardisation studies, professional reviews[13]	high-functioning ASD ($N = 45$) and mild-moderate ID ($N = 66$) [13]
MABC-	Expert Panel, Stakeholder feedback, Literature review [18]	Factor analysis, correlation coefficients [36] Subtest correlations 0.65-0.76 <i>p</i> <0.001. Discriminates
2		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 (η 2 =
	Expert panel - clarity (validity content index 71.8-93.9. Kappa	0.63) p < 0.0001 [40]. UK norms not appropriate to use with Dutch/Flemish children as under/over-
	0.76-0.88) and pertinence (98.5-99.3 and kappa 0.83-0.92)	estimate risk of motor impairment [15]. In Chinese population: CFA initially rejected. Acceptable fit
	<i>p</i> <0.001 [40]	achieved after 2 items removed [14]. Age band 2 shows good validity in Japanese population [37].
	p k - k	
MAND	Based on neuropsychological theory. Several rounds of	Factor analysis [29] [42]. Scores increase with age, and discriminate between typically developing
	revision/trials of tasks during development [29]	children and those with head trauma or neurological dysfunction as well as gender [29] [42]
NSMDA	Literature review. Developed by an experienced paediatric	Factor analysis (up to 2 years of age) [45] [46]. Stability of test results over time (up to 2 years) [45] [46].
	physiotherapist [45]	
PDMS-	Literature review. Created by experts in the field. Revised with	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with
2	feedback from therapists guided revision. Hierarchical sequence	age (r=0.80-0.93) [31]
5		
	For peer review only - h	http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

	of items [31]	
TGMD-	Expert Panel (3 PE teachers with post-grad qualifications) [25].	Exploratory and confirmatory factor analysis [50] [52] [25] [22] [53] [51] High and significant correlation
2	Translated version (Brazilian Portuguese) language clarity 0.96,	of increasing age and increasing scores [47]. Age and disability differentiation [25] [51] Subtest
	pertinence >0.89. Experts CVI for clarity and pertinence were also	correlation 0.41 [25]
	strong- α = 0.93 clarity and α =0.91 pertinence [52]	
		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 3 rd edition;[28] BOT	T-2, Bruininks-Oseretsky Test of Motor Proficiency 2 nd edition;[13] MABC-2, Movement Assessment Battery fo
Children 2	2 nd edition;[24] MAND, McCarron Assessment of Neuromuscular Dev	elopment;[29] NSMDA, Neurological Sensory Motor Developmental Assessment;[30] PDMS-2, Peabody
Developn	nental Motor Scales 2 nd edition;[31] TGMD-II, Test of Gross Motor De	velopment 2 nd edition;[25]; H _i , scalability coefficient; CFA, Confirmatory Factor Analysis; TD, Typically
Developir	ng; ASD, Autism Spectrum Disorder, ID, Intellectual Disability; WPPSI,	Wechsler Preschool and Primary Scale of Intelligence; WISC-R, Wechsler Preschool and Primary Scale of
Cupala	menter stable - Criterian and avadiative validity of	
Supple Test	mentary table 5: Criterion and predictive validity of Concurrent/criterion	assessment tools Predictive
Supple Test	ementary table 5: Criterion and predictive validity of Concurrent/criterion	assessment tools Predictive Motor impairment at 4 years: Bayley III at 2 years <15D = sensitivity 0.32-0
Supple Test BAYLEY III	ementary table 5: Criterion and predictive validity of Concurrent/criterion Given but mean age <22 months. Not relevant to stuc	F assessment tools Predictive dy population. [28] Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.320 specificity 0.97 <2SD sensitivity 0.18-0.21 specificity 1.00.
Supple Test BAYLEY III 6	ementary table 5: Criterion and predictive validity of Concurrent/criterion Given but mean age <22 months. Not relevant to stuc	assessment tools Predictive dy population. [28] Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.320 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. <2SI
		sensitivity 0.67 specificity 1.0 [4]
BOT-2	MABC-2 <i>p</i> = 0.92 PDMS-2 <i>p</i> = 0.88 (<i>N</i> = 38) [17]. PDMS-2 Total motor composite <i>r</i> = 0.77 [13].	- -
		Classification groups (DCD, at visit and TD) remained some quarting (C
	PDIVIS-2 $p = 0.051 - 0.04$ [17] [14]. TOIND-2 $p = 0.45$ [5]. TOIND-2 Standard scores ($r = 0.5, p < 0.02$)	Classification groups (DCD, at fisk and TD) remained same over time (6 months) $y_2 = 0.67$ n = 0.72 [40]. Predictive of motor impoirment over 6.12
	[40]. BO1-2 $p = 0.90 - 0.92$ [17].	months) $\chi_2 = 0.67 p = 0.72$ [40]. Predictive of motor impairment over 6-1.
		months (N =41) ICC 0.88 p < 0.007 [40]. Scores at 4 years predictive of moto
		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%, specificit
		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 <i>ρ</i> = 0.63- 0.84, [14, 17] MABC-2 gross motor composite <i>ρ</i> = 0.743 [14]	
	BOT-2 ρ = 0.88 [17]. Mullen Scales of Early Learning GMQ = 0.86 FMQ = 0.80 [31]	
7		
/		
	For peer review only - http://bmjopen.bmj.com/site	e/about/guidelines.xhtml
BMJ Open

Rayley-III			or 0.63 object control			
Bayloy-III		0.41 [25]				
Dayley-III	, Bayley Scale of Infant and Toddler De	evelopment 3 rd edition;[28] BOT-2, Bruir	ninks-Oseretsky Test o	f Motor Proficiency 2 ⁿ	^d edition;[13] MABC-2, Movement A	ssessment Battery for
Children	2 nd edition;[24] MAND, McCarron Asse	essment of Neuromuscular Developmen	nt;[29] NSMDA, Neuro	logical Sensory Motor	Developmental Assessment;[30] PD	MS-2, Peabody
Developr	nental Motor Scales 2 nd edition;[31] TG	GMD-II, Test of Gross Motor Developme	ent 2 nd edition;[25] NS,	, Not Specified; SD, Sta	indard Deviation; CP, Cerebral Palsy;	; TD, Typically
Developi	ng; ICC, Intraclass Correlation Coefficie	ent; χ2, Chi Squared; NDI, Neurodevelop	mental Index; CSSA, (Comprehensive Scales	of Student Abilities	
Supple	mentary table 6: Reliability c	of assessment tools				
Supple	ementary table 6: Reliability c	of assessment tools Test-Retest	Intra-rater	Inter-rater	Minimal detectable change	Minimal clinical
Supple Test	ementary table 6: Reliability c Internal Consistency	of assessment tools Test-Retest	Intra-rater	Inter-rater	Minimal detectable change	Minimal clinical important difference
Supple Test BAYLEY	ementary table 6: Reliability c Internal Consistency GM α = 0.87-0.93 MC: α 0.90-0.96	of assessment tools Test-Retest Gross Motor subtest (N=47)	Intra-rater	Inter-rater	Minimal detectable change SEM Gross motor subtest	Minimal clinical important difference
Supple Test BAYLEY III	ementary table 6: Reliability c Internal Consistency GM α = 0.87-0.93 MC: α 0.90-0.96 (24-42 months) [28]	of assessment tools Test-Retest Gross Motor subtest (N=47) r=0.79 Motor component r=0.80	Intra-rater -	Inter-rater	Minimal detectable change SEM Gross motor subtest 0.85-1.08. of Motor	Minimal clinical important differenc

					42 months) [28]	
BOT-2	(<i>N</i> = 100) <i>α</i> = 0.92 [35]	(<i>N</i> = 100) ICC = 0.99 [35] (<i>N</i> =	-	Total motor composite	4.18 (sensitivity 55.10%	6.53 (sensitivity 48.989
		141) ICC = 0.97 [17] 4-7 yrs (N =		4-21 yrs (<i>N</i> = 47) <i>r</i> =	specificity 72.55%) [35] 7.43	specificity 76.47%) [35
	$(N = 141) \alpha = 0.86 [17]$	43) r = 0.81 (8-12 yrs (N= 44) r =		0.98 [13]	(sensitivity 42.49% specificity	6.55 (sensitivity 49.99
	4-7 yrs (N= 620) α = 0.95 8-11 yrs (N=	0.80 [13]			65.72%) [17]	specificity 58.78%) [17
	450) <i>α</i> = 0.95 [13]					
MABC-	$(N = 60)$ M.D $\alpha = 0.51$, A&C $\alpha = 0.70$,	(N=60) ICC = 0.85 [36] Item	(<i>N</i> =28) <i>κ</i> = 0.71	Item ICC's range 0.892-	(N=28) Intrarater MDC =	-
2 (AB 1)	Bal α = 0.66 [36] (N = 1823) α = 0.502	ICC's 0.830-0.985 [14] ICC test-	[23]	0.998 [14] (<i>N</i> =22) <i>κ</i> =	3.43	
	[14] (<i>N</i> =50) <i>α</i> = 0.81-0.87 [23]	retest = 0.83 [23] Inter-rater		0.60 [23]		
		test-retest ICC = 0.79 [23]			(N=22) Inter-tester MDC =	
					3.81 [23]	
MABC-	Translated version (Japanese)	-	ICC = 0.64 [18]	ICC 0.63 [18]	Intra-rater SDC TTS: +/- 11.7	-
2 (AB 2)	(N=132) <i>α</i> = 0.602 [37]				TSS +/- 3.3. Inter-rater SDC	
					TTS +/-16.0 TSS +/- 3.8 [18]	
MABC-	Subscales α = 0.78 (M.D = 0.77, BS =	N=60 (all 3 age bands) r=0.80	ICC 0.88 [40]	ICC 0.96-0.99 [40]	SEM 1.34 (95%CI) = 3 [24]	1.39 (sensitivity 72.47%
2	0.52, Bal = 0.77) [40] α = 0.88 [41]	[24]			1.83 (95%Cl) [41] 1.83	specificity 46.18%) [17 41]
	$(N = 141) \alpha = 0.88 [17]$	<i>r</i> =0.74 p<0.0001 (standard			(sensitivity 69.69% specificity	
		score). ICC standard score = 0.85			52.10%) [17]	
		[40]				
		ICC 0.96 [41]				
		<i>N</i> = 141 ICC =0.96 [17]				
9						
		- · · · · · · · · · · · · · · · · · · ·			Let 1	

MAND	-	-	-	-	-	-
NSMDA	Cross correlation matrix Item scoring (12+24months) 0.73 <i>p</i> <0.001, Functional grade (12+24months) 0.87 <i>p</i> <0.001 [45]		-	-	-	-
PDMS- 2	$(N=141) \alpha=0.89 [17] 24-35m \alpha=0.97,$ 36-47m $\alpha=0.95, 48-59m \alpha=0.97, 60-71m \alpha= 0.98.$ For subgroups† $\alpha=0.99$ [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%) [1]
TGMD- 2	$(N=1438) \alpha=0.80 [47] N=75$ Locomotor subset $\alpha=0.71$ object control $\alpha=0.72 [22] N=120 \alpha = 0.72$ $[49] N= 99 \alpha = 0.90 [51] N = 1208$ Cronbach's $\alpha = 0.91$ (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-I edition; edition; Aiming a Locomo	II, Bayley Scale of Infant and Toddler Develo [24] MAND, McCarron Assessment of Neuro [31] TGMD-II, Test of Gross Motor Developn and catching; SDC, Smallest Detectable Char tion Subset	pment 3 rd edition;[28] BOT-2, Bruinin omuscular Development;[29] NSMDA, nent 2 nd edition;[25] GM, Gross Moto nge; TTS, Total Test Score; TSS, Total S	ks-Oseretsky Test of Mor Neurological Sensory Mo r Subset; MC, Motor Con tandard Score; †, gender	tor Proficiency 2 nd edition;[1 otor Developmental Assessm nponent; <i>K</i> , Kappa Coefficien [.] , ethnicity, speech/language	3] MABC-2, Movement Assessmen hent;[30] PDMS-2, Peabody Develo t; M.D, Manual Dexterity; BS, Ball or physical disorder; Obj, Object O	t Battery for Children 2 nd opmental Motor Scales 2 nd Skills; BAL, Balance; A&C, Control Subset; L.S,



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
	·		
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
s 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
4 5		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	•

Page 39 of 40

BMJ Open



3

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.	-
		Page 1 of 2	•
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]).	-
5 6 7	1	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	<u>I</u>



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. 18 doi:10.1371/journal.pmed1000097

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

Page 2 of 2

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

SCHOLARONE[™] Manuscripts

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

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

Alison Griffiths^{1, 2, 3}, Rachel Toovey^{3, 4}, Prue E. Morgan¹, 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

1 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.

7 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. Testretest 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 ans.

1 Introduction

Motor function promotes cognitive and perceptual development in children and contributes to their ability to participate in their home, school and community environments¹. Motor impairment can negatively affect activity and participation levels of children², which may lead to lower levels of physical activity, fitness and health into adulthood³. 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 peers³. 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 children⁴. Further, assessment tools are essential to monitor the effects of interventions⁴. 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.

BMJ Open

A systematic review in 2010 by Slater⁸ 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 Lalor⁹ 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. 2.

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 \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; \geq 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 responsiveness⁷ (Supplementary Table 5). Cross-cultural validity, structural validity and hypothesis testing are all considered to be a component of construct validity⁷. Whilst predictive validity is considered to be a component of content validity, it is reported on separately in this paper for interpretability of results⁷. 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.

BMJ Open

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.

7 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	Age	Diagnostic criteria	Primary	Secondary	Type of test	Normative
1001		components tested	Talige		purpose	pulpose		sample (year)
Bayley-III	Gross motor, fine	Static postures, dynamic	1 mth – 3	Developmental delay:	Discriminative	Predictive,	Norm	1700 children
	motor, cognitive,	movement, balance	yrs	<25th centile or below		Evaluative,		from the USA
	communication,			2SD. *		Research tool		(2000)
	social/emotional,							
12	adaptive							
BOT-2	Gross motor, fine	Coordination, balance,	4 – 21 yrs	*	Discriminative	Research tool	Norm	1520 children
	motor	running speed and			Evaluative			from the USA
14		agility, strength						(2005)
MABC-2 ¹⁴	Gross motor, fine	Aiming and catching,	3 – 16 yrs	Traffic light system: Green	Discriminative	Intervention	Norm	1172 children
	motor, balance	static and dynamic		= normal, amber = 'at risk'	Evaluative	planning,		from United
		balance		and red = definite motor		Research tool		Kingdom (2006)
				impairment (<15%). *				
MAND ¹⁵	Gross and fine motor	Coordination, jumping,	3 yrs – 25	NDI 70-85 = mild	Evaluative	Research tool	Norm	2000 3-35 yrs
		static and dynamic	yrs	55-69 = moderate				from the USA
		balance		<55 = severe disability *				(1970's)
NSMDA 16	Gross Motor, Fine	Sitting, kneeling, walking,	1 mth – 6	Total score 6-8 normal, 9-	Evaluative	Predictive,	Criterion	N/A
	Motor, Neurological,	balance, running,	yrs	11 minimal, 12-14 mild, 15-	Discriminative	Research tool		
	Postural	hopping, jumping,		19 moderate, 20-25				
	Development, Infant	catching, motor planning		severe, >25 profound				
	Patterns of			disability *				
	Movement, Sensory							
	Motor. †							
PDMS-2 17	Gross motor, fine	Stationary (standing	Birth – 5	*	Discriminative	Predictive,	Norm	2003 USA and
	motor	balance, sit-ups, push-	yrs		Evaluative	Research tool		Canada (1997-8)
		ups), locomotion	-					
		(walking, running,						
		jumping, hopping, etc.),						
		object manipulation						
		(kick, throw, hit, catch)						

TGMD-2 ¹⁸	Gross Motor	Locomotion (run, gallop,	3 – 10 yrs	*	Discriminative	Outcome	Norm	1208 USA
		hop, leap, jump, slide)			Evaluative	measure,		children (199
		and Object control				research tool,		1998)
		(batting, dribbling, catch,				intervention		
		kick, throw, roll)				planning		
hildren 2 nd edition; ¹⁴ lotor Scales 2 nd editio inical reasoning; †, ro	MAND, McCarron Asse on; ¹⁷ TGMD-II, Test of equires some manual h	Severophient 3 edition, BOT-2, essment of Neuromuscular Develo Gross Motor Development 2 nd edit andling; USA, United States of Am	bi unfinitis-Oser etsk pment; ¹⁵ NSMDA, N ion; ¹⁸ NDI, Neurode erica	eurological Sensory M evelopmental Index; Si	lotor Developmental Asse D, Standard Deviation; mtl	ssment; ¹⁶ PDMS-2, Pe	, Advisable to u	pmental ise
0								
		For peer review only - ht	tp://bmiopen.br	ni.com/site/about	/quidelines xhtml			

Page 11 of 45

BMJ Open

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.

14 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).

24 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.

1 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 Bruininks¹³ 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-2⁵ (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	Test Procedure	Target Examiner population	Training	Equipment/Manual
	administer (min)				
Bayley-III ¹²	30-90	Therapist administers in	Paediatric health professionals	Formal training not	Comprehensive manual/kit: £1089
		standardised order	early childhood specialists	required. DVD, webinars	Test kit provides most equipment
				and workshops available	
BOT-2 ¹³	40-60	Therapist administered in	Paediatric health professionals	Formal training not	Comprehensive manual/kit: £961
		standardised order	early childhood specialists	required	Test kit provides most equipment
MABC-2 ¹⁴	20-40	Therapist administers items in	Research psychologists, OT, PT,	Formal training not	Comprehensive manual/ kit: £1191
		standardised order. Some	Paediatricians	required.	Test kit provides most equipment
		flexibility allowed.			
MAND 15	15-20	Therapist administers items in	Professionals e.g. education,	Formal training not	Manual and test kit: £1366 includes
		standardised order.	neurology, OT, PT, psychology etc.	required.	equipment
NSMDA ¹⁶	20-45	Observation followed by	PT, OT	Formal training not	Comprehensive manual: £35.
		therapist administration of test		required (but is available)	Equipment not included
		items.			
PDMS-2 17	45-60 (20-30 for	Standardised procedure.	Paediatric health professionals, PE	Formal training not	Comprehensive manual/kit: £553
	GM only)		teachers, early intervention	required	Includes some but not all equipment
			specialists		required
TGMD-2 ¹⁸	15-20	Standardised procedure.	Teachers, health professionals (OT,	Formal training not	Kit includes manual and record form: £128.
			PT, doctors)	required	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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 3. Methodological quality of included articles

Test	First author, Year	Country	Population	Internal	Reliability	Measurement	Content	Structural	Hypothesis	Cross-	Criterion	Responsive -
			(Age,	consistency		error	validity	validity	testing	cultural	validity	ness
			Diagnosis)							validity		
BAYLEY III	Bayley 12	USA	1-42 mths	Fair	Fair	Good	Excellent	Good	Good	-	Good	-
-	Spittle, et al. ⁴	Australia	2,4 yrs, Ex	-	-	-	-	-	-	-	Good	-
			prem									
	Visser, et al. ²⁶	Netherlands	2.2-10.8 yrs,	-	-	-	Excellent	Poor	-	-	-	-
			GDD, L.I.									
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	USA	4-21 yrs	Good	Fair (interrater)	Good	Excellent	Good	-	-	Good	-
	Bruininks ¹³				Fair (test-retest)							
MABC-2	Ellinoudis, et al. 29	Greece	3-5.5 yrs	Excellent	Good	-	-	-	-	-	-	-
(AB 1)	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	Belgium	3-4 yrs	Poor	Poor	Poor	-	-	-	-	-	-
	al. ³⁰											
MABC-2	Holm, et al. ³¹	Norway	7-9 yrs	-	Fair (interrater)	Poor		-	-	-	-	-
(AB 2)					Poor (intrarater)							
	Kita, et al. ³²	Japan	7-10 yrs	Excellent	-	-	-	-/1	-	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	-	-	-	-	-	-	-	Poor	-	-
		+ Belgium										
	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	-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 15 of 45

BMJ Open

	Wuang, et al. ²⁸	Taiwan	3-6 yrs, ID	Fair	Good	Good	-	-	-	-	Good	Fa
	Wuang, et al. ³⁶	Taiwan	6-12 yrs DCD	Poor	Fair	Good	-	-	-	-	-	Fa
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	-	9/	-	-	Poor	-	-	Fair	-
DMS-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	Fa
	Folio and Fewell 17	USA	0-71 mths	Good	-	Poor	Excellent	Good	Good	-	Poor	-
GMD-2	Barnett, et al. ⁴²	Australia	4-8 yrs	-	Fair		<u></u>	-	-	-	-	-
	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	-	. •	D 1	-	-	-	-
	Kim, et al. ⁴⁵	Korea	5-6 yrs	Poor	Fair	-	-	Poor	-	-	Poor	-
	Logan, et al. ⁵	USA	3-6 yrs	-	-	-	-	- 5	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	-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;¹² BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2rd edition;¹³ MABC-2, Movement Assessment Battery for Children 2rd edition;¹⁴ MAD, McCarron Assessment of Neuromuscular Developments¹⁴ NSMDA, Neurological Sensory Motor Developmental Assessment, ¹⁴ PDMS-2, Peabody Developmental Motor Scales 2rd edition;¹⁷; TGMD-II, Test of Gross Motor Development 2rd edition;¹⁸ Mths, Months; yrs, years; DCD, Developmental Coordination Disorder; VI, Vision Impairment; ID, Intellectual Disability; GDD, global developmental delay; L1, Language Impairment; ELBW, Extremely Low Birth Weight; VLBW, Very Low Birth Weight; CP, Cerebral Palsy; prem, premature; USA, United States of America



1 2	1	Reliability
3 ∡	2	Internal consistency of assessments are summarised in Table 6. The BOT-2's high internal
5	3	consistency is well supported, including for children with an intellectual disability ^{28 46} . The MABC-2
6 7	4	appears to have lower internal consistency than the BOT-2, which may relate to the limited number
8 9	5	of test items (eight) on the MABC-2. The highest values for internal consistency for the MABC-2
10	6	were obtained in specific populations (intellectual disability and developmental coordination
12	7	disorder) with poor to fair methodology only. Conversely the highest quality articles reported the
13 14	8	lowest values, although it should be noted that these assessed age band 1 (3-6 years) only. Internal
15 16	9	consistency is reported to be high for the PDMS-2, while the Bayley-III is shown to have excellent
17	10	internal consistency in children aged 24-42 months The TGMD-2 is reported by two good quality
18 19	11	(and four poor to fair quality) articles to have excellent internal consistency, including for children
20 21	12	with vision impairment and intellectual disability. The MAND is the only assessment tool included in
22	13	this review without published data of internal consistency or reliability in this age group.
23 24	-	
25 26	14	The reliability findings are summarised in Table 6 and in Figures 2 and 3. Test-retest reliability was
27	15	excellent in the Bayley-III (Table 6), BOT-2 and PDMS-2; and was good to excellent in the MABC-2
28 29	16	and TGMD-2 (Figure 2). Intra-rater reliability was rarely investigated or reported for most tools, with
30 31 32	17	the TGMD-2 demonstrating better results than the MABC-2 (Figure 3). Only the TGMD-2 and
	18	MABC-2 report inter-rater reliability values using an ICC (Figure 3) ^{31 42} . Inter-rater reliability is also
33 34	19	supported in the BOT-2 with Pearson Correlation Coefficient and Kappa respectively. The studies
35 36	20	referred to in the test manuals for the TGMD-2, Bayley-III, BOT-2 and MABC-2 all report reliability
37	21	findings using Pearson's correlation, which is less ideal than an ICC or weighted kappa for statistical
38 39	22	analysis ^{47 48} . Only studies reporting ICC's are visually represented in Figures 2 (test-retest) and 3
40 41	23	(inter and intra-rater). The TGMD-2 test-retest reliability results from Houwen, et al. 21 were
42 42	24	believed to contain an error as the reported ICC was outside of the reported confidence intervals
43 44	25	(ICC 0.92, 0.82-0.91). This data set was therefore excluded from Figure 2.
45 46	-6	Perpensiveness was reported for the Paulov III, POT 2, MARC 2 and PDMS 2 with minimal
47 48	20	detectable change (MDC) or a standard error of measurement (SEM) ²⁸ Sensitivity and execisivity
49	27	detectable change (MDC) of a standard error of measurement (SEM) . Sensitivity and specificity
50 51	28	for detecting change was shown to be satisfactory in the MABC-2, PDMS-2 and MABC-2 (Table 6).
52 53	29	There have been no studies to date on the responsiveness of the TGMD-2, NSMDA or MAND.
54		
55 56		

Table 4: Content and c	construct validity of assessment tools
------------------------	--

Test	Content	Construct
BAYLEY	Expert opinion for standard and low verbal version ^{12 26} . Literature	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language impairment 12
III	reviews. Gross motor score correlated with Motor component 0.70 $^{\rm 12}$	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	Factor analysis, scores increase with age, discriminates between normal and children with DCD (N=50), high-
	standardisation studies, professional reviews ¹³	functioning ASD ($N = 45$) and mild-moderate ID ($N = 66$) ¹³
ИАВС-2	Expert Panel, Stakeholder feedback, Literature review ³¹	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 .
	Expert panel - clarity (validity content index 71.8-93.9, Kappa 0.76-	Significant difference between TD, DCD and at risk DCD scores (η 2 = 0.63) p < 0.0001 ³⁵ . UK norms not appropriate to
	0.88) and pertinence (98.5-99.3 and kappa 0.83-0.92) $ ho$ <0.001 35	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	Factor analysis ^{15 37} . Scores increase with age, and discriminate between typically developing children and those with
	of tasks during development ¹⁵	head trauma or neurological dysfunction as well as gender ^{15 37}
NSMDA	Literature review. Developed by an experienced paediatric	Factor analysis (up to 2 years of age) ^{40 41} . Stability of test results over time (up to 2 years) ^{40 41} .
	physiotherapist ⁴⁰	
DMS-2	Literature review. Created by experts in the field. Revised with	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with age (r=0.80-
	feedback from therapists guided revision. Hierarchical sequence of	0.93) 17
	items ¹⁷	
GMD-2	Expert Panel (3 PE teachers with post-grad qualifications) ¹⁸ .	Exploratory and confirmatory factor analysis ^{19 20 18 21 22 23} High and significant correlation of increasing age and
	Translated version (Brazilian Portuguese) language clarity 0.96,	increasing scores 43 . Age and disability differentiation $^{18\ 23}$ Subtest correlation 0.41 18
	pertinence >0.89. Experts CVI for clarity and pertinence were also	Galloping, running and leaping not well correlated with locomotion subscale. Object control significant & highly
	strong- α = 0.93 clarity and α =0.91 pertinence ²⁰	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 3 edition; BOT-2, Brui	ninks-Oseretsky lest of Motor Proficiency 2 edition; MABC-2, Movement Assessment Battery for Children 2
edition; ⁻ I	MAND, McCarron Assessment of Neuromuscular Development; NSMDA	, Neurological Sensory Motor Developmental Assessment; ²⁷ PDMS-2, Peabody Developmental Motor Scales 2 ²⁴ edition; ²
TGMD-II, T	Test of Gross Motor Development 2 nd edition; ²⁰ ; H _i , scalability coefficient;	CFA, Confirmatory Factor Analysis; TD, Typically Developing; ASD, Autism Spectrum Disorder, ID, Intellectual Disability;
WPPSI <i>, We</i>	echsler Preschool and Primary Scale of Intelligence; WISC-R, Wechsler Pres	school and Primary Scale of Intelligence-R; NDI, Neurodevelopmental Index; ANOVA, Analysis of Variance
10		
	For peer review only	/ - http://bmiopen.bmi.com/site/about/guidelines.xhtml
		,

 BMJ Open

Table 5: Criterion and predictive validity of assessment tools

Test	Criterion	Predictive
BAYLEY	Given but mean age <22 months. Not relevant to study population. ¹²	Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.32037
ш		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, ρ < 0.02) ³⁵ . BOT-2 ρ	Classification groups (DCD, at risk and TD) remained same over time (6 months) $\chi 2$
	=0.90 - 0.92 ²⁸ .	= 0.67 p = 0.72 ³⁵ . Predictive of motor impairment over 6-12 months (<i>N</i> =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%)) $^{ m 33}$
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, I	Bayley Scale of Infant and Toddler Development 3 rd edition; ¹² BOT-2, Bruininks-Oseretsky Test of Motor Prof	iciency 2 nd edition; ¹³ MABC-2, Movement Assessment Battery for Children 2 nd
edition; ¹⁴ N	MAND, McCarron Assessment of Neuromuscular Development; ¹⁵ NSMDA, Neurological Sensory Motor Devel	lopmental Assessment; ¹⁶ PDMS-2, Peabody Developmental Motor Scales 2 nd edition; ¹⁷
TGMD-II, T	est of Gross Motor Development 2 nd edition; ¹⁸ NS, Not Specified; SD, Standard Deviation; CP, Cerebral Palsy;	; TD, Typically Developing; ICC, Intraclass Correlation Coefficient; $\chi 2$, Chi Squared; NDI,
Neurodeve	lopmental Index; CSSA, Comprehensive Scales of Student Abilities	
19		
	For peer review only - http://bmionon.hmi.com/s	ite/about/quidelines.yhtml
	i or peer review only - http://binjopen.binj.com/s	nc, about guidelines.xittini

Table 6: Reliability of assessment tools

5	Test	Internal Consistency	Test-Retest	Intra-rater	Inter-rater	Minimal detectable change	Minimal clinical
6 7							important difference
8	BAYLEY	GM α = 0.87-0.93 MC: α 0.90-0.96 (24-	Gross Motor subtest (N=47) r=0.79	-	-	SEM Gross motor subtest 0.85-	-
9	Ш	42 months) ¹²	Motor component r=0.80 ¹²			1.08. of Motor component =	
10						3.00-4.74 (24-42 months) ¹²	
12	BOT-2	$(N = 100) \alpha = 0.92^{27}$	(N = 100) ICC = 0.99 ²⁷ $(N = 141)$ ICC	-	Total motor composite	4.18 (sensitivity 55.10%	6.53 (sensitivity 48.98%
13		$(N = 141) \alpha = 0.86^{28}$	= 0.97 ²⁸ 4-7 yrs (<i>N</i> = 43) <i>r</i> = 0.81 (8-		4-21 yrs (<i>N</i> = 47) <i>r</i> = 0.98	specificity 72.55%) ²⁷ 7.43	specificity 76.47%) ²⁷ 6.55
14		4-7 yrs (N= 620) α = 0.95 8-11 yrs (N=	12 yrs (<i>N</i> = 44) <i>r</i> = 0.80 ¹³		13	(sensitivity 42.49% specificity	(sensitivity 49.99%
15		450) α = 0.95 ¹³				65.72%) ²⁸	specificity 58.78%) ²⁸
17	MABC-2	(<i>N</i> = 60) M.D α = 0.51, A&C α = 0.70, Bal	(N=60) ICC = 0.85 ²⁹ Item ICC's	$(N=28) \kappa = 0.71^{30}$	Item ICC's range 0.892-	(N=28) Intrarater MDC = 3.43	-
18	(AB 1)	$\alpha = 0.66^{29} (N = 1823) \alpha = 0.502^{24} (N = 50)$	0.830-0.985 ²⁴ ICC test-retest =		0.998 ²⁴ ($N=22$) $\kappa = 0.60$	(N=22) Inter-tester MDC = 3.81	
19 20		$\alpha = 0.81 - 0.87^{30}$	0.83 ³⁰ Inter-rater test-retest ICC =		30	30	
21			0.79 ³⁰				
22	MABC-2	Translated version (Japanese) (N=132) α	-	ICC = 0.64 ³¹	ICC 0.63 31	Intra-rater SDC TTS: +/- 11.7	-
23 24	(AB 2)	= 0.602 ³²				TSS +/- 3.3. Inter-rater SDC	
24						TTS +/-16.0 TSS +/- 3.8 ³¹	
26	MABC-2	Subscales α = 0.78 (M.D = 0.77, BS =	<i>N</i> =60 (all 3 age bands) <i>r</i> =0.80 ¹⁴	ICC 0.88 35	ICC 0.96-0.99 35	SEM 1.34 (95%Cl) = 3 ¹⁴	1.39 (sensitivity 72.47%
27		0.52, Bal = 0.77) $^{35} \alpha$ = 0.88 36	<i>r</i> =0.74 p<0.0001 (standard score).			1.83 (95%Cl) ³⁶ 1.83 (sensitivity	specificity 46.18%) ^{28 36}
20 29		$(N = 141) \alpha = 0.88^{28}$	ICC standard score = 0.85^{35}			69.69% specificity 52.10%) ²⁸	
30			ICC 0.96 36				
31			N = 141 ICC =0.96 ²⁸				
32 . 33	MAND	-	-	-	-	<u>.</u>	-
34	NSMDA	Cross correlation matrix Item scoring	-	-	-	-	-
35		(12+24months) 0.73 <i>p</i> <0.001, Functional					
36 37		grade (12+24months)					
38	PDMS-2	(<i>N</i> =141) α=0.89 ²⁸ 24-35m α=0.97, 36-	<i>N</i> =141 ICC= 0.97 ²⁸	unable to extract	unable to extract data	7.76 (sensitivity 60.65%	8.39 (sensitivity 61.65%
39		47m α=0.95, 48-59m α=0.97, 60-71m α=		data for ≥24months	for \geq 24months ¹⁷	specificity 74.13%) ²⁸ SEM 24-	specificity 71.34%) ²⁸
40 · ⁄11							
42	20						
43							
44			For neer review only - http	·//hmionen.hmi.cor	n/site/about/quidelines	vhtml	
45 46			i or peer review only - http	.,, onjopen.onj.com	n, site, about, guideilles.		
47							

2							
3 ⁻ 1		0.98. For subgroups† α =0.99 ¹⁷		17		59 months = 3, 60-71m = 2 ¹⁷	
-+ - 5	TGMD-2	(<i>N</i> =1438) α=0.80 ⁴³ <i>N</i> =75 Locomotor	<i>N</i> =63 ICC=0.81 95% Cl ⁴³	N=32 ICC=0.97 95%	Obj ICC=0.93 ⁴² (<i>N</i> =50)		
6		subset α =0.71 object control α =0.72 ²¹	N=23 ICC=0.92 total 95% Cl ²¹	CI ⁴³	ICC=0.89 ²¹ ICC=0.75 ⁴⁴		
7		<i>N</i> =120 α = 0.72 ⁴⁵ <i>N</i> =99 α = 0.90 ²³ <i>N</i> =	<i>N</i> =99 <i>r</i> =0.98 23 Locomotor test <i>r</i> =	N=25 ICC=0.95 95%	<i>N</i> =8 <i>r</i> =1.00 ²³		
8		1208 Cronbach's α = 0.91 (gross motor	0.90 <i>p</i> <0.0001 object control test <i>r</i>	Cl ²¹ ICC = 0.78 ⁴⁴	L.S ICC=0.88 Obj		
9 10		quotient). Locomotor 0.85 and object	= 0.91 <i>p</i> <0.001 ²⁰ <i>N</i> = 75 <i>r</i> =0.96	ICC=0.92-0.99 ²⁰	ICC=0.89 ²⁰ N = 30		
11		control o.88. Note SEM GMQ = 4-5 SEM	overall (3-5 yrs <i>r</i> = 0.91), 6-8 years <i>r</i>		<i>r</i> =0.98 ¹⁸		
12		subsets=1 ¹⁸	= 0.95), (9-10 years <i>r</i> = 0.94) ¹⁸				
13 - 14	Bayley-II	I, Bayley Scale of Infant and Toddler Developm	ent 3 rd edition; ¹² BOT-2, Bruininks-Oseret	sky Test of Motor Proficie	ncy 2 nd edition; ¹³ MABC-2, Mov	vement Assessment Battery for Children 2 nd edition; ¹⁴ MAND,	
14	McCarro	n Assessment of Neuromuscular Development;	¹⁵ NSMDA, Neurological Sensory Motor D	Developmental Assessmen	t; ¹⁶ PDMS-2, Peabody Develop	mental Motor Scales 2 nd edition; ¹⁷ TGMD-II, Test of Gross Mo	tor
16	Develop	ment 2 nd edition; ¹⁸ GM, Gross Motor Subset; M	C, Motor Component; K, Kappa Coefficier	nt; M.D, Manual Dexterity	; BS, Ball Skills; BAL, Balance; A	&C, Aiming and catching; SDC, Smallest Detectable Change; T	ΤS,
17	Total Tes	st Score; TSS, Total Standard Score; †, gender, e	thnicity, speech/language or physical dis	order; Obj, Object Control	Subset; L.S, Locomotion Subse	t	
18							
20							
21							
22							
23 24							
24 25							
26							
27							
28							
29 30							
31							
32							
33							
34 35							
36							
37							
38							
39 40							
41	21						
42	21						
43							
44 45			For peer review only - http	://bmiopen.bmi.con	n/site/about/guidelines.	xhtml	
46			peer enerroiny inte		,, as e a garacines.		
47							

1 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.

Page 23 of 45

BMJ Open

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 clinicians⁷. 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,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Disability and Health (ICF) domains of participation, personal factors and environment². Clinicians should
 utilise appropriate assessments or questionnaires to ensure that these domains of health are also
 addressed in line with World Health Organisation guidelines².

4 When considering a test's reliability all three elements of test error should be taken into account – these 5 can be described as time sampling (assessed with test-retest reliability), content sampling (assessed as 6 internal consistency), and inter-scorer difference (or interrater reliability) ¹⁸. This is one of the reasons that 7 clinicians should consider repeating assessments and/or completing a second alternative assessment. All 8 assessments should be interpreted in conjunction with clinical reasoning and observation. Included 9 assessment tools are not intended to be diagnostic on their own; results need to be combined with other 10 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 (\geq 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,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 25 of 45

BMJ Open

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 times³⁰.

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 MABC¹⁴, 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 poor¹⁰. 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.

28 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.

7 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.

18 Figures

- 19 Figure 1. PRISMA flow diagram detailing study selection
- 20 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	1			
2				
3 ⊿				
4 5				
6				
7				
8				
9				
10 11				
12				
13				
14				
15				
16				
17 18				
19				
20				
21				
22				
23 24				
25				
26				
27				
28				
30				
31				
32				
33				
35				
36				
37				
38				
39 40				
41				
42				
43				
44 45				
45 46				
47				

1 References

2		
3	2	1. Piek JP, Baynam GB, Barrett NC. The relationship between fine and gross motor ability, self-perceptions and
4 5	3	self-worth in children and adolescents. Hum Mov Sci 2006;25(1):65-75. doi:
6 7	4	http://dx.doi.org/10.1016/j.humov.2005.10.011
8	5	2. World Health Organization. International Classification of Functioning, Disability and Health: ICF: World
9 10	6	Health Organization 2001.
11	7	3. Magalhaes LC, Cardoso AA, Missiuna C. Activities and participation in children with developmental
12 13	8	coordination disorder: a systematic review. Res Dev Disabil 2011;32(4):1309-16. doi:
14	9	10.1016/j.ridd.2011.01.029 [published Online First: 2011/02/19]
15 16	10	4. Spittle AJ, Spencer-Smith MM, Eeles AL, et al. Does the Bayley-III Motor Scale at 2 years predict motor
17	11	outcome at 4 years in very preterm children? Developmental Medicine And Child Neurology 2013;55(5):448-
18 19	12	52. doi: 10.1111/dmcn.12049
20	13	5. Logan SW, Robinson LE, Getchell N. The Comparison of Performances of Preschool Children on Two Motor
21	14	Assessments. Perceptual and Motor Skills 2011;113(3):715-23.
23 24	15	6. Venetsanou F, Kambas A. Environmental factors affecting preschoolers' motor development. Early
25	16	Childhood Education Journal 2010;37(4):319-27.
26 27	17	7. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy,
28	18	terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin
29 30	19	Epidemiol 2010;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006 [published Online First: 2010/05/25]
31 22	20	8. Slater LM, Hillier SL, Civetta LR. The clinimetric properties of performance-based gross motor tests used for
32 33	21	children with developmental coordination disorder: A systematic review. PEDIATRIC PHYSICAL THERAPY
34 35	22	2010;22(2):170.
36	23	9. Brown T, Lalor A. The Movement Assessment Battery for Children—Second Edition (MABC-2): A Review and
37 38	24	Critique. Phys Occup Ther Pediatr 2009;29(1):86-103.
39	25	10. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies
40 41	26	on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21(4):651-7. doi:
42 43	27	10.1007/s11136-011-9960-1 [published Online First: 2011/07/07]
44	28	11. Law M. Outcome measures rating form. Ontario, Canada: CanChild Centre for Disability Research, 2004.
45 46	29	12. Bayley N. Bayley Scales of Infant Development and Toddler Development: Technical Manual: The
47	30	PsychCorp 2006.
48 49	31	13. Bruininks R, Bruininks B. Bruininks-Oseretsky Test of Motor Proficiency–2nd Edition (BOT-2): Manual.
50	32	Circle Pines: MN: AGS Publishing 2005.
52	33	14. Henderson SE, Sugden DA, Barnett AL. Movement assessment battery for children-2: Movement ABC-2:
53 54	34	Examiner's manual: Pearson 2007.
55	35	15. McCarron LT. MAND: McCarron assessment of neuromuscular development, fine and gross motor abilities:
56 57	36	McCarron-Dial Systems, Incorporated 1997.
58		28
59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 29 of 45

BMJ Open

	1		16. Burns YR. N.S.M.D.A Physiotherapy Assessment for Infants and Young Children Second Edition. Brisbane,
1 2	2		Queensland: CopyRight Publishing Company 2014.
3	3		17. Folio M, Fewell R. Peabody Developmental Motor Scales. Examiner's Manual . 2nd Edition. Austin, Texas.:
4 5	4		Pro-Ed. 2000.
6	5		18. Ulrich DA. Test of gross motor development-2. Austin: Prod-Ed 2000
/ 8	6		19. Rudd J, Butson ML, Barnett L, et al. A holistic measurement model of movement competency in children.
9 10	7		Journal of Sports Sciences 2016;34(5):477-85.
11	8		20. Valentini NC. Validity and reliability of the TGMD-2 for Brazilian children. Journal of Motor Behavior
12 13	9		2012;44(4):275-80.
14	10		21. Houwen S, Hartman E, Jonker L, et al. Reliability and validity of the TGMD-2 in primary-school-age children
15 16	11		with visual impairments. Adapted Physical Activity Quarterly 2010;27(2):143-59.
17	12		22. Wong KYA, Yin Cheung S. Confirmatory factor analysis of the Test of Gross Motor Development-2.
18 19	13		Measurement in Physical Education & Exercise Science 2010;14(3):202-09. doi:
20	14		10.1080/10913671003726968
21 22	15		23. Simons J. Daly D. Theodorou F. et al. Validity and reliability of the TGMD-2 in 7-10-year-old Flemish
23 24	16		children with intellectual disability. Adapted physical activity guarterly : APAO 2008:25(1):71-82.
24 25	17		24. Hua J. Gu G. Meng W. et al. Age band 1 of the Movement Assessment Battery for Children-Second Edition:
26 27	_, 18		exploring its usefulness in mainland China. Research in Developmental Disabilities 2013:34(2):801-8.
28	10		25. Niemeijer AS, van Waelvelde H, Smits-Engelsman BC, Crossing the North Sea seems to make DCD
29 30	-J 20		disappear: cross-validation of Movement Assessment Battery for Children-2 norms. Hum Mov Sci
31	21		2015:39:177-88. doi: 10.1016/i.humov.2014.11.004
32 33	22		26. Visser L, Ruiter SAJ, Van der Meulen BF, et al. Low verbal assessment with the Bayley-III. Research in
34 25	23		Developmental Disabilities 2015:36:230-43.
36	24		27. Wuang YP. Su CY. Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency-
37 38	-4		Second Edition in children with intellectual disability. <i>Research in Developmental Disabilities</i> 2009:30(5):847-
39	-5 26		
40 41	20		28 Wuang VP, Su CY, Huang MH, Psychometric comparisons of three measures for assessing motor functions
42	~/ 28		in preschoolers with intellectual disabilities. <i>Journal of Intellectual Disability Research</i> 2012;56(6):567-78
43 44	20		29 Ellinoudis T. Evaggelinou C. Kourtessis T. et al. Reliability and validity of age hand 1 of the Movement
45	29		Assessment Battery for Children-second edition Res Dev Disabil 2011:32(3):1046-51 doi:
46 47	30 21		http://dx.doi.org/10.1016/i.ridd.2011.01.035
48 ⊿0	3± 22		30 Smits-Engelsman BCM. Niemeijer AS, van Waelvelde H. Is the Movement Assessment Battery for Children-
49 50	32		2nd edition a reliable instrument to measure motor performance in 2 year old shildren? <i>Research in</i>
51 52	33		Developmental Disabilities 2011;22(4):1270-77
53	34		Developmental Disabilities 2011, 52(4).1570-77.
54 55	35		batten for shildren 2, ageband 2, Besagreb in Developmental Disabilities 2012;24(2):705, 800
56	30		שמננפיץ וסי נוחוטופון 2, מצפטמוט 2. הפצעונוז זוז שפיפוסטוזופוונטו שוגטטווונופג 2013;34(2):795-800.
57 58		20	
59		29	For neer review only - http://bmionen.hmi.com/site/about/quidelines.yhtml
60			for peer review only intep.//binjopen.binj.com/site/about/guidennes.xitum
	1	32. Kita Y, Suzuki K, Hirata S, et al. Applicability of the Movement Assessment Battery for Children-Second	
----------	----------	--	------
1 2	2	Edition to Japanese children: A study of the Age Band 2. Brain & Development 2016;38(8):706-13.	
3	3	33. Griffiths A, Morgan P, Anderson PJ, et al. Predictive value of the Movement Assessment Battery for	
4 5	4	Children - Second Edition at 4 years, for motor impairment at 8 years in children born preterm. Dev Med	
6	5	Child Neurol 2017;59(5):490-96. doi: 10.1111/dmcn.13367 [published Online First: 2017/01/10]	
7 8	6	34. Schulz J, Henderson SE, Sugden DA, et al. Structural validity of the Movement ABC-2 test: factor structur	e
9 10	7	comparisons across three age groups. Research in Developmental Disabilities 2011;32(4):1361-9.	
10	8	35. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation,	
12 13	9	reliability, and validity for Brazilian children. <i>Res Dev Disabil</i> 2014;35(3):733-40. doi:	
14	10	http://dx.doi.org/10.1016/j.ridd.2013.10.028	
15 16	11	36. Wuang YP, Su JH, Su CY. Reliability and responsiveness of the Movement Assessment Battery for Childre	n-
17	12	Second Edition Test in children with developmental coordination disorder. Developmental Medicine & Chil	d
18 19	13	Neurology 2012;54(2):160-5.	
20	14	37. Hands B, Larkin D, Rose E. The psychometric properties of the McCarron Assessment of Neuromuscular	
21 22	, 15	Development as a longitudinal measure with Australian youth. Human Movement Science 2013:32(3):485-	
23	16	97.	
24 25	17	38 Danks M Maideen ME Burns YR et al. The long-term predictive validity of early motor development in	
26 27	-/ 18	"apparently normal" FLBW survivors <i>Early Human Development</i> 2012:88(8):637-41	
28	10	39 MacDonald L Burns Y. Performance on the NSMDA During the First and Second Year of Life to Predict	
29 30	-9 20	Functional Ability at the Age Of 4 in Children with Cerebral Palsy. Hong Kong Physiotherany Journal	
31	20	2005:23(1):40-45_doi: 10.1016/S1013-7025(09)70058-2	
32 33	22	40. Burns YR. Ensbey RM. Norrie MA. The Neuro-sensory motor developmental assessment part 1:	
34 25	22	development and administration of the test. Aust J Physiother 1989;35(3):141-49	
35 36	-5	41 Burns YR Ensbey RM Norrie MA. The neuro-sensory motor developmental assessment part II: predictiv	P
37 38	-4 2E	and concurrent validity. Aust I Physiother 1989:35(3):151-57	-
39	25 26	42 Barnett I.M. Minto C. Lander N. et al. Interrater reliability assessment using the Test of Gross Motor	
40 41	20	Development-2. Journal of Science and Medicine in Sport 2014:17(6):667-70. doi:	
42	2/ 28	bevelopment-2. Sournal of science and medicine in Sport $2014,17(0).007-70.001.$	
43 44	20	12 Earrokhi A. Zarob Zadob M. Karimi Alvar L. et al. Poliability and validity of test of gross motor developme	nt
45	29	45. Fair Okin A, Zaren Zauen M, Karmi Alvar L, et al. Kenability and valuity of test of gross motor developme	III-
46 47	30	2 (Officit, 2000) among 5-10 aged children of remain City. <i>Journal of Physical Education and Sports</i>	
48	31	Munugement 2014;5(2):18-28. 001: 10.5897/JPESW12.003	
49 50	32	44. Kim Y, Park I, Kang M. Examining rater effects of the TGMD-2 on children with intellectual disability.	
51 52	33	Adaptea Physical Activity Quarterly 2012;29(4):346-65.	
52 53	34	45. Kim Ci, Han Dw, Park IH. Reliability and validity of the test of gross motor development-II in Korean	
54 55	35	preschool children: applying AHP. Research in Developmental Disabilities 2014;35(4):800-7.	
56	36	46. Wuang YP, Lin YH, Su CY. Rasch analysis of the Bruininks-Oseretsky Test of Motor Proficiency-Second	
57 58	37	Edition in intellectual disabilities. <i>Research in Developmental Disabilities</i> 2009;30(6):1132-44.	
59		30	
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

Page	e 31 of 4	5 BMJ Open
1 2	1 2	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. <i>Dev Med Child Neurol</i> 2008;50(4):254-66. doi:
3	3	10.1111/j.1469-8749.2008.02025.x [published Online First: 2008/01/15]
5	4	48. McDowell I. Measuring health: a guide to rating scales and questionnaires: Oxford university press 2006.
6 7	5	49. Cools W, De Martelaer K, Samaey C, et al. Movement skill assessment of typically developing preschool
8	6	children: a review of seven movement skill assessment tools. (Report). Journal of Sports Science and Medicine
9 10	7	2009;8(2):154.
11 12	8	50. Adair B, Said CM, Rodda J, et al. Psychometric properties of functional mobility tools in hereditary spastic
13	9	paraplegia and other childhood neurological conditions. Dev Med Child Neurol 2012;54(7):596-605.
14 15 16 17 18 19 20 22 23 26 27 28 29 30 32 33 45 36 7 89 40 42 43 44 50 51 52 34 56 78 90 51 52 54 55 57 58 59	10	









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)

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
5	gross motor.mp.	3821
7	Psychomotor Disorders/	2609
8	Motor Skills Disorders/	2580
Ð	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
25	21 and 22 and 23 and 24	1152

Supplementary Table 1 OVID Medline database search (1006 to present)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Yield 2/7/17

	Search
S1	(MH "Child")
S2	(MH "Child, Preschool")
S3	"paediatric"
S4	(MH "Motor Skills")
S5	(MH "Motor Activity")
S6	(MH "Psychomotor Performance")
S7	(MH "Motor Skills Disorders")
S8	(MH "Developmental Disabilities")
S9	(MH "Child Development Disorders")
S10	"gross motor"
S11	(MH "Clinical Assessment Tools")
S12	(MH "Outcome Assessment")
\$13	(MH Physical Therapy Assessment")
\$14	"scale"
\$15	instrument*
S16	outcome*
\$17	measure*
S18	evaluat*
S19	assess*
S20	(MH Reliability and Validity")
S21	S1 OR S2 OR S3
\$22	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10
\$23	S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17
	S18 OR S19

2: CINAHL plus database search

Search No	Search	Search Yield
1	Child/	12/66/1
1		1340041
2		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	C 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	10121631
	19 or 20 or 21 or 22	
31	23 or 24 or 25 or 26 or 27	324779
32	28 and 29 and 30 and 31	1105

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

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

Search	Search	Yield 2/7/17
Number		
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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

	Measurement Property	Definition	Example/explanation
Validity	Content	The degree to which an assessment tool's content measures the construct that it intends to measure	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 validit (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, 0.4 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 erro
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)

 BMJ Open

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
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 instrue diagrams and photos of test proce
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 (scr 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 hyperactivity, encephalitis, mild h trauma, CP and muscular dystrop
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 strengt 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 (interven 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
TGMD-2 ¹⁸ ayley-III, Bayley S hildren 2 nd editio evelopmental M	Locomotion (6 items) and Object Control (6 items). Separate male/female norms for object control subset Scale of Infant and Toddler Development 3 rd edition ¹² ; BOT- on ¹⁴ ; MAND, McCarron Assessment of Neuromuscular Developtor Scales 2 nd edition ¹⁷ ;; TGMD-II, Test of Gross Motor Dev	Raw scores, standard scores, percentile rank, age equivalent, Gross Motor Quotient. Simple conversion. 2, Bruininks-Oseretsky Test of Motor Proficiency 2 nd edition opment ¹⁵ ; NSMDA, Neurological Sensory Motor Developme elopment 2 nd edition ¹⁸ ; GM, Gross Motor; FM, Fine Motor;	Simple to administer ¹³ ; MABC-2, Movement Assessment Batte ental Assessment ¹⁶ ; PDMS-2, Peabody NDI, Neurodevelopmental Index

. . . . d a dual a lateration of ~ مام **c**

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

to beet terien only

Page 43 of 45



PRISMA 2009 Checklist

4 5 6	Section/topic	#	Checklist item	Reported on page #
7	TITLE			
8 9	Title	1	Identify the report as a systematic review, meta-analysis, or both.	1 Title page
10	ABSTRACT			
112 12 13 14	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
15	INTRODUCTION			
17	, Rationale	З	Describe the rationale for the review in the context of what is already known.	4-5
18 19	B Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
20 21	METHODS			
22 23	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.	-
24 25 26	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
27 28	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
29 30 31 32	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
33	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	
36 37	, 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
38	³ 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
40 41 42	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
43	Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7
44 45	· · · · · · · · · · · · · · · · · · ·		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	

- 46
- 47



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.	
		Page 1 of 2	
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.b sensitivity or subgroup analyses meta-regression [see Item 16]).	-

Page 45 of 45



PRISMA 2009 Checklist

<u> </u>				
4	DISCUSSION			
5 6 7	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
8 9	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
10 11 12	Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	25-26
13	FUNDING			
14 15 16	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

BMJ Open

18 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

Page 2 of 2 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

SCHOLARONE[™] Manuscripts

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

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

Alison Griffiths^{1, 2, 3}, Rachel Toovey^{3, 4}, Prue E. Morgan¹, 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

1 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.

7 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. Testretest 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 ans.

1 Introduction

Motor function promotes cognitive and perceptual development in children and contributes to their ability to participate in their home, school and community environments¹. Motor impairment can negatively affect activity and participation levels of children², which may lead to lower levels of physical activity, fitness and health into adulthood³. 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 peers³. 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 children⁴. Further, assessment tools are essential to monitor the effects of interventions⁴. 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.

BMJ Open

A systematic review in 2010 by Slater⁸ 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 Lalor⁹ 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. 2.

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 \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; \geq 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 responsiveness⁷ (Supplementary Table 5). Cross-cultural validity, structural validity and hypothesis testing are all considered to be a component of construct validity⁷. Whilst predictive validity is considered to be a component of content validity, it is reported on separately in this paper for interpretability of results⁷. 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.

BMJ Open

3	
1	
4	
5	
6	
7	
8	
9	
10	
11	
12	
12	
13	
14	
15	
16	
17	
18	
19	
20	
20 21	
∠ I 22	
22	
23	
24	
25	
26	
27	
28	
20	
29	
20	
31	
32	
33	
34	
35	
36	
37	
30	
20	
39	
40	
41	
42	
43	
44	
45	
46	
17	
47	
4ð	
49	
50	
51	
52	
53	
54	
55	
55	
50	
5/	
58	
50	

60

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.

7 Patient and Public Involvement

8 As this was a systematic review of existing papers there was no patients or public involvement.

9 Results

Figure 1 provides details of study selection. Seven assessment tools were identified for 10 inclusion; Bayley Scale of Infant and Toddler Development III (Bayley-III), Bruininks-11 Oseretsky Test of Motor Proficiency 2 (BOT-2), Movement Assessment Battery for Children 2 12 (MABC-2), McCarron Assessment of Neuromuscular Development (MAND), Neurological 13 Sensory Motor Developmental Assessment (NSMDA), Peabody Developmental Motor Scales 14 2 (PDMS-2), and Test of Gross Motor Development 2 (TGMD-2). The corresponding manuals 15 were then added to the final yield resulting in thirty papers and seven manuals. Twenty 16 assessment tools were excluded (Supplementary Table 6). 17

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

3 some age groups).

4 The number of gross motor items for assessment vary both within and between the tools

5 (Table 1). For example, the number of items tested in the Bayley-III and the PDMS-2

6 depends on the age and ability of the child. Several assessments report criteria for

7 describing gross motor delay, although all test manuals warn against diagnosing delay based

8 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	Type of test	Normative sample (year)
Bayley-III ¹²	Gross motor fine	Static postures dynamic	1 mth _ 3	Developmental delay:	Discriminative	Predictive	Norm	1700 children
Dayley-III	motor cognitive		I IIIII – J	<25th centile or below	Discriminative	Evaluative,	NOITH	from the USA
	communication	niovement, balance	yı s	250 *		Research tool		(2000)
	social/emotional			230.		Research tool		(2000)
	adantive							
BOT-2 ¹³	Gross motor fine	Coordination balance	4 – 21 vrs	*	Discriminative	Research tool	Norm	1520 children
50. 2	motor	running speed and	,		Evaluative		Norm	from the USA
	motor	agility, strength			Evaluative			(2005)
MABC-2 ¹⁴	Gross motor, fine	Aiming and catching,	3 – 16 yrs	Traffic light system: Green	Discriminative	Intervention	Norm	1172 children
	motor, balance	static and dynamic		= normal, amber = 'at risk'	Evaluative	planning,		from United
		balance		and red = definite motor		Research tool		Kingdom (2006)
				impairment (<15%). *				
MAND ¹⁵	Gross and fine motor	Coordination, jumping,	3 yrs – 25	NDI 70-85 = mild	Evaluative	Research tool	Norm	2000 3-35 yrs
		static and dynamic	yrs	55-69 = moderate				from the USA
		balance		<55 = severe disability *				(1970's)
NSMDA ¹⁶	Gross Motor, Fine	Sitting, kneeling, walking,	1 mth – 6	Total score 6-8 normal, 9-	Evaluative	Predictive,	Criterion	N/A
	Motor, Neurological,	balance, running,	yrs	11 minimal, 12-14 mild, 15-	Discriminative	Research tool		
	Postural	hopping, jumping,		19 moderate, 20-25				
	Development, Infant	catching, motor planning		severe, >25 profound				
	Patterns of			disability *				
	Movement, Sensory							
	Motor. †							
PDMS-2 ¹⁷	Gross motor, fine	Stationary (standing	Birth – 5	*	Discriminative	Predictive,	Norm	2003 USA and
	motor	balance, sit-ups, push-	yrs		Evaluative	Research tool		Canada (1997-8)
		ups), locomotion						
		(walking, running,						
		jumping, hopping, etc.),						
		object manipulation						
		(kick, throw, hit, catch)						

TGMD-2 ¹⁸	Gross Motor	Locomotion (run, gallop,	3 – 10 yrs	*	Discriminative	Outcome	Norm	1208 USA
		hop, leap, jump, slide)			Evaluative	measure,		children (199
		and Object control				research tool,		1998)
		(batting, dribbling, catch,				intervention		
		kick, throw, roll)				planning		
hildren 2 nd edition; ¹⁴ lotor Scales 2 nd editio inical reasoning; †, ro	MAND, McCarron Asse on; ¹⁷ TGMD-II, Test of equires some manual h	Sevenopment 3' edition, BOT-2, essment of Neuromuscular Develo Gross Motor Development 2 nd edit andling; USA, United States of Am	bi unfinitis-Oser etsk pment; ¹⁵ NSMDA, N ion; ¹⁸ NDI, Neurode erica	eurological Sensory M evelopmental Index; Si	lotor Developmental Asse D, Standard Deviation; mtl	ssment; ¹⁶ PDMS-2, Pe	, Advisable to u	pmental ise
0								
		For peer review only - ht	tp://bmiopen.br	ni.com/site/about	/quidelines xhtml			

Page 11 of 45

BMJ Open

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.

14 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).

24 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.

1 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 Bruininks¹³ 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-2⁵ (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	Test Procedure	Target Examiner population	Training	Equipment/Manual
	administer (min)				
Bayley-III ¹²	30-90	Therapist administers in	Paediatric health professionals	Formal training not	Comprehensive manual/kit: £1089
		standardised order	early childhood specialists	required. DVD, webinars	Test kit provides most equipment
				and workshops available	
BOT-2 ¹³	40-60	Therapist administered in	Paediatric health professionals	Formal training not	Comprehensive manual/kit: £961
		standardised order	early childhood specialists	required	Test kit provides most equipment
MABC-2 ¹⁴	20-40	Therapist administers items in	Research psychologists, OT, PT,	Formal training not	Comprehensive manual/ kit: £1191
		standardised order. Some	Paediatricians	required.	Test kit provides most equipment
		flexibility allowed.			
MAND 15	15-20	Therapist administers items in	Professionals e.g. education,	Formal training not	Manual and test kit: £1366 includes
		standardised order.	neurology, OT, PT, psychology etc.	required.	equipment
NSMDA ¹⁶	20-45	Observation followed by	PT, OT	Formal training not	Comprehensive manual: £35.
		therapist administration of test		required (but is available)	Equipment not included
		items.			
PDMS-2 17	45-60 (20-30 for	Standardised procedure.	Paediatric health professionals, PE	Formal training not	Comprehensive manual/kit: £553
	GM only)		teachers, early intervention	required	Includes some but not all equipment
			specialists		required
TGMD-2 ¹⁸	15-20	Standardised procedure.	Teachers, health professionals (OT,	Formal training not	Kit includes manual and record form: £128.
			PT, doctors)	required	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

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 3. Methodological quality of included articles

Test	First author, Year	Country	Population	Internal	Reliability	Measurement	Content	Structural	Hypothesis	Cross-	Criterion	Responsive -
			(Age,	consistency		error	validity	validity	testing	cultural	validity	ness
			Diagnosis)							validity		
BAYLEY III	Bayley 12	USA	1-42 mths	Fair	Fair	Good	Excellent	Good	Good	-	Good	-
-	Spittle, et al. ⁴	Australia	2,4 yrs, Ex	-	-	-	-	-	-	-	Good	-
			prem									
	Visser, et al. ²⁶	Netherlands	2.2-10.8 yrs,	-	-	-	Excellent	Poor	-	-	-	-
			GDD, L.I.									
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	USA	4-21 yrs	Good	Fair (interrater)	Good	Excellent	Good	-	-	Good	-
	Bruininks ¹³				Fair (test-retest)							
MABC-2	Ellinoudis, et al. 29	Greece	3-5.5 yrs	Excellent	Good	-	-	-	-	-	-	-
(AB 1)	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	Belgium	3-4 yrs	Poor	Poor	Poor	-	-	-	-	-	-
	al. ³⁰											
MABC-2	Holm, et al. ³¹	Norway	7-9 yrs	-	Fair (interrater)	Poor		-	-	-	-	-
(AB 2)					Poor (intrarater)							
	Kita, et al. ³²	Japan	7-10 yrs	Excellent	-	-	-	-/1	-	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	-	-	-	-	-	-	-	Poor	-	-
		+ Belgium										
	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	-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 15 of 45

BMJ Open

	Wuang, et al. ²⁸	Taiwan	3-6 yrs, ID	Fair	Good	Good	-	-	-	-	Good	Fa
	Wuang, et al. ³⁶	Taiwan	6-12 yrs DCD	Poor	Fair	Good	-	-	-	-	-	Fa
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	-	9/	-	-	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	Fa
	Folio and Fewell 17	USA	0-71 mths	Good	-	Poor	Excellent	Good	Good	-	Poor	-
GMD-2	Barnett, et al. ⁴²	Australia	4-8 yrs	-	Fair		<u></u>	-	-	-	-	-
	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	-	. •	D 1	-	-	-	-
	Kim, et al. ⁴⁵	Korea	5-6 yrs	Poor	Fair	-	-	Poor	-	-	Poor	-
	Logan, et al. ⁵	USA	3-6 yrs	-	-	-	-	- 5	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	-

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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

Bayley-III, Bayley Scale of Infant and Toddler Development 3rd edition;¹² BOT-2, Bruininks-Oseretsky Test of Motor Proficiency 2rd edition;¹³ MABC-2, Movement Assessment Battery for Children 2rd edition;¹⁴ MAD, McCarron Assessment of Neuromuscular Developments¹⁴ NSMDA, Neurological Sensory Motor Developmental Assessment, ¹⁴ PDMS-2, Peabody Developmental Motor Scales 2rd edition;¹⁷; TGMD-II, Test of Gross Motor Development 2rd edition;¹⁸ Mths, Months; yrs, years; DCD, Developmental Coordination Disorder; VI, Vision Impairment; ID, Intellectual Disability; GDD, global developmental delay; L1, Language Impairment; ELBW, Extremely Low Birth Weight; VLBW, Very Low Birth Weight; CP, Cerebral Palsy; prem, premature; USA, United States of America



1 2	1	Reliability
3 ⊿	2	Internal consistency of assessments are summarised in Table 6. The BOT-2's high internal
5	3	consistency is well supported, including for children with an intellectual disability ^{28 46} . The MABC-2
6 7	4	appears to have lower internal consistency than the BOT-2, which may relate to the limited number
8 9	5	of test items (eight) on the MABC-2. The highest values for internal consistency for the MABC-2
10	6	were obtained in specific populations (intellectual disability and developmental coordination
11 12	7	disorder) with poor to fair methodology only. Conversely the highest quality articles reported the
13 14	8	lowest values, although it should be noted that these assessed age band 1 (3-6 years) only. Internal
15	9	consistency is reported to be high for the PDMS-2, while the Bayley-III is shown to have excellent
16	10	internal consistency in children aged 24-42 months The TGMD-2 is reported by two good quality
18 19	11	(and four poor to fair quality) articles to have excellent internal consistency, including for children
20	12	with vision impairment and intellectual disability. The MAND is the only assessment tool included in
21	12	this review without published data of internal consistency or reliability in this age group
23 24	-5	this review without published data of internal consistency of reliability in this use group.
25 26	14	The reliability findings are summarised in Table 6 and in Figures 2 and 3. Test-retest reliability was
20 27	15	excellent in the Bayley-III (Table 6), BOT-2 and PDMS-2; and was good to excellent in the MABC-2
28 29	16	and TGMD-2 (Figure 2). Intra-rater reliability was rarely investigated or reported for most tools, with
30 21	17	the TGMD-2 demonstrating better results than the MABC-2 (Figure 3). Only the TGMD-2 and
32	18	MABC-2 report inter-rater reliability values using an ICC (Figure 3) ^{31 42} . Inter-rater reliability is also
33 34	19	supported in the BOT-2 with Pearson Correlation Coefficient and Kappa respectively. The studies
35 36	20	referred to in the test manuals for the TGMD-2, Bayley-III, BOT-2 and MABC-2 all report reliability
37	21	findings using Pearson's correlation, which is less ideal than an ICC or weighted kappa for statistical
38 39	22	analysis ^{47 48} . Only studies reporting ICC's are visually represented in Figures 2 (test-retest) and 3
40 41	23	(inter and intra-rater). The TGMD-2 test-retest reliability results from Houwen, et al. ²¹ were
42	24	believed to contain an error as the reported ICC was outside of the reported confidence intervals
43 44	25	(ICC 0.92, 0.82-0.91). This data set was therefore excluded from Figure 2.
45 46		
47	26	Responsiveness was reported for the Bayley-III, BOT-2, MABC-2 and PDMS-2 with minimal
48 49	27	detectable change (MDC) or a standard error of measurement (SEM) ^{2°} . Sensitivity and specificity
50 51	28	for detecting change was shown to be satisfactory in the MABC-2, PDMS-2 and MABC-2 ²⁸ (Table 6).
52	29	There have been no studies to date on the responsiveness of the TGMD-2, NSMDA or MAND.
53 54		
55 56		

Test	Content	Construct
BAYLEY	Expert opinion for standard and low verbal version ^{12 26} . Literature	Factor analysis. Difference in mean scores with pervasive developmental disorder, and specific language impairment ¹²
ш	reviews. Gross motor score correlated with Motor component 0.70 $^{\rm 12}$	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	Factor analysis, scores increase with age, discriminates between normal and children with DCD (N=50), high-
	standardisation studies, professional reviews ¹³	functioning ASD ($N = 45$) and mild-moderate ID ($N = 66$) ¹³
VABC-2	Expert Panel, Stakeholder feedback, Literature review ³¹	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 .
	Expert panel - clarity (validity content index 71.8-93.9, Kappa 0.76-	Significant difference between TD, DCD and at risk DCD scores (η 2 = 0.63) p < 0.0001 ³⁵ . UK norms not appropriate to
	0.88) and pertinence (98.5-99.3 and kappa 0.83-0.92) $p{<}0.001^{35}$	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	Factor analysis ^{15 37} . Scores increase with age, and discriminate between typically developing children and those with
	of tasks during development ¹⁵	head trauma or neurological dysfunction as well as gender ^{15 37}
NSMDA	Literature review. Developed by an experienced paediatric	Factor analysis (up to 2 years of age) 40 41. Stability of test results over time (up to 2 years) 40 41.
	physiotherapist ⁴⁰	
DMS-2	Literature review. Created by experts in the field. Revised with	Item response modelling. Factor analysis. Differential item functioning analysis. Scores correlated with age (r=0.80-
	feedback from therapists guided revision. Hierarchical sequence of	0.93) 17
	items ¹⁷	
GMD-2	Expert Panel (3 PE teachers with post-grad qualifications) ¹⁸ .	Exploratory and confirmatory factor analysis ^{19 20 18 21 22 23} High and significant correlation of increasing age and
	Translated version (Brazilian Portuguese) language clarity 0.96,	increasing scores 43 . Age and disability differentiation $^{18\ 23}$ Subtest correlation 0.41 18
	pertinence >0.89. Experts CVI for clarity and pertinence were also	Galloping, running and leaping not well correlated with locomotion subscale. Object control significant & highly
	strong- α = 0.93 clarity and α =0.91 pertinence ²⁰	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
		and we have a set and we have a set and a set and
Bayley-III,	Bayley Scale of Infant and Toddler Development 3 rd edition; BOT-2, Brui	Ininks-Oseretsky Test of Motor Proficiency 2 rd edition; ²¹ MABC-2, Movement Assessment Battery for Children 2 rd
edition; ¹⁴ I	MAND, McCarron Assessment of Neuromuscular Development; NSMDA	, Neurological Sensory Motor Developmental Assessment; ²⁰ PDMS-2, Peabody Developmental Motor Scales 2 ¹⁰ edition; ²
TGMD-II, T	Fest of Gross Motor Development 2 ^{nu} edition; ¹⁵ ; H _i , scalability coefficient;	CFA, Confirmatory Factor Analysis; TD, Typically Developing; ASD, Autism Spectrum Disorder, ID, Intellectual Disability;
WPPSI, We	echsler Preschool and Primary Scale of Intelligence; WISC-R, Wechsler Pres	school and Primary Scale of Intelligence-R; NDI, Neurodevelopmental Index; ANOVA, Analysis of Variance
18		
	For neer review only	y - http://bmiopen.hmi.com/site/about/guidelines.yhtml
	To peer review only	y maps, singepending.com/ste/ubout/guidennes.nttilli

 BMJ Open

Table 5: Criterion and predictive validity of assessment tools

Test	Criterion	Predictive
BAYLEY	Given but mean age <22 months. Not relevant to study population. ¹²	Motor impairment at 4 years: Bayley III at 2 years <1SD = sensitivity 0.32037
ш		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, ρ < 0.02) ³⁵ . BOT-2 ρ	Classification groups (DCD, at risk and TD) remained same over time (6 months) $\chi 2$
	=0.90 - 0.92 ²⁸ .	= 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%)) $^{ m ^{33}}$
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 (<i>N</i> = 148) predictive of medical diagnosis χ^2 = 0.08 <i>p</i> = 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, Ba	ayley Scale of Infant and Toddler Development 3 rd edition; ¹² BOT-2, Bruininks-Oseretsky Test of Motor Profi	iciency 2 nd edition; ¹³ MABC-2, Movement Assessment Battery for Children 2 nd
edition; ¹⁴ M	AND, McCarron Assessment of Neuromuscular Development; ¹⁵ NSMDA, Neurological Sensory Motor Devel	lopmental Assessment; ¹⁶ PDMS-2, Peabody Developmental Motor Scales 2 nd edition; ¹⁷
TGMD-II, Te	st of Gross Motor Development 2 nd edition; ¹⁸ NS, Not Specified; SD, Standard Deviation; CP, Cerebral Palsy;	; TD, Typically Developing; ICC, Intraclass Correlation Coefficient; χ 2, Chi Squared; NDI,
Neurodevel	opmental Index; CSSA, Comprehensive Scales of Student Abilities	
19		
	For pear vertices only http://humics.org/articles/	ite/about/quidalines.vhtml
	For peer review only - http://bmjopeh.bmj.com/s	ite/about/guidelines.xntmi
Table 6: Reliability of assessment tools

5	Test	Internal Consistency	Test-Retest	Intra-rater	Inter-rater	Minimal detectable change	Minimal clinical
6 7							important difference
8	BAYLEY	GM α = 0.87-0.93 MC: α 0.90-0.96 (24-	Gross Motor subtest (N=47) r=0.79	-	-	SEM Gross motor subtest 0.85-	-
9	Ш	42 months) ¹²	Motor component <i>r</i> =0.80 ¹²			1.08. of Motor component =	
10 11						3.00-4.74 (24-42 months) ¹²	
12	BOT-2	$(N = 100) \alpha = 0.92^{27}$	(N = 100) ICC = 0.99 ²⁷ $(N = 141)$ ICC	-	Total motor composite	4.18 (sensitivity 55.10%	6.53 (sensitivity 48.98%
13		$(N = 141) \alpha = 0.86^{28}$	= 0.97 28 4-7 yrs (N = 43) r = 0.81 (8-		4-21 yrs (<i>N</i> = 47) <i>r</i> = 0.98	specificity 72.55%) ²⁷ 7.43	specificity 76.47%) ²⁷ 6.55
14		4-7 yrs (N= 620) α = 0.95 8-11 yrs (N=	12 yrs (<i>N</i> = 44) <i>r</i> = 0.80 ¹³		13	(sensitivity 42.49% specificity	(sensitivity 49.99%
15		450) <i>α</i> = 0.95 ¹³				65.72%) ²⁸	specificity 58.78%) ²⁸
17	MABC-2	(<i>N</i> = 60) M.D α = 0.51, A&C α = 0.70, Bal	(N=60) ICC = 0.85 ²⁹ Item ICC's	$(N=28) \kappa = 0.71^{30}$	Item ICC's range 0.892-	(N=28) Intrarater MDC = 3.43	-
18	(AB 1)	$\alpha = 0.66^{29} (N = 1823) \alpha = 0.502^{24} (N = 50)$	0.830-0.985 ²⁴ ICC test-retest =		0.998 ²⁴ ($N=22$) $\kappa = 0.60$	(N=22) Inter-tester MDC = 3.81	
19 20		$\alpha = 0.81 - 0.87^{30}$	o.83 ³⁰ Inter-rater test-retest ICC =		30	30	
21			0.79 ³⁰				
22	MABC-2	Translated version (Japanese) (N=132) α	-	ICC = 0.64 ³¹	ICC 0.63 31	Intra-rater SDC TTS: +/- 11.7	-
23 24	(AB 2)	= 0.602 ³²				TSS +/- 3.3. Inter-rater SDC	
24						TTS +/-16.0 TSS +/- 3.8 ³¹	
26	MABC-2	Subscales α = 0.78 (M.D = 0.77, BS =	<i>N</i> =60 (all 3 age bands) <i>r</i> =0.80 ¹⁴	ICC 0.88 35	ICC 0.96-0.99 35	SEM 1.34 (95%Cl) = 3 ¹⁴	1.39 (sensitivity 72.47%
27		0.52, Bal = 0.77) $^{35} \alpha$ = 0.88 36	<i>r</i> =0.74 p<0.0001 (standard score).			1.83 (95%Cl) ³⁶ 1.83 (sensitivity	specificity 46.18%) ^{28 36}
28 29		$(N = 141) \alpha = 0.88^{28}$	ICC standard score = 0.85^{35}			69.69% specificity 52.10%) ²⁸	
30			ICC 0.96 36				
31			<i>N</i> = 141 ICC =0.96 ²⁸				
32	MAND	-		-	-	<u>-</u>	-
34	NSMDA	Cross correlation matrix Item scoring	-	-	-	-	-
35		(12+24months) 0.73 <i>p</i> <0.001, Functional					
36		grade (12+24months)					
38	PDMS-2	(<i>N</i> =141) α=0.89 ²⁸ 24-35m α=0.97, 36-	N=141 ICC= 0.97 ²⁸	unable to extract	unable to extract data	7.76 (sensitivity 60.65%	8.39 (sensitivity 61.65%
39		47m α=0.95, 48-59m α=0.97, 60-71m α=		data for ≥24months	for ≥24months ¹⁷	specificity 74.13%) ²⁸ SEM 24-	specificity 71.34%) ²⁸
40 ·							
41	20						
43							
44			Ear poor raviaw aply bttp	·//bmionon.hmi.com	n/sita/about/quidalinas	vhtml	
45 46			i or peer review only - http	.// 5111000011.01111.001	n/ site/ about/ guideillies.		
47							

2							
3 ⁻ 1		0.98. For subgroups $\dagger \alpha$ = 0.99 17		17		59 months = 3, 60-71m = 2 ¹⁷	
-+ - 5	TGMD-2	(<i>N</i> =1438) α=0.80 ⁴³ <i>N</i> =75 Locomotor	<i>N</i> =63 ICC=0.81 95% Cl 43	N=32 ICC=0.97 95%	Obj ICC=0.93 ⁴² (<i>N</i> =50)		
6		subset α =0.71 object control α =0.72 ²¹	N=23 ICC=0.92 total 95% Cl ²¹	CI 43	ICC=0.89 ²¹ ICC=0.75 ⁴⁴		
7		<i>N</i> =120 α = 0.72 ⁴⁵ <i>N</i> =99 α = 0.90 ²³ <i>N</i> =	<i>N</i> =99 <i>r</i> =0.98 ²³ Locomotor test <i>r</i> =	N=25 ICC=0.95 95%	N=8 r= 1.00 ²³		
8		1208 Cronbach's α = 0.91 (gross motor	0.90 <i>p</i> <0.0001 object control test <i>r</i>	Cl ²¹ ICC = 0.78 ⁴⁴	L.S ICC=0.88 Obj		
9 10		quotient). Locomotor 0.85 and object	= 0.91 <i>p</i> <0.001 ²⁰ <i>N</i> = 75 <i>r</i> =0.96	ICC=0.92-0.99 ²⁰	ICC=0.89 ²⁰ N = 30		
11		control o.88. Note SEM GMQ = 4-5 SEM	overall (3-5 yrs r = 0.91), 6-8 years r		<i>r</i> =0.98 ¹⁸		
12		subsets=1 ¹⁸	= 0.95), (9-10 years <i>r</i> = 0.94) ¹⁸				
13 - 14	Bayley-II	I, Bayley Scale of Infant and Toddler Developm	ent 3 rd edition; ¹² BOT-2, Bruininks-Oseret	sky Test of Motor Proficier	ncy 2 nd edition; ¹³ MABC-2, Mov	vement Assessment Battery for Children 2 nd edition; ¹⁴	MAND,
14	McCarro	n Assessment of Neuromuscular Development,	¹⁵ NSMDA, Neurological Sensory Motor E	Developmental Assessmen	t; ¹⁶ PDMS-2, Peabody Develop	mental Motor Scales 2 nd edition; ¹⁷ TGMD-II, Test of Gr	oss Motor
16	Developr	ment 2 nd edition; ¹⁸ GM, Gross Motor Subset; M	IC, Motor Component; K, Kappa Coefficier	nt; M.D, Manual Dexterity	; BS, Ball Skills; BAL, Balance; A	&C, Aiming and catching; SDC, Smallest Detectable Ch	ange; TTS,
17	Total Tes	st Score; TSS, Total Standard Score; †, gender, e	ethnicity, speech/language or physical dis	order; Obj, Object Control	Subset; L.S, Locomotion Subse	t	
18							
20							
21							
22							
23 24							
24 25							
26							
27							
28							
29 30							
31							
32							
33							
34 35							
36							
37							
38							
39 40							
41	21						
42	27						
43							
44 45			For peer review only - httr	://bmiopen.bmi.con	n/site/about/guidelines.	xhtml	
46			· · · · · · · · · · · · · · · · · · ·			-	
47							

1 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.

Page 23 of 45

BMJ Open

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 clinicians⁷. 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 environment². Clinicians should
utilise appropriate assessments or questionnaires to ensure that these domains of health are also
addressed in line with World Health Organisation guidelines².

4 When considering a test's reliability all three elements of test error should be taken into account – these 5 can be described as time sampling (assessed with test-retest reliability), content sampling (assessed as 6 internal consistency), and inter-scorer difference (or interrater reliability) ¹⁸. This is one of the reasons that 7 clinicians should consider repeating assessments and/or completing a second alternative assessment. All 8 assessments should be interpreted in conjunction with clinical reasoning and observation. Included 9 assessment tools are not intended to be diagnostic on their own; results need to be combined with other 10 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.

Page 25 of 45

BMJ Open

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 times³⁰.

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 MABC¹⁴, 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 poor¹⁰. 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.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

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 rez 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 for beer terien only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1 References

1	1	References
2 3	2	1. Piek JP, Baynam GB, Barrett NC. The relationship between fine and gross motor ability, self-perceptions and
4 5	3	self-worth in children and adolescents. Human Movement Science 2006;25(1):65-75. doi:
6	4	http://dx.doi.org/10.1016/j.humov.2005.10.011
7 8	5	2. World Health Organization. International Classification of Functioning, Disability and Health: ICF: World
9	6	Health Organization 2001.
10 11	7	3. Magalhaes LC, Cardoso AA, Missiuna C. Activities and participation in children with developmental
12	8	coordination disorder: a systematic review. Research in developmental disabilities 2011;32(4):1309-16. doi:
14	9	10.1016/j.ridd.2011.01.029 [published Online First: 2011/02/19]
15 16	10	4. Spittle AJ, Spencer-Smith MM, Eeles AL, et al. Does the Bayley-III Motor Scale at 2 years predict motor
17	11	outcome at 4 years in very preterm children? Developmental Medicine And Child Neurology 2013;55(5):448-
18 19	12	52. doi: 10.1111/dmcn.12049
20	13	5. Logan SW, Robinson LE, Getchell N. The Comparison of Performances of Preschool Children on Two Motor
21 22	14	Assessments. Perceptual and Motor Skills 2011;113(3):715-23.
23 24	15	6. Venetsanou F, Kambas A. Environmental factors affecting preschoolers' motor development. Early
24	16	Childhood Education Journal 2010;37(4):319-27.
26 27	17	7. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy,
28	18	terminology, and definitions of measurement properties for health-related patient-reported outcomes.
29 30	19	Journal of clinical epidemiology 2010;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006 [published Online
31	20	First: 2010/05/25]
32 33	21	8. Slater LM, Hillier SL, Civetta LR. The clinimetric properties of performance-based gross motor tests used for
34 35	22	children with developmental coordination disorder: A systematic review. PEDIATRIC PHYSICAL THERAPY
36	23	2010;22(2):170.
37 38	24	9. Brown T, Lalor A. The Movement Assessment Battery for Children—Second Edition (MABC-2): A Review and
39	25	Critique. Physical & Occupational Therapy in Pediatrics 2009;29(1):86-103.
40 41	26	10. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies
42 43	27	on measurement properties: a scoring system for the COSMIN checklist. Quality of life research : an
43 44	28	international journal of quality of life aspects of treatment, care and rehabilitation 2012;21(4):651-7. doi:
45 46	29	10.1007/s11136-011-9960-1 [published Online First: 2011/07/07]
47	30	11. Law M. Outcome measures rating form. Ontario, Canada: CanChild Centre for Disability Research, 2004.
48 49	31	12. Bayley N. Bayley Scales of Infant Development and Toddler Development: Technical Manual: The
50	32	PsychCorp 2006.
51 52	33	13. Bruininks R, Bruininks B. Bruininks-Oseretsky Test of Motor Proficiency–2nd Edition (BOT-2): Manual.
53 54	34	Circle Pines: MN: AGS Publishing 2005.
55	35	14. Henderson SE, Sugden DA, Barnett AL. Movement assessment battery for children-2: Movement ABC-2:
56 57	36	Examiner's manual: Pearson 2007.
58		28
59 60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 29 of 45

60

BMJ Open

2	McCarron-Dial Systems Incorporated 1997
3	16. Burns YR. N.S.M.D.A Physiotherapy Assessment for Infants and Young Children Second Edition. Brisbane,
4	Queensland: CopyRight Publishing Company 2014.
5	17. Folio M, Fewell R. Peabody Developmental Motor Scales. Examiner's Manual . 2nd Edition. Austin, Texas.:
6	Pro-Ed. 2000.
7	18. Ulrich DA. Test of gross motor development-2. Austin: Prod-Ed 2000
8	19. Rudd J, Butson ML, Barnett L, et al. A holistic measurement model of movement competency in children.
9	Journal of Sports Sciences 2016;34(5):477-85.
10	20. Valentini NC. Validity and reliability of the TGMD-2 for Brazilian children. Journal of Motor Behavior
11	2012;44(4):275-80.
12	21. Houwen S, Hartman E, Jonker L, et al. Reliability and validity of the TGMD-2 in primary-school-age children
13	with visual impairments. Adapt Phys Act Q 2010;27(2):143-59.
14	22. Wong KYA, Yin Cheung S. Confirmatory factor analysis of the Test of Gross Motor Development-2.
15	Measurement in Physical Education & Exercise Science 2010;14(3):202-09. doi:
16	10.1080/10913671003726968
17	23. Simons J, Daly D, Theodorou F, et al. Validity and reliability of the TGMD-2 in 7-10-year-old Flemish
18	children with intellectual disability. Adapted physical activity quarterly : APAQ 2008;25(1):71-82.
19	24. Hua J, Gu G, Meng W, et al. Age band 1 of the Movement Assessment Battery for Children-Second Edition:
20	exploring its usefulness in mainland China. Research in developmental disabilities 2013;34(2):801-8.
21	25. Niemeijer AS, van Waelvelde H, Smits-Engelsman BC. Crossing the North Sea seems to make DCD
22	disappear: cross-validation of Movement Assessment Battery for Children-2 norms. Hum Mov Sci
23	2015;39:177-88. doi: 10.1016/j.humov.2014.11.004
24	26. Visser L, Ruiter SAJ, Van der Meulen BF, et al. Low verbal assessment with the Bayley-III. Research in
25	developmental disabilities 2015;36:230-43.
26	27. Wuang YP, Su CY. Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency-
27	Second Edition in children with intellectual disability. Research in developmental disabilities 2009;30(5):847-
28	55.
29	28. Wuang YP, Su CY, Huang MH. Psychometric comparisons of three measures for assessing motor functions
30	in preschoolers with intellectual disabilities. Journal of Intellectual Disability Research 2012;56(6):567-78.
31	29. Ellinoudis T, Evaggelinou C, Kourtessis T, et al. Reliability and validity of age band 1 of the Movement
32	Assessment Battery for Childrensecond edition. Research in developmental disabilities 2011;32(3):1046-51.
33	doi: http://dx.doi.org/10.1016/j.ridd.2011.01.035
34	30. Smits-Engelsman BCM, Niemeijer AS, van Waelvelde H. Is the Movement Assessment Battery for Children-
25	2nd edition a reliable instrument to measure motor performance in 3 year old children? Research in
35	
35 36	developmental disabilities 2011;32(4):1370-77.
35 36	developmental disabilities 2011;32(4):1370-77.
	4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34

	1	31	Holm I, Tveter AT, Aulie VS, et al. High intra- and inter-rater chance variation of the movement assessment
1 2	2	b	pattery for children 2, ageband 2. Research in developmental disabilities 2013;34(2):795-800.
3	3	32	. Kita Y, Suzuki K, Hirata S, et al. Applicability of the Movement Assessment Battery for Children-Second
4 5	4	E	Edition to Japanese children: A study of the Age Band 2. <i>Brain & Development</i> 2016;38(8):706-13.
6	5	33	. Griffiths A, Morgan P, Anderson PJ, et al. Predictive value of the Movement Assessment Battery for
/ 8	6	C	Children - Second Edition at 4 years, for motor impairment at 8 years in children born preterm. Dev Med
9	7	C	Child Neurol 2017;59(5):490-96. doi: 10.1111/dmcn.13367 [published Online First: 2017/01/10]
10 11	8	34	. Schulz J, Henderson SE, Sugden DA, et al. Structural validity of the Movement ABC-2 test: factor structure
12 13	9	C	comparisons across three age groups. Research in developmental disabilities 2011;32(4):1361-9.
14	10	35	. Valentini NC, Ramalho MH, Oliveira MA. Movement assessment battery for children-2: translation,
15 16	11	r	reliability, and validity for Brazilian children. Research in developmental disabilities 2014:35(3):733-40. doi:
17	12	h	http://dx.doi.org/10.1016/i.ridd.2013.10.028
18 19	13	- 36	Wuang YP. Su JH. Su CY. Reliability and responsiveness of the Movement Assessment Battery for Children-
20	-J 14	S	Second Edition Test in children with developmental coordination disorder. Developmental Medicine & Child
21 22	15	^	Neurology 2012:54(2):160-5
23	-J 16	37	Hands B Larkin D. Rose F. The psychometric properties of the McCarron Assessment of Neuromuscular
24 25	17	с, Г	Development as a longitudinal measure with Australian youth Human Movement Science 2013;32(3):485-
26 27	-7 18	- 0	
28	10	38	B Danks M Maideen ME Burns VB et al. The long-term predictive validity of early motor development in
29 30	-9		annarently normal" ELBW survivors <i>Farly Human Development</i> 2012:88(8):637-41
31	20	39	MacDonald L Burns Y. Performance on the NSMDA During the First and Second Year of Life to Predict
32 33	22	F	Functional Ability at the Age Of 4 in Children with Cerebral Palsy. Hong Kong Physiotherapy Journal
34 25	22		2005:23(1):40-45 doi: 10.1016/S1013-7025(09)70058-2
35 36	-J 24	40	Burns YB. Ensbey RM. Norrie MA. The Neuro-sensory motor developmental assessment part 1:
37 38	-+ 2E	d.	levelopment and administration of the test. Australian Journal of Physiotherapy 1989:35(3):141-49
39	25 26	۲ 41	Burns VR Enshey RM Norrie MA. The neuro-sensory motor developmental assessment part II: predictive
40 41	20	1T 2	and concurrent validity. Australian Journal of Physiotherapy 1989;35(3):151-57
42	2/ 28	42	Barnett I.M. Minto C. Lander N. et al. Interrater reliability assessment using the Test of Gross Motor
43 44	20	-τ <u>2</u> Γ	Development-2 Journal of Science and Medicine in Sport 2014;17(6):667-70 doi:
45	29	L	$\frac{1}{2}$
46 47	30	<u>י</u> גו	$\frac{1}{1}$
48 40	31	45	2. Farrowin A, Zaren Zauen W, Karinn Alvar L, et al. Kenability and validity of test of gloss motor development-
49 50	32	2	2 (Onich, 2000) among 5-10 aged children of Tenran City. <i>Journal of Physical Education and Sports</i>
51 52	33	л а а	Vanagement 2014,5(2).18-28. doi: 10.3897/JPESM12.003
52	34	44	A KINET, PARKI, KANG WI, EXAMINING FALEF ENELLS OF THE FOMD-2 ON CHILDREN WITH INTELECTUAL DISADILITY. Adapt
54 55	35	P 4 F	mys Act Q 2012;29(4):346-65.
56	30	45	. KITI CI, Hari DW, Park IH. Kellability and validity of the test of gross motor development-II in Korean
57 58	37	p	preschool children: applying AHP. <i>Research in developmental disabilities</i> 2014;35(4):800-7.
59		30	For near review only - http://bmionen.hmi.com/site/shout/guidalings.yhtml
60			for peer review only intep.//binjopen.binj.com/site/about/guidelines.xittin

Page	e 31 of 45	BMJ Open
	1	46. Wuang YP, Lin YH, Su CY. Rasch analysis of the Bruininks-Oseretsky Test of Motor Proficiency-Second
1 2	2	Edition in intellectual disabilities. Research in developmental disabilities 2009;30(6):1132-44.
3	3	47. Spittle AJ, Doyle LW, Boyd RN. A systematic review of the clinimetric properties of neuromotor
4 5	4	assessments for preterm infants during the first year of life. Dev Med Child Neurol 2008;50(4):254-66. doi:
6	5	10.1111/j.1469-8749.2008.02025.x [published Online First: 2008/01/15]
8	6	48. McDowell I. Measuring health: a guide to rating scales and questionnaires: Oxford university press 2006.
9 10	7	49. Cools W, De Martelaer K, Samaey C, et al. Movement skill assessment of typically developing preschool
11	8	children: a review of seven movement skill assessment tools. (Report). Journal of Sports Science and Medicine
12 13	9	2009;8(2):154.
14	10	50. Adair B, Said CM, Rodda J, et al. Psychometric properties of functional mobility tools in hereditary spastic
15 16	11	paraplegia and other childhood neurological conditions. Developmental Medicine & Child Neurology
17	12	2012;54(7):596-605.
18 19		
20	13	
21		
23 24		
25		
26 27		
28		
29 30		
31		
32 33		
34 25		
36		
37 38		
39		
40 41		
42		
43 44		
45		
46 47		
48 40		
49 50		
51 52		
53		
54 55		
56		
57 58		
59	31	For poor rovious only http://bmionon.hmi.com/site/ahast/avidalines.shtml
60		For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xntml









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)

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
5	gross motor.mp.	3821
7	Psychomotor Disorders/	2609
8	Motor Skills Disorders/	2580
Ð	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
25	21 and 22 and 23 and 24	1152

Supplementary Table 1 OVID Medline database search (1006 to present)

Yield 2/7/17

	Search		
S1	(MH "Child")		
S2	(MH "Child, Preschool")		
S3	"paediatric"		
S4	(MH "Motor Skills")		
S5	(MH "Motor Activity")		
S6	(MH "Psychomotor Performance")		
S7	(MH "Motor Skills Disorders")		
S8	(MH "Developmental Disabilities")		
S9	(MH "Child Development Disorders")		
S10	"gross motor"		
S11	(MH "Clinical Assessment Tools")		
S12	(MH "Outcome Assessment")		
\$13	(MH Physical Therapy Assessment")		
\$14	"scale"		
\$15	instrument*		
S16	outcome*		
\$17	measure*		
S18	evaluat*		
S19	assess*		
S20	(MH Reliability and Validity")		
S21	S1 OR S2 OR S3		
\$22	S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10		
\$23	S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17		
	S18 OR S19		

2: CINAHL plus database search

Search No	Search	Search Yield
1	Child/	12/66/1
1		1340041
2		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	C 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	10121631
	19 or 20 or 21 or 22	
31	23 or 24 or 25 or 26 or 27	324779
32	28 and 29 and 30 and 31	1105

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

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

Search	Search	Yield 2/7/17
Number		
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

	Measurement Property	Definition	Example/explanation
Validity	Content	The degree to which an assessment tool's content measures the construct that it intends to measure	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 validit (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, 0.4 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 erro
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)

 BMJ Open

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
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 instrue diagrams and photos of test proce
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 (scr 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 hyperactivity, encephalitis, mild h trauma, CP and muscular dystrop
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 strengt 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 (interven 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
TGMD-2 ¹⁸ ayley-III, Bayley S hildren 2 nd editio evelopmental M	Locomotion (6 items) and Object Control (6 items). Separate male/female norms for object control subset Scale of Infant and Toddler Development 3 rd edition ¹² ; BOT- on ¹⁴ ; MAND, McCarron Assessment of Neuromuscular Devel otor Scales 2 nd edition ¹⁷ ;; TGMD-II, Test of Gross Motor Dev	Raw scores, standard scores, percentile rank, age equivalent, Gross Motor Quotient. Simple conversion. 2, Bruininks-Oseretsky Test of Motor Proficiency 2 nd edition opment ¹⁵ ; NSMDA, Neurological Sensory Motor Developme elopment 2 nd edition ¹⁸ ; GM, Gross Motor; FM, Fine Motor;	Simple to administer ¹³ ; MABC-2, Movement Assessment Batte ental Assessment ¹⁶ ; PDMS-2, Peabody NDI, Neurodevelopmental Index

. . . . d a dual a lateration of ~ مام **c**

to beet terien only

Page 43 of 45



PRISMA 2009 Checklist

4 5 Se	ction/topic	#	Checklist item	Reported on page #
7 TIT	ſLE			
9 Title	e	1	Identify the report as a systematic review, meta-analysis, or both.	1 Title page
10 AB	ABSTRACT			
12 Stru 13 14	uctured 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
15 INT	TRODUCTION			
16 17 Rat	tionale	3	Describe the rationale for the review in the context of what is already known.	4-5
18 Obj 19	jectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
20 21 ME	THODS			
22 Pro 23	otocol 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.	-
²⁴ Elig 25	gibility 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
27 Infc 28	ormation 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
²⁹ Sea 30 31	arch	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
32 Stu 33	idy 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
34 35 Dat 36	ta 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
37 Dat 38	ta 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
³⁹ Ris 40 stuo	k of bias in individual dies	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
42 Sur	mmary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	7
43 Syr 44 45	nthesis 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

Page	- 1	of	2
i auc		UI.	_

4 -	4 Page 1 of 2			
5 6 7	Section/topic	#	Checklist item	Reported on page #
/ 8 9	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
10	Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	-
13	RESULTS			
14 15 16	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
17 18 19 20 21	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
22	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)
24 25 26 27 28 30 31 32 33 34 35 36	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
37 38	Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	-
39	Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	-
40 41	Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	-
42 ⊿3	DISCUSSION			
44 45 46	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., inclutionarie providers, jusers part opolicy indexes) juidelines.xhtml	22-25
.0				

Page 45 of 45

BMJ Open



PRISMA 2009 Checklist

3				
4 5	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
6 7 8	Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	26
9	FUNDING			
10 11 12	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
13 14	<i>From:</i> Moher D, Liberati A, Tetzlaff doi:10.1371/journal.pmed1000097	J, Altma	an DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS	Med 6(7): e1000097.
15	,		For more information, visit: www.prisma-statement.org.	
16				
17	7		Page 2 of 2	
18	3			
19)			
20)			
21				
22	2			
23	8			
24	ł			
25				
26				
27	7			
28	3			
29)			
30)			
31				
32	2			
33	8			
34	ł			
35	5			
36				
37	7			
38	3			
39)			
40)			
41				
42	2			
43	8			
44	ŀ			
45			For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml	
46				
47	7			