PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Psychometric properties of gross motor assessment tools for children: a systematic review
AUTHORS	Griffiths, Alison; Toovey, Rachel; Morgan, Prue; Spittle, Alicia

VERSION 1 – REVIEW

REVIEWER	M.M. Schoemaker University of Groningen, University Medical Center Groningen, Center for Human Movement Sciences
REVIEW RETURNED	06-Feb-2018

GENERAL COMMENTS	Psychometric properties of gross motor assessment tools for children: a systematic review. The aim of the review is to systematically evaluate the psychometric properties and clinical utility of gross motor assessment tools for children 2-12 years of age. A review of the psychometric property and clinical utility of gross motor tests can be very useful for clinicians and researchers. Overall, the most important qualities of the included tests have been assessed in this paper, and the methodological quality has also been thoroughly assessed. My compliments to the authors for all the work they have done to present the data in a systematic way. A former review of gross motor tests has already been conducted in 2010 by Slater et al. In addition, another review has been conducted in 2009 (Cools, W., Martelaer, K. D., Samaey, C., & Andries, C. (2009). Movement Skill Assessment of Typically Developing Preschool Children: A Review of Seven Movement Skill Assessment Tools. Journal of Sports Science & Medicine, 8(2), 154–168.) However, this review was not referred to in the paper (which raises some doubts about the thoroughness of the literature selection process). If a new review has been done within 8 years of a former review, the authors need to make very clear what their review adds to the existing literature. They give several arguments in the introduction (the MABC2 was not reviewed by Slater at al, and a lot of papers have been published about the psychometric properties of tests). However, in my view, the paper can be further improved, by discussing more clearly in the discussion, what kind of new evidence has been gathered since the last review. In addition, some tests were included in both the present and 2010 review, but there is also a difference between both papers in the inclusion of tests (ZNA, TMP). Why? Apart from this, there are several inconsistencies in the text, that need to be corrected before publication. 1. According to figure 1, Seven tests have been included in the review. However, on page 7 line 2, the authors report that 8 assessment tools were identified. Also in the discussion (page 16, lines 29 and 30), eight tests are mentioned. In addition, according

	figure 1, 20 assessment tools were excluded. However, we read in the text on page 7, line 8, that 19 assessment tools were excluded. 2. Supplementary table 1 provides a list of measurement properties, On page 6, the authors report that 9 measurement properties will be evaluated. However, in Supplementary table 1 only 7 measurement properties are defined (cross-cultural validity and hypothesis testing are lacking). In addition, construct validity is called structural validity in Table 3. Also: in Supplementary table 5, concurrent and predictive validity are reported. Both properties have not been introduced before in Supplementary Table 1. 3. The numbers in Supplementary Table 4 do not correspond to the right papers (For instance in the row MABC-2, reference [36] should be [31]. I did not check all references, but several others are incorrect too, also in other tables. Introduction: well written Method: Titles and abstract were screened by one author, who also excluded assessment tools which did not meet the inclusion criteria. This could have created a bias, one person might miss relevant papers (for instance the review mentioned above). I would suggest to discuss this as a limitation of the study. Results: • Page 7, line 19: the PDMS-2 and the MAND also include strength assessments. However, also the BOT-2 includes strength assessment. • The TGMD-2 assesses quality of performance. It would be helpful to mention this in the text. • Page 10, line 14: 'The TGMD-2 which has 10 items' ◊ the TGMD-2 has 12 items. • Page 10, line 13: The shortest duration time is for the TGMD-2. However, assessment with the MAND takes the same amount of time (Table 2). • Page 11, line 2: the MABC-2 and the TGMD-2 have the most evidence for construct validity. It would be helpful if the authors would make a distinction between what aspect of construct validity has been investigated. For instance, only one paper addressed factor analysis of the MABC-2. More research is definitely needed. • Page 11, line 4: discriminant validity ◊ is not mentioned before as a measurement property. • Page 16: Reliability of the Bailey is not discussed. • In table 6 for some measures sensitivity and specificity are reported. These findings are not mentioned in the results or discussion. Discussion: • My main concern is the discussion of the results. The present discussion is a collection of short loose paragraphs. I miss a structured discussion about 'What's new" in this review compared to the 2010 review. In addition, a more systematic overview of the results in the discussion would be helpful: which tests are reliable, and which aspects of reliability still need to be investigated? For each aspect of validity: which test demonstrates good validity, and what needs to be investigated in future studies. I do realize that the information is in the tables, but the authors now leave it to the reader to draw conclusions. They do conclude that the available tools demonstrate adequate validity and reliability. Yet, when I scan the tables, some aspects of validity have hardly been assessed for some tests (for instance concurrent validity of the MAND and Bailey). The authors have done such a good job in gathering all this information and composing all the tables, that it is a pity not to

	discuss the results in a more thorough and constructive way. • The discussion also lacks depth at some places. For instance: page 17, line 7: “all assessment tools were found to have merits and limitations….” What merits and limitations? Are the findings in line with the purposes of a test (according to table 1)?

REVIEWER	rameckers Eugene Maastricht University and Hasselt University Netherlands and Belgium
REVIEW RETURNED	21-Mar-2018

GENERAL COMMENTS	I have a few questions and hope they can be addressed easily: Based on the title to review gross motor assessment tools I am surprised to see that the NSMDA is included. NSMDA just partly assess gross motor function but far more development including several different domains. Otherwise the title is not correct and should be gross, fine motor and development assessment Please clarify how this NSMDA fit in the inclusion. My advice would be a 3 step inclusion of the data: 1. Inclusion of the assessment as you did in this paper 2. Inclusion of only combined gross and fine motor assessments – excluding other area’s - 3. Inclusion of only gross motor assessments – excluding fine motor items- The tables and conclusions will be the same, and you can add/ mark this 3 step choices. I really think it will help the clinicians to be decide what assessment is needed based on the question in motor assessments. Furthermore your discussion is exactly about these topics and could be more structured related to the 3 step question – gross motor, gross + fine motor and gross + other domains assessments This leads to the second question: Please create a decision scheme for choosing a most adequate assessment based on the COSMIN grading and the clinical question and order the assessments accordingly. This will add an extremely important guideline for clinicians and help both researchers and clinicians to choose adequately.

REVIEWER	Corinna Gerber Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland
REVIEW RETURNED	03-Apr-2018

GENERAL COMMENTS	In general, this is a very nice paper and there is a lot of work behind this manuscript. The topic is actual, interesting, and important. The language is adequate and the paper is very well written and understandable. However, there are some important methodological issues and other minor comments to be addressed. General: Very thorough review with many results. I’d suggest or to split the review in two parts (two publications – e.g. validity and rest of psychometric properties) or to include some of the supplementary tables as normal tables as they contain important study results and not only supplementary information. Number of assessment tools included is not consistent in the

	manuscript – sometimes authors write 8 assessment tools were included and 19 excluded but in the figure 27-20 -> 7 assessment tools. Please clarify and correct. Introduction: Well-structured and complete. Page 4, line 13: I'd suggest: "…effects of an intervention." OR "…effects of interventions." Page 5, lines 11-14: To be able to evaluate critically clinical utility and psychometric properties from studies, the evaluation of the methodological quality of these studies is crucial. Therefore, at least part of the secondary aim should be included in the primary aim of the study. Methods: Clearly written but with some information missing. See also comments below. However, it would have helped for overall interpretability of the results to perform a best evidence synthesis (see for example Terwee et al. "Quality criteria were proposed for measurement properties of health status questionnaires" and van Tulder et al. "Updated method guidelines for systematic reviews in the Cochrane collaboration back review group"). Page 5, line 19: ' missing before motor activity' Page 5 lines 25 ff: It becomes not very clear how you extracted the tools because following the methods you screened the abstracts only after this step. Wouldn't it have made sense to do a two-step search as we did in our work ("Reliability and Responsiveness of Upper Limb Motor Assessments for Children with Central Neuromotor Disorders: A Systematic Review" Gerber et al. 2016) or Elvrum et al did in their systematic review (Outcome measures evaluating hand function in children with bilateral cerebral palsy: a systematic review)? Page 6, line 4: Why did authors include "questionnaires" in the search terms when questionnaires were excluded from the study? Page 6, line 5: Authors write they did exclude assessment tools "…only applicable to children with a specific diagnosis". Just to be sure, I didn't misunderstand: Authors exclude tools if they were developed for a specific diagnosis but, in theory, could be used with children of other diagnosis (but for the other diagnosis, there was no evidence for validity at this time)? Page 6, lines 26-27: Clinical utility involves more than only what has been listed by the authors (see for example "A multi-dimensional model of clinical utility" by A. Smart 2006). I understand the choice of the authors but would suggest to re-formulate the sentence (e.g. authors chose cost of manual, kits, training requirements, time to administer the assessment and the ease of scoring as the important points of clinical utility of an assessment tool) or to explain the term clinical utility and that authors chose only part of the model "clinical utility". Page 6, lines 27-28: I don't really understand this sentence. Were all

	values for reliability reported as ICC? If so, this should rather be stated in the results section. And is there an “and” missing before “…directly compared”? Supplementary Table 2: If I search with your search terms (Medline), papers including the Gross Motor Function Classification System, the Pediatric Balance Scale, and the Gross Motor Function Measure result in the search. They probably might not fit your inclusion criteria for assessment tools (especially the classification system) but I do not understand why they do not figure in the excluded assessments. Thank you for clarification. Results: Text is clearly written and summarizes the most important findings. However, there are a lot of results and it is quite difficult to keep the overview. Especially because there is no direct link made between methodological quality of papers and evidence of psychometric properties. A best evidence synthesis (see comment for methods section) would help to gather the information and help the interpretability of the results. Tables are well structured and complete. However, some of the supplementary tables are essential for the study and thus should be provided as “normal” tables in the study results section. Discussion: In general well written. However, there is a lack of discussion of study limitations – for example, authors could discuss their methodology with the one of other systematic reviews on outcome measures in a pediatric (e.g. Ammann-Reiffer et al 2014, Elvrum et al 2016, or Gerber et al 2016).

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: M.M. Schoemaker
Institution and Country: University of Groningen, University Medical Center Groningen, Center for Human Movement Sciences
Please state any competing interests or state ‘None declared’: None declared

1. A former review of gross motor tests has already been conducted in 2010 by Slater et al. In addition, another review has been conducted in 2009 (Cools, W., Martelaer, K. D., Samaey, C., & Andries, C. (2009). Movement Skill Assessment of Typically Developing Preschool Children: A Review of Seven Movement Skill Assessment Tools. Journal of Sports Science & Medicine, 8(2), 154–168.) However, this review was not referred to in the paper (which raises some doubts about the thoroughness of the literature selection process). If a new review has been done within 8 years of a former review, the authors need to make very clear what their review adds to the existing literature. They give several arguments in the introduction (the MABC2 was not reviewed by Slater at al, and a lot of papers have been published about the psychometric properties of tests). However, in my view, the paper can be further improved, by discussing more clearly in the discussion, what kind of new evidence has been gathered since the last review.

We have clarified the key changes in evidence since the 2010 review have added a discussion of the Cools et al paper on page 22 lines 17-27 and page 24 line 20:

"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….."

2. In addition, some tests were included in both the present and 2010 review, but there is also a difference between both papers in the inclusion of tests (ZNA, TMP). Why?

In order to thoroughly assess the tests, including the manuals we were restricted to those with manuals published in English. For example the ZNA did not have an English manual available at the time of assessment. The Cools et al (2009) paper had an emphasis on assessment in an educational context, whilst Slater et al (2010) were only looking at assessments for DCD, only included assessments with 50% or greater gross motor component and only for children over 4 years of age. The different inclusion/exclusion criteria as noted above and in our paper account for the differences of included tests between the papers.

We have added to the discussion that a limitation of this review is that we only included assessments tools published in English, which excluded some assessments used commonly in Europe (page 24 lines 25-29). The rationale behind the exclusion of assessments is also demonstrated in Figure 1.

3. There are several inconsistencies in the text that need to be corrected before publication.
1. According to figure 1, Seven tests have been included in the review. However, on page 7 line 2, the authors report that 8 assessment tools were identified. Also in the discussion (page 16, lines 29 and 30), eight tests are mentioned. In addition, according figure 1, 20 assessment tools were excluded. However, we read in the text on page 7, line 8, that 19 assessment tools were excluded.

These errors have been amended to reflect the seven assessment tools throughout the paper, in the results on page 7 line 9 and 15, as well as in the discussion on page 22 lines 2 and 3.

4. Supplementary table 1 provides a list of measurement properties, On page 6, the authors report that 9 measurement properties will be evaluated. However, in Supplementary table 1 only 7 measurement properties are defined (cross-cultural validity and hypothesis testing are lacking). In addition, construct validity is called structural validity in Table 3. Also: in Supplementary table 5, concurrent and predictive validity are reported. Both properties have not been introduced before in Supplementary Table 1.

Thank you for drawing our attention to these inconsistencies. This has been amended to reflect the 7 measurement properties in the COSMIN checklist, with cross cultural validity, hypothesis testing and structural validity considered an aspect of construct validity (page 6, lines 20-21). We have removed the reference to concurrent validity in supplementary table 1, and added definitions for cross cultural validity, structural validity and hypothesis testing.

We have also added a reference to predictive validity on page 6, lines 21-23:

"Whilst predictive validity is considered to be a component of content validity, it is reported on separately in this paper for interpretability of results[7]."

5. The numbers in Supplementary Table 4 do not correspond to the right papers (For instance in the row MABC-2, reference [36] should be [31]. I did not check all references, but several others are incorrect too, also in other tables.

These have been amended.

6. Method: Titles and abstract were screened by one author, who also excluded assessment tools which did not meet the inclusion criteria. This could have created a bias, one person might miss

relevant papers (for instance the review mentioned above). I would suggest to discuss this as a limitation of the study.

We have added a discussion about this limitation on page 24, line 22-24.

7. Results:
• Page 7, line 19: the PDMS-2 and the MAND also include strength assessments. However, also the BOT-2 includes strength assessment.

Thank-you for bringing our attention to this oversight, the BOT-2 has been included in this list on page 8, line 2.

8. The TGMD-2 assesses quality of performance. It would be helpful to mention this in the text.

This has been added (page 11, lines 4-7): "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."

9. Page 10, line 14: 'The TGMD-2 which has 10 items' à the TGMD-2 has 12 items.

We have reworded this sentence for readability, which also meant removing the reference to the number of items on the TGMD-2.

10. Page 10, line 13: The shortest duration time is for the TGMD-2. However, assessment with the MAND takes the same amount of time (Table 2).

We have added the MAND (page 11, line 17)

11. Page 11, line 2: the MABC-2 and the TGMD-2 have the most evidence for construct validity. It would be helpful if the authors would make a distinction between what aspect of construct validity has been investigated. For instance, only one paper addressed factor analysis of the MABC-2. More research is definitely needed.

Thank you for the suggestion. We have clarified this on page 12, lines 8-14 "The TGMD-2 has the most evidence for construct validity with several papers performing confirmatory and exploratory factor analysis [19 2018 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."

12. Page 11, line 4: discriminant validity à is not mentioned before as a measurement property.

This has been reworded for clarity on page 12 lines 12-14: "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."

13. Page 16: Reliability of the Bailey is not discussed.

We have added some comments on the reliability of the Bayley-III on page 17 lines 9-10 ("the Bayley-III is shown to have excellent internal consistency in children aged 24-42 months"), 17 ("Test-retest reliability was excellent in the Bayley-III (Table 6)….."), and 22 ("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…").

14. In table 6 for some measures sensitivity and specificity are reported. These findings are not mentioned in the results or discussion.

We have added a comment in the results (page 17 lines 29-30) and in the discussion (page 24, lines 1-3)

15. My main concern is the discussion of the results. The present discussion is a collection of short loose paragraphs. I miss a structured discussion about 'What's new" in this review compared to the 2010 review. In addition, a more systematic overview of the results in the discussion would be helpful: which tests are reliable, and which aspects of reliability still need to be investigated? For each aspect of validity: which test demonstrates good validity, and what needs to be investigated in future studies. I do realize that the information is in the tables, but the authors now leave it to the reader to draw conclusions. They do conclude that the available tools demonstrate adequate validity and reliability. Yet, when I scan the tables, some aspects of validity have hardly been assessed for some tests (for instance concurrent validity of the MAND and Bailey). The authors have done such a good job in gathering all this information and composing all the tables, that it is a pity not to discuss the results in a more thorough and constructive way.
• The discussion also lacks depth at some places. For instance: page 17, line 7: "all assessment tools were found to have merits and limitations…." What merits and limitations? Arethe findings in line with the purposes of a test (according to table 1)?

We have added to and modified the existing information on "what's new" in the discussion page 22 lines 15-27: "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."

We also added a more in depth discussion of the validity and reliability of the tests on pages 22 and 23:

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

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

the gross motor subsections of the Bayley-III and the NSMDA with the other assessment tools in this study in the future.

The BOT-2 was the only assessment tool to have its reliability assessed with excellent methodology. In conjunction with its reported resultsit 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."

Reviewer: 2
Reviewer Name: rameckers eugene
Institution and Country: Maastricht University and Hasselt University
Netherlands and Belgium
Please state any competing interests or state 'None declared': none

1. Based on the title to review gross motor assessment tools I am surprised to see that the NSMDA is included. NSMDA just partly assess gross motor function but far more development including several different domains. Otherwise the title is not correct and should be gross, fine motor and development assessment
Please clarify how this NSMDA fit in the inclusion.

Based on our inclusion criteria developmental assessments (such as the NSMDA or the Bayleys) were included if they were able to provide a meaningful gross motor subscore. While we agree that the NSMDA would not be selected for a gross motor assessment in isolation, it is useful to understand its psychometric properties and how the gross motor domain compares with other gross motor assessments.

2. My advice would be a 3 step inclusion of the data:
1. Inclusion of the assessment as you did in this paper
2. Inclusion of only combined gross and fine motor assessments – excluding other area's -
3. Inclusion of only gross motor assessments – excluding fine motor items-
The tables and conclusions will be the same, and you can add/ mark this 3 step choices.
I really think it will help the clinicians to be decide what assessment is needed based on the question in motor assessments.
Furthermore your discussion is exactly about these topics and could be more structured related to the 3 step question – gross motor, gross + fine motor and gross + other domains assessments

Thank you for the suggestion. We have emphasised these differences on page 7, lines 22-25 and in the discussion on page 22, lines 2-8. We hope that clinicians and researchers will refer to the clinical utility table (Table 2) to narrow down the selection of their assessment tool based on the domains included (gross, gross + fine, gross + developmental) as well as the age of the participant as the first stage of tool selection. Our view is that the body of the discussion needs to emphasise the evidence and methodological rigour of the psychometric properties, to guide this second, often more complex stage of assessment tool selection.

3, Please create a decision scheme for choosing a most adequate assessment based on the COSMIN grading and the clinical question and order the assessments accordingly.
This will add an extremely important guideline for clinicians and help both researchers and clinicians to choose adequately.

Thank-you for your suggestion, while we agree that this would be a very valuable tool for clinicians we feel that it is beyond the scope of this paper due to the complexity of the decision making process and

variety of clinical questions and contexts requiring consideration. We will consider this as an idea for a future paper.

Reviewer: 3
Reviewer Name: Corinna Gerber
Institution and Country: Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland
Please state any competing interests or state 'None declared': None declared

1. I'd suggest or to split the review in two parts (two publications – e.g. validity and rest of psychometric properties) or to include some of the supplementary tables as normal tables as they contain important study results and not only supplementary information.

Thank you for the suggestion. We have embedded the supplementary tables for validity and reliability into the paper as Tables 4, 5 and 6 on pages 18-21, rather than splitting into two papers.

2. Number of assessment tools included is not consistent in the manuscript – sometimes authors write 8 assessment tools were included and 19 excluded but in the figure 27-20 -> 7 assessment tools. Please clarify and correct.

Apologies for this oversight, this has been amended throughout the text.

3. Page 4, line 13: I'd suggest: "…effects of an intervention." OR "…effects of interventions."

This has been changed to "…effects of interventions." Thank you for the suggestion.

4. Page 5, lines 11-14: To be able to evaluate critically clinical utility and psychometric properties from studies, the evaluation of the methodological quality of these studies is crucial. Therefore, at least part of the secondary aim should be included in the primary aim of the study.

We agree and have amended the aims to reflect this (page 5, lines 13-15)

5. Methods: Clearly written but with some information missing. See also comments below. However, it would have helped for overall interpretability of the results to perform a best evidence synthesis (seefor example Terwee et al. "Quality criteria were proposed for measurement properties of health status questionnaires" and van Tulder et al. "Updated method guidelines for systematic reviews in the Cochrane collaboration back review group").

While the authors understand the reasoning for this suggestion, we feel that the addition of any more tables would make this paper unreasonably long. We have added further discussion to clarify the key points on page 22 lines 22-28 and page 23.

6. Page 5, line 19: ' missing before motor activity'

This has been amended.

7. Page 5 lines 25 ff: It becomes not very clear how you extracted the tools because following the methods you screened the abstracts only after this step. Wouldn't it have made sense to do a two-step search as we did in our work ("Reliability and Responsiveness of Upper Limb Motor Assessments for Children with Central Neuromotor Disorders: A Systematic Review" Gerber et al. 2016) or Elvrum et al did in their systematic review (Outcome measures evaluating hand function in children with bilateral cerebral palsy: a systematic review)?

We have clarified the methodology on page 6 lines 10-11, and in Figure 1. Any assessment tools known not to fit the inclusion criteria (such as the GMFM) were excluded during the screening of abstracts by the first author. We would certainly consider a two-step search for future reviews.

8. Page 6, line 4: Why did authors include "questionnaires" in the search terms when questionnaires were excluded from the study?

This is a MESH search term that we were advised may include assessment tools relevant to our search.

9. Page 6, line 5: Authors write they did exclude assessment tools "…only applicable to children with a specific diagnosis". Just to be sure, I didn't misunderstand: Authors exclude tools if they were developed for a specific diagnosis but, in theory, could be used with children of other diagnosis (but for the other diagnosis, there was no evidence for validity at this time)?

Yes, we excluded tools that were developed for a specific diagnosis.

10. Page 6, lines 26-27: Clinical utility involves more than only what has been listed by the authors (see for example "A multi-dimensional model of clinical utility" by A. Smart 2006). I understand the choice of the authors but would suggest to re-formulate thesentence (e.g. authors chose cost of manual, kits, training requirements, time to administer the assessment and the ease of scoring as the important points of clinical utility of an assessment tool) or to explain the term clinical utility and that authors chose only part of the model "clinical utility".

The wording has been changed accordingly (page 7 line 3-4): "Items chosen to represent the clinical utility of the assessment tools were…"

11. Page 6, lines 27-28: I don't really understand this sentence. Were all values for reliability reported as ICC? If so, this should rather be stated in the results section. And is there an "and" missing before "…directly compared"?

All reported values for reliability (Kappa, ICC etc) were collected, however only the ICCs were directly compared (in Figure 2). We have reworded this sentence for clarity (page 7, line 5-7): "All reported values for reliability were collected, however, only those papers reporting anIntraclass Correlation Coefficient (ICC) were directly compared."

12. Supplementary Table 2: If I search with your search terms (Medline), papers including the Gross Motor Function Classification System, the Pediatric Balance Scale, and the Gross Motor Function Measure result in the search. They probably might not fit your inclusion criteria for assessment tools (especially the classification system) but I do not understand why they do not figure in the excluded assessments. Thank you for clarification.

We have clarified the methodology in the paper (page 6 lines 10-11), and in Figure 1 in response to your comment above and at the start of your feedback on our methods. These assessment tools did not fit the inclusion criteria during the screening of abstracts, and as such were not listed with the assessment tools that were excluded during screening of the full texts.

13. Results: Text is clearly written and summarizes the most important findings. However, there are a lot of results and it is quite difficult to keep the overview. Especially because there is no direct link made between methodological quality of papers and evidence of psychometric properties. A best evidence synthesis (see comment for methods section) would help to gather the information and help the interpretability of the results.
Tables are well structured and complete. However, some of the supplementary tables are essential for the study and thus should be provided as "normal" tables in the study results section.

We have now embedded the supplementary tables for validity and reliability into the paper as Tables 4, 5 and 6 on pages 18-21. The authors are concerned that the addition of any more tables would make this paper unreasonably long. Instead, we have added further discussion highlighting the key points regarding methodological quality of papers and psychometric properties on page 22 lines 29-30 and page 23 lines 1-28.

14. Discussion: In general well written. However, there is a lack of discussion of study limitations – for example, authors could discuss their methodology with the one of other systematic reviews on outcome measures in a pediatric (e.g. Ammann-Reiffer et al 2014, Elvrum et al 2016, or Gerber et al 2016).

We have added a discussion of the study limitations on page 24 lines 22-26:

“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.”

VERSION 2 – REVIEW

REVIEWER	Marina M. Schoemaker Centre for Human Movement Science, University Medical Centre, Groningen, The Netherlands
REVIEW RETURNED	21-Jun-2018

GENERAL COMMENTS	The authors changed the paper in line with the comments I made. I noticed 4 typo's: Page 22, line 24: 'Papers that had were given.....' Please check sentence, words are lacking or words should be deleted. Page 23, line 4: 'of the each test's' --> 'of each test' Page 23 , line 24: 'Test-rest' --> 'Test-retest' Page 24, line 18: 'All of the included assessment tools .....' Please check sentence, words are lacking.

REVIEWER	Corinna Gerber Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland
REVIEW RETURNED	08-Jul-2018

GENERAL COMMENTS	In general, the review, especially the discussion, has improved considerably. However, there are still some major methodological issues to be discussed. Methods: General: The methodology of a first screening of title and abstracts with exclusion of some of the tools and papers that then do not figure in the excluded tools/papers is very unusual and does not follow guidelines for a systematic search/review. Furthermore, the fact that two steps were combined in one (exclusion of papers not fitting and exclusion of assessment tools not fitting) is a major weakness of this paper. This have to be clearly pointed out in the limitations section. Page 6, lines 22-23: I don’t understand what authors want to say with “…it is reported on separately in this paper for interpretability of results“ (maybe you meant “… is reported separately in this paper…” OR is reported on ? in this paper…”?) Supplementary table 1: In your end search you have 21 and 22 and 23 and 24 but in search 21 you only have „reproducibility of results“ don’t you exclude validity with this search strategy?

	Results: Incongruence between methods and results on what is considered as gross motor function. In the methods section authors write: "Discriminative, predictive or evaluative of gross motor skills, 2. Assessed ≥ two gross motor (e.g. balance, jumping etc.)..." thus, balance is considered a gross motor function. But then in the results section they write: "The other gross motor assessments were either in conjunction with assessment of fine motor and/or balance..." thus balance is not considered a gross motor function. Discussion: Page 24, lines 22: The limitation is not that much that only one author did the first inclusion/exclusion but that this process has been separated from the second inclusion/exclusion and thus, for example, tools excluded in this step are not reported as excluded tools. Authors should better compare with high quality systematic reviews what their limitations are, and state their limitations based on these reflections. Page 22, line 24: "Papers that had were given lower methodological scores..." please correct. Page 23, line 24: Please correct "Test-rest, inter or intra-rater…"

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Marina M. Schoemaker

Institution and Country: Centre for Human Movement Science, University Medical Centre, Groningen, The Netherlands

Typographical errors

1. Page 22, line 24: 'Papers that had were given.....' Please check sentence, words are lacking or words should be deleted.

This has been changed to "Papers that received lower methodological scores..." (page 22, line 24).

2. Page 23, line 4: 'of the each test's' --> 'of each test'

This has been changed to "of each tests'…" (PubMed Page 23, line 4)

3. Page 23 , line 24: 'Test-rest' --> 'Test-retest'

This has been changed to "test-retest, inter or intra-rater reliability…"

4. Page 24, line 18: 'All of the included assessment tools .....' Please check sentence, words are lacking.

This paragraph has been modified – with part of it deleted (page 24 lines 17-19) and part of it moved to the conclusion (Page 26 lines 20-22)

Reviewer: 3
Reviewer Name: Corinna Gerber
Institution and Country: Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland

Methods:
1. General: The methodology of a first screening of title and abstracts with exclusion of some of the tools and papers that then do not figure in the excluded tools/papers is very unusual and does not follow guidelines for a systematic search/review. Furthermore, the fact that two steps were combined in one (exclusion of papers not fitting and exclusion of assessment tools not fitting) is a major weakness of this paper. This have to be clearly pointed out in the limitations section.

Thank you for this feedback. The authors feel that the exclusion of papers and assessment tools that clearly did not meet the inclusion criteria during screening of titles and abstracts is in keeping with the Cochrane and COSMIN guidelines. The COSMIN guidelines suggest documenting the total number of abstracts selected, the total number of full-text articles that were selected, and the main reasons for excluding full-text articles, which we have done. While some systematic reviews of outcome measures perform a two-tiered search, the search strategy presented by Terwee, et al. [1] is, as we have performed, a single tiered search. Furthermore, our flowchart is based on the PRISMA and COSMIN guidelines which do not require a rationale for papers/assessment tools excluded during the screening phase. An additional comment elaborating on the search process has been included in the limitations section (page 24, lines 18-20).

2. Page 6, lines 22-23: I don't understand what authors want to say with "…it is reported on separately in this paper for interpretability of results" (maybe you meant "… is reported separately in this paper…" OR is reported on ? in this paper…"?)

This has been changed to "… it is reported separately" on page 6 line 22

3. Supplementary table 1:
In your end search you have 21 and 22 and 23 and 24 but in search 21 you only have „reproducibility of results" don't you exclude validity with this search strategy?

The MESH search term "reproducibility of results" encompasses the following terms, thus including both validity and reliability in the search strategy:

- Reproducibility of Findings
- Reliability of Results
- Reliability (Epidemiology)
- Validity (Epidemiology)
- Validity of Results
- Face Validity

- Validity, Face
- Reliability and Validity
- Validity and Reliability
- Test-Retest Reliability
- Reliabilities, Test-Retest
- Reliability, Test-Retest
- Test Retest Reliability

Results:
4. Incongruence between methods and results on what is considered as gross motor function. In the methods section authors write: “Discriminative, predictive or evaluative of gross motor skills, 2. Assessed ≥ two gross motor (e.g. balance, jumping etc.)...“ thus, balance is considered a gross motor function. But then in the results section they write: “The other gross motor assessments were either in conjunction with assessment of fine motor and/or balance...“ thus balance is not considered a gross motor function.

Thank you for this comment. Balance is considered separate to gross motor skills in this paper. To clarify and more accurately reflect this, the wording on page 5 line 27 has been changed to “Assessed ≥ two gross motor (e.g. hopping, jumping etc)…..”

Discussion:
5. Page 24, lines 22: The limitation is not that much that only one author did the first inclusion/exclusion but that this process has been separated from the second inclusion/exclusion and thus, for example, tools excluded in this step are not reported as excluded tools. Authors should better compare with high quality systematic reviews what their limitations are, and state their limitations based on these reflections.

Thank you for this comment. We respectfully refer the reviewer back to our comments above in regards to our methodology. The first exclusions of papers/assessment tools was during the screening of the titles and abstracts, following the COSMIN guidelines.

6. Page 22, line 24: “Papers that had were given lower methodological scores...“ please correct.

This has been changed to “Papers that received lower methodological scores on the COSMIN” (page 22, line 24)

7. Page 23, line 24: Please correct “Test-rest, inter or intra-rater…”

This has been changed to “Test-retest, inter or intra-rater…” (page 23, line 24)

Reference List

1. Terwee CB, Jansma EP, Riphagen II, et al. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research*2009;18(8):1115-23. doi: 10.1007/s11136-009-9528-5