PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

Title (Provisional)

Psychometric properties of early childhood development assessment tools in low-and-middle-income countries: a systematic review

Authors

Bliznashka, Lilia; Hentschel, Elizabeth; Ali, Nazia Binte; Hunt, Xanthe; Neville, Sarah Elizabeth; Olney, D.; Pitchik, Helen O; Roy, Aditi; Seiden, Jonathan; Solís-Cordero, Katherine; Thapa, Aradhana; Jeong, Joshua

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Brinkman, Sally**
Affiliation	**University of Adelaide CHRD, Public Health**
Date	**29-Oct-2024**
COI	**I developed one of the child development instruments included in the review.**

The premise of the paper needs to be clear upfront. It needs to be clear that you are seeking to promote instruments that can/aim to be used for cross country comparison.

Within the paper it needs to be stated that if an instrument shows poor invariance across countries (or there is no evidence either way) but shows good invariance within a country then it's likely to still be a good instrument for that country to use. Ultimately it is a choice for stakeholders - do they care about cross country comparison? Often not, and if not, then within country validity is more important. Some commentary on this would be prudent.

Why wouldn't you include RCTs where an ECD instrument was the primary outcome.

If the RCT showed impact, then this inherently let's us know that the instrument is sensitive to change, which is a crucial aspect of validity and utility in it's own right. Further, if an ECD measure was used at baseline and those results then predict later outcomes then this provides evidence of predictive validity - often trials report the relationship between baseline and midline/endline measures. I understand that the authors may not be keen on adding more papers to their review at this stage - and if not, then the exclusion of RCTs and their potential value to the validity results needs to be included within the limitations section in the discussion.

It would be of value to add to the table that defines the different aspects of validity a column on why that aspect is important, so readers can understand the relevance of each aspect of validity to their own circumstance/intent of use.

Please provide more detail on why/how bias is associated with training in the Introduction. This may not be obvious to some readers.

It would be worth undertaking a grammatical check before publication. I am not noting all errors here, but just for example in the Abstract - Objective - "To systematically appraised" should be To systematically appraise. There are quite a few edits like this worth picking up before publication.

Key messages, What is already known on this topic: I don't understand the value/point of the second dot point.

Introduction, second paragraph: Directly applying HIC instruments in LMICs is more problematic than just psychometrics - it is also about relevance i.e. congruous with local values. This deserves some attention.

Introduction, last paragraph: The reasons why different tools have differing levels of psychometric evidence published has little to do with the quality of the instrument, but more to do with the time it's been in the field, the financial support backing it's development/use and the number of western academics involved. This needs to be deeply reflected upon within the paper. So, instead of stating that stakeholders can use the results of this paper to select instruments - and thus perpetuating existing global biases - it would be more appropriate to promote the value of greater supports for validation work of existing and new instruments in LMICs. Maybe even reflect on the facilitators and barriers to validation work in LMICs.

I would suggest deleting all reference to invariance over time - this makes little sense for measures of ECD - why even mention it. The relevance of time invariance relates to the study design - if longitudinal, or an RCT, for example, then it is irrelevant as children age through the study. So this is a far more nuanced discussion that either needs a lot more commentary, or just simply delete.

VERSION 1 - AUTHOR RESPONSE

Reviewer 1

We thank the reviewer for their time and thoughtful comments to improve the paper. Our point-by-point responses follow with line numbers referring to the tracked changes version.

The premise of the paper needs to be clear upfront. It needs to be clear that you are seeking to promote instruments that can/aim to be used for cross country comparison.

Thank you for raising this concern. We have clarified in the Introduction that: *"We sought to deepen our understanding and consistently summarise whether psychometric evidence exists. Our findings can assist stakeholders in selection of ECD tools with adequate psychometric evidence for the intended use, including for comparison between countries, and inform what research is needed to improve how we track ECD-related SDGs, programmes, and policies in LMICs." (58-62)*

In the Discussion, we highlight the importance of cross country comparisons on lines 189-193: *"While several studies reported on samples drawn from multiple countries, few conducted statistical analysis to test for equivalence across countries, thus providing no evidence of measurement invariance.*[8] *Cross-country invariance, which guarantees assessment of the same construct across countries, is key for tracking global SDG goals."*

Lastly in the Conclusion we note that the results can help stakeholders not only select tools with evidence for the country, but also select tools that can be used for cross-country comparison: *"Nevertheless, the results by country and ECD tool presented here can serve stakeholders in selecting tools with at least some available psychometric evidence, and tools that can be used for cross-country comparison."* (lines 257-259)

Within the paper it needs to be stated that if an instrument shows poor invariance across countries (or there is no evidence either way) but shows good invariance within a country then it's likely to still be a good instrument for that country to use. Ultimately it is a choice for stakeholders - do they care about cross country comparison? Often not, and if not, then within country validity is more important. Some commentary on this would be prudent.

Thanks for raising this point, we have reflected in on lines 193-197: *“Nevertheless, if an ECD tool shows poor invariance or lack of invariance across countries, it can still have strong psychometric properties and be useful for a specific country if the purpose of measurement is not cross-country comparison. In practice, when selecting an ECD tool, stakeholders should consider such potential trade-offs in light of their objectives and purpose of measurement.”*

Why wouldn't you include RCTs where an ECD instrument was the primary outcome. If the RCT showed impact, then this inherently let's us know that the instrument is sensitive to change, which is a crucial aspect of validity and utility in it's own right. Further, if an ECD measure was used at baseline and those results then predict later outcomes then this provides evidence of predictive validity - often trials report the relationship between baseline and midline/endline measures. I understand that the authors may not be keen on adding more papers to their review at this stage - and if not, then the exclusion of RCTs and their potential value to the validity results needs to be included within the limitations section in the discussion.

Thanks for raising this concern. First, we have clarified in the methods that we excluded any study where ECD was the primary outcome, not just RCTs (lines 76-78): *“We excluded articles where a tool was used to assess an outcome measure (e.g., trials reporting impacts on ECD outcomes, cross-sectional studies examining ECD predictors), but the article did not include measurement objectives.[17]”*

Then, we clarify in the limitations why we could not include studies were ECD was an outcome (lines 243-248): *“Further, by excluding trials where ECD was the primary outcome we could not take stock of the evidence on whether tools were sensitive to change. This is an important psychometric property, albeit beyond the scope of the current paper. To adequately capture all studies where ECD was an outcome and assess evidence on sensitivity to change, a different and more extended search strategy should have been used as the measurement properties search filter we used was not designed for this purpose[17].”*

It would be of value to add to the table that defines the different aspects of validity a column on why that aspect is important, so readers can understand the relevance of each aspect of validity to their own circumstance/intent of use.

Thanks for this suggestion. We have added the column to Table 1. For brevity, please refer to the revised manuscript.

Please provide more detail on why/how bias is associated with training in the Introduction. This may not be obvious to some readers.

Thanks for this suggestion. We have clarified on lines 94-95: *"Poor or inconsistent training can result in different assessors administering the ECD tool in different ways, which undermines the accuracy and consistency of the scores obtained."*

It would be worth undertaking a grammatical check before publication. I am not noting all errors here, but just for example in the Abstract - Objective - "To systematically appraised" should be To systematically appraise. There are quite a few edits like this worth picking up before publication.

Thanks for catching this typo. We have done a grammatic check of the manuscript to eliminate similar issues.

Key messages, What is already known on this topic: I don't understand the value/point of the second dot point.

We have removed this bullet since this section is no longer required.

Introduction, second paragraph: Directly applying HIC instruments in LMICs is more problematic than just psychometrics - it is also about relevance i.e. congruous with local values. This deserves some attention.

Thanks for raising this point. We agree and have reflected it in the Introduction (lines 37-39): *"Directly applying ECD tools from HICs in LMICs can be problematic without psychometric evidence for new cultures and contexts, and without considering local norms and values."*

Introduction, last paragraph: The reasons why different tools have differing levels of psychometric evidence published has little to do with the quality of the instrument, but more to do with the time it's been in the field, the financial support backing it's development/use and the number of western academics involved. This needs to be deeply reflected upon within the paper. So, instead of stating that stakeholders can use the results of this paper to select instruments - and thus perpetuating existing global biases - it would be more appropriate to promote the value of greater supports for validation work of existing and new instruments in LMICs. Maybe even reflect on the facilitators

and barriers to validation work in LMICs.

We absolutely agree with you on the points raised here. We have now clarified in the Introduction that *"We sought to deepen our understanding and consistently summarise whether psychometric evidence exists, rather than what evidence exists."* (lines 58-59).

Further in the Discussion, we highlight that *"This fragmentation is evidenced here by included articles focusing on individual countries, limited age ranges, and single developmental domains. At least partially, this fragmentation may be due to the time ECD tools have been in the field and the financial support for their development and use."* (lines 165-168). We also note that *"Understanding the facilitators and barriers of measurement work can help foster greater support for this type of research in LMICs."* (207-208). However, we do not reflect on specific facilitators or barriers to validation work as this was beyond the scope of the review.

Lastly, we highlighted that *"Further, more work is needed to understand publication bias and whether the amount of psychometric evidence published is associated with the quality of the ECD instrument. Future studies should consider reviewing grey literature as well, which may contribute valuable information on the psychometric quality of ECD tools." (220-223)* While we agree with you that the level/amount of psychometric evidence is often unrelated to the quality of the tool, it is also plausible that publication bias has prevented authors from publishing null findings or findings of poor psychometric properties. We therefore recommend more work to adequately address this question.

I would suggest deleting all reference to invariance over time - this makes little sense for measures of ECD - why even mention it. The relevance of time invariance relates to the study design - if longitudinal, or an RCT, for example, then it is irrelevant as children age through the study. So this is a far more nuanced discussion that either needs a lot more commentary, or just simply delete.

We opted to keep the invariance over time and results discussion. We believe it is an important psychometric property that is very understudied. To your point above, we have distinguished measurement invariance over time from sensitivity to capture change in ECD. Invariance ensures that the same construct is measured over time, which is irrespective of children's age or whether the tool is sensitive to change. ECD tools contain different items for different age groups precisely because children age and different skills become relevant. However, the underlying construct/concept remains the same. If the tool measures a different construct over time, then longitudinal results on intervention impacts are meaningless and cannot be attributed solely to the intervention but rather also to a change in what is measured. For these reasons, we have kept measurement invariance over time. If the reviewers and editor feel strongly that the paper would be much improved by removing it, we kindly ask that you let us know and we can revise the manuscript.

VERSION 2 - REVIEW

Reviewer	**1**
Name	**Brinkman, Sally**
Affiliation	**University of Adelaide CHRD, Public Health**
Date	**11-Feb-2025**
COI	

Thank you for addressing the initial concerns. The revisions have resulted in a more balanced presentation of the paper.

I would still recommend either removing the measurement invariance over time or significantly expanding the commentary around it. While I understand your perspective, establishing measurement invariance over time in our field presents substantial challenges. Developmental constructs can evolve in both form and function over time, with the extent of this variation depending on the specific domain under consideration. For instance, literacy skills tend to follow a relatively predictable cumulative trajectory, making them more amenable to measurement invariance testing. In contrast, social and emotional constructs are inherently more complex due to heterotypic continuity. Additionally, milestone- or screening-based instruments differ in fundamental ways from early childhood development (ECD) measures that assess both strengths and weaknesses. Consequently, the feasibility of demonstrating measurement invariance is influenced by factors such as the breadth and scope of the instrument. In summary, my primary concern is that this issue is more nuanced than currently described in the paper.

Finally, regarding my initial review comments on potential biases related to funding, duration in the field, and the individuals involved in instrument development—while I appreciate that this has been addressed to some extent, I believe further clarity is needed. Specifically, the absence of published psychometric evidence for an instrument should not be equated with inferior quality compared to those with documented evidence. While greater psychometric validation is certainly necessary, and this is a valid conclusion, encouraging readers to select instruments based solely on the reported results moves beyond scientific analysis into advocacy. Striking the right balance between these perspectives is challenging, but I believe a more measured and nuanced conclusion would be appropriate and more helpful for readers.

Reviewer	**2**
Name	**Spittle, Alicia**
Affiliation	**University of Melbourne, Physiotherapy**
Date	**27-Jan-2025**
COI	**None**

This is the first time I have reviewed this manuscript and have looked at the previous reviewers comments. In general, as per previous reviewers comments, it's important to clarify that measures validated in HMIC countries may be applicable for use in LMIC but why it's important to understand the psychometric properties in LMIC. This needs to come out in the abstract background. It has been adequately addressed in the main paper. It's also not clear how many assessment tools were included, rather the number of articles. For clinicians, researchers, policy makers – they will want to know which tools have the best psychometrics not just a general number of articles on each area.

Abstract:

1. From the results it appears there are many articles that examine psychometric properties, yet the conclusion is "Psychometric evidence is fragmented, limited, and heterogeneous." Can the results section be more specific (if the word count allows) as to why this is limited or what tools had the best evidence.

Introduction

1. Line 59, page 5. "We reviewed available evidence on 11 psychometric properties of tools used to assess ECD in children 0-6 years old living in LMICs". I would like to know which psychometrics and why there were chosen.

Methods:

1. Page 7, line 92: Why did you use your own risk of bias checklist and not a validated checklist such as COSMIN (which is included in your references)?

Results

1. Would benefit from subheadings to guide the reader as to which aspect you are reporting on (e.g. country, age, etc)

Discussion

1. As above, it's more important to focus on the tool and then whether it has psychometric properties, rather than the number of articles on an area. It's the quality of the evidence

rather than number of articles that's important. You need to guide the reader as to which tools they can use, why, which context.

VERSION 2 - AUTHOR RESPONSE

REVIEWER 1

Thank you for your time and for reviewing the paper twice. We appreciate your comments and thoughtful suggestions for improving our paper. Our point-by-point responses follow. Line numbers refer to the clean version of the manuscript.

Thank you for addressing the initial concerns. The revisions have resulted in a more balanced presentation of the paper.

I would still recommend either removing the measurement invariance over time or significantly expanding the commentary around it. While I understand your perspective, establishing measurement invariance over time in our field presents substantial challenges. Developmental constructs can evolve in both form and function over time, with the extent of this variation depending on the specific domain under consideration. For instance, literacy skills tend to follow a relatively predictable cumulative trajectory, making them more amenable to measurement invariance testing. In contrast, social and emotional constructs are inherently more complex due to heterotypic continuity. Additionally, milestone- or screening-based instruments differ in fundamental ways from early childhood development (ECD) measures that assess both strengths and weaknesses. Consequently, the feasibility of demonstrating measurement invariance is influenced by factors such as the breadth and scope of the instrument. In summary, my primary concern is that this issue is more nuanced than currently described in the paper.

> Thank you for this thoughtful suggestion. We agree with you on the difficulties of establishing measurement invariance over time in our field. We have now removed measurement invariance over time from the paper. Even after updating the search, there was only one article providing evidence on measurement invariance over time. As a result, we thought it more appropriate to remove the topic from the paper rather than to expand on it. Overall, this did not change any other results or conclusions presented in the paper since the one article (Chen et al. 2017) reporting on measurement invariance over time also reported on structural validity and remained included. For brevity, please refer to the revised manuscript, tables and figures, and supplement.

Finally, regarding my initial review comments on potential biases related to funding, duration in the field, and the individuals involved in instrument development—while I appreciate that this has been addressed to some extent, I believe further clarity is needed. Specifically, the absence of published psychometric evidence for an instrument should not be equated with inferior quality compared to those with documented evidence. While greater psychometric validation is certainly necessary, and this is a valid conclusion, encouraging readers to select instruments based solely on the reported results moves beyond scientific analysis into advocacy. Striking the right balance between these perspectives is challenging, but I believe a more measured and nuanced conclusion would be appropriate and more helpful for readers.

Thank you for reiterating this point. We have revised the Discussion and Conclusion to provide a more nuanced commentary. We agree with you the availability of psychometric evidence may be related to factors that are unrelated to the quality of the tool itself, and have noted this on lines 290-293: *"Likewise, ample psychometric evidence was available for ECD tools that have been implemented for longer duration (e.g., ASQ and BSID), or have had more available funding (e.g., GSED). As a result, the quantity of psychometric evidence available should not be the criterion used to determine the psychometric quality or useability of an ECD tool."*

We also more clearly state that the availability of psychometric evidence should not be equated with the quality of psychometric evidence (lines 298-304): *"In addition, since our review focused on whether psychometric evidence exists, our findings on the availability of psychometric evidence do not inform our understanding of the underlying quality, strength, or rigor of the psychometric evidence. An important next step in this line of work is to fully unpack the utility of existing psychometric evidence. The results of psychometric analyses along with other characteristics of an ECD tool (e.g., domain assessed, age range, and administration time and cost among others) should be used to determine the most relevant ECD tool for the given context."*

We have also revised the conclusion to clarify that psychometric evidence should be considered as only one of several selection criteria for ECD tools. We also recognize that more evidence is needed beyond what we present in the paper: *"Psychometric evidence on ECD tools used in LMICs is fragmented, limited, and heterogeneous. More research is warranted to establish the applicability of existing tools in diverse populations, including urban and rural settings, and on establishing measurement invariance over countries. Nevertheless, the results by country and ECD tool presented here can serve stakeholders by providing a database of available psychometric evidence for ECD tools in LMICs. To improve monitoring, evaluation, and accountability for ECD globally, psychometric evidence should be a key consideration when selecting ECD tools together with other important consideration including the purpose of measurement, available resources for training and administration, and the population and developmental domain of interest. As psychometric properties can vary by geography, population, and age, among other characteristics, greater psychometric validation can help facilitate ECD tool selection across diverse contexts in LMICs. Improved reporting for psychometric studies can help ensure transparency, replication, and adequate ability to assess the quality of evidence."* (lines 319-331)

REVIEWER 2

Thank you for your time and thoughtful suggestions for improving our paper. Our point-by-point responses follow. Line numbers refer to the clean version of the manuscript.

This is the first time I have reviewed this manuscript and have looked at the previous reviewers comments. In general, as per previous reviewers comments, it's important to clarify that measures validated in HMIC countries may be applicable for use in LMIC but why it's important to understand the psychometric properties in LMIC. This needs to come out in the abstract background. It has been adequately addressed in the main paper. It's also not clear how many assessment tools were included, rather the number of articles. For clinicians, researchers, policy makers – they will want to know which tools have the best psychometrics not just a general number of articles on each area.

Thanks for raising these important points. We have clarified in the abstract (lines 3-6) that: *"Although ECD tools developed in high-income countries may be applicable to low- and middle-income countries (LMICs), directly applying them in LMICs can be problematic without psychometric evidence for new cultures and contexts."*

We have also reported the number of tools include:

- In the abstract (line 19): *"A total of 160 articles covering 117 tools met inclusion criteria."*
- In the results (line 173): *"We included 160 articles, covering 117 tools."*
- In the discussion (lines 228-229): *"Based on 160 articles, available evidence on 10 psychometric properties of 117 ECD tools for children 0-6 years old is fragmented, limited, and heterogeneous."*

Abstract:

1. From the results it appears there are many articles that examine psychometric properties, yet the conclusion is "Psychometric evidence is fragmented, limited, and heterogeneous." Can the results section be more specific (if the word count allows) as to why this is limited or what tools had the best evidence.

Thanks for raising these points. We did not have the space to include these points in the abstract. We have included a paragraph in the Results describing the psychometric evidence available by ECD tool (lines 199-207): *"For development articles, a single article reported psychometric properties for all tools except for the CREDI covered in two articles (Supplemental Table 7). For adaptation articles, ASQ-3 and BSID-III were most often studied (14 and 12 articles, respectively). For 75% of tools only a single article provided psychometric evidence (Supplemental Table 8). Other frequently studied tools included the Mullen Scales of Early Learning (n=5 articles), the Alberta Infant Motor Scale (n=4 articles), the Bayley Infant Neurodevelopmental Screener (n=4 articles), Denver Developmental Screening Test (Denver-II, n=4 articles), and International Development and Early Learning Assessment (IDELA, n=4 articles) (Supplemental Table 5)."*

In the discussion, we have commented on why we interpret results as fragmented, limited, and heterogenous:

- *"Our findings support ECD measurement trends found in other work:*[27] much *of the work on psychometric properties is recent, with ECD tools being developed and adapted concurrently. Psychometric efforts remain limited to a few ECD tools,*[28,29] *individual ECD domains,*[12–14] *and few psychometric properties.*[10,12] *This fragmentation is evidenced by included articles focusing on individual countries, limited age ranges, and single developmental domains. In addition, included studies focused on individual reliability or validity properties (e.g., internal consistency reliability and concurrent reliability, respectively), thus providing a limited picture of reliability and validity as a whole. The resulting heterogeneous psychometric evidence can hinder comparability and large-scale monitoring of ECD policies and programmes within and across LMICs, which is crucial for identifying and implementing effective approaches to support ECD."* (lines 235-244)

- *"This review highlights four important limitations of existing psychometric evidence for ECD tools in LMICs. First, although most tools are designed for a wide age range, the psychometric evidence behind most tools pertained to narrower age ranges and in some cases as narrow as one month. This may have limited applicability to diverse age ranges (given that there is a natural variability in child development in the early years[13]) in these specific contexts. Relatedly, most psychometric evidence pertained to urban contexts. Given existing urban-rural disparities in ECD[15,16] and increasingly diverse young child populations in urban settings,[13] ECD tools whose psychometric properties were examined only in urban settings might be inadequate for rural settings."* (lines 251-259)
- *"Third, consistent with existing literature, we found limited psychometric evidence on tools measuring socio-emotional and personal-social development.[13,19] This is surprising given that these two domains are among the most culturally specific,[13] implying they require more comprehensive and rigorous adaptation. In addition, psychometric evidence on tools to assess attention/executive function and academic/preacademic development was very limited. Without additional work to establish a psychometric base, this poses major challenges for those seeking to monitor these domains in early life."* (lines 270-276)

Introduction

1. Line 59, page 5. "We reviewed available evidence on 11 psychometric properties of tools used to assess ECD in children 0-6 years old living in LMICs". I would like to know which psychometrics and why there were chosen.

Thanks for raising this point. The list of psychometric properties and their definitions are provided in Table 1. Please note that we have reduced it to 10 based on Dr Brinkman's suggestion (see her comments and our response below). We have reported in the methods how these psychometric properties were chosen (lines 96-98): *"The 10 psychometric properties were selected based on prior systematic reviews on psychometrics properties of ECD tools,[9,10,12,18,19] classical test theory,[20] and reviews of measurement in cross-cultural psychology.[21,22]"*

Methods:

1. Page 7, line 92: Why did you use your own risk of bias checklist and not a validated checklist such as COSMIN (which is included in your references)?

Thank you for raising this point. Our objective as stated in the abstract and introduction was to review available evidence on 10 psychometric properties of tools used to assess ECD in children 0-6 years old living in LMICs. We have clarified that our objective was to understand *whether* psychometric evidence exists, not *what* psychometric evidence exists (lines 72-73): *"We sought to deepen our understanding and consistently summarise whether psychometric evidence exists for tools used to measure ECD in LMICs."*

At the start of the work, we reviewed available tools to assess risk of bias of psychometric studies, including the one developed by COSMIN. We also consulted with a psychometrician. Since we aimed to assess the underlying quality of psychometric studies rather than the psychometric properties themselves, none of the existing tools were deemed adequate. As a result, we developed our own checklist which includes key aspects that may bias psychometric studies. We have now clarified this in the methods (lines 135-137): *"Since we aimed to assess the quality of the underlying studies, rather than the quality of the psychometric properties, we could not use a validated risk of bias tool and instead developed a new one."*

We recognize that this can be perceived as a limitation of our study, and have acknowledge it in the "Strengths & Limitations section" (lines 40-41): *"We did not use a validated tool to assess risk of bias, which may limit the comparability of our findings."* and the Discussion (lines 312-313): *"Further, we did not use a validated risk of bias tool, which may limit the comparability of our findings."*

Results

1. Would benefit from subheadings to guide the reader as to which aspect you are reporting on (e.g. country, age, etc)

Thanks for this suggestion! We have added subheadings to the results section.

Discussion

1. As above, it's more important to focus on the tool and then whether it has psychometric properties, rather than the number of articles on an area. It's the quality of the evidence rather than number of articles that's important. You need to guide the reader as to which tools they can use, why, which context.

Thank you for raising this point. We have clarified in the introduction that our objective was to understand *whether* psychometric evidence exists, not *what* psychometric evidence exists (lines 72-73): *"We sought to deepen our understanding and consistently summarise whether psychometric evidence exists for tools used to measure ECD in LMICs."* While we agree with you that the quality of psychometric properties is important, evaluating it was beyond the scope of this review. We acknowledge that this an important line of future research on lines 298-304: *"In addition, since our review focused on whether psychometric evidence exists, our findings on the availability of psychometric evidence do not inform our understanding of the underlying quality, strength, or rigor of the psychometric evidence. An important next step in this line of work is to fully unpack the utility of existing psychometric evidence. The results of psychometric analyses along with other characteristics of an ECD tool (e.g., domain assessed, age range, and administration time and cost among others) should be used to determine the most relevant ECD tool for the given context."*

Guidance on the selection of ECD tools already exists (lines 60-61): *"Prior reviews have provided guidance for selecting ECD tools for use in LMICs[5,9–11] and underscored that evidence on tool reliability and validity is fundamental.[9,10]"* We aimed to further strengthen this guidance by summarising available psychometric evidence by various characteristics that implementers, researchers, and policy makers may consider in selecting an ECD tool based on their aims and objectives. As psychometric results for each tool can vary by geography, population, and child age, among other characteristics, our review was meant to provide a comprehensive list of available psychometric evidence that can be reviewed by stakeholders when determining which tool to use in their context, as well as highlight gaps where more psychometric evidence is needed to allow for the appropriate selection of ECD tools across diverse contexts. We have stated this in the conclusion on lines 322-329: *"Nevertheless, the results by country and ECD tool presented here can serve stakeholders by providing a database of available psychometric evidence for ECD tools in LMICs. To improve monitoring, evaluation, and accountability for ECD globally, psychometric evidence should be a key consideration when selecting ECD tools together with other important consideration including the purpose of measurement, available resources for training and administration, and the population and developmental domain of interest. As psychometric properties can vary by geography, population, and age, among other characteristics, greater psychometric validation can help facilitate ECD tool selection across diverse contexts in LMICs."* Given these considerations, we did not consider ourselves in a position to recommend specific tools.

VERSION 3 - REVIEW

Reviewer	**1**
Name	**Brinkman, Sally**
Affiliation	**University of Adelaide CHRD, Public Health**
Date	**07-Apr-2025**
COI	

No further comments

Reviewer	**2**
Name	**Spittle, Alicia**
Affiliation	**University of Melbourne, Physiotherapy**
Date	**24-Apr-2025**
COI	

Thank you for your response to my comments.

I think it would be worth clarifying in the abstract and conclusion of the paper that there is a large volume of tools already in existence and the quality of the tools needs to be further examined rather than new tools.