




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

Lilia Bliznashka ^{1,2}, Elizabeth Hentschel,³ Nazia Binte Ali,³ Xanthe Hunt,^{4,5} Sarah Elizabeth Neville,⁶ Deanna Olney,¹ Helen O Pitchik ⁷, Aditi Roy,⁸ Jonathan Seiden,⁹ Katherine Solís-Cordero,¹⁰ Aradhana Thapa,¹¹ Joshua Jeong ¹¹

To cite: Bliznashka L, Hentschel E, Ali NB, *et al*. Psychometric properties of early childhood development assessment tools in low- and middle-income countries: a systematic review. *BMJ Open* 2025;**15**:e096365. doi:10.1136/bmjopen-2024-096365

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<https://doi.org/10.1136/bmjopen-2024-096365>).

Received 09 November 2024
Accepted 25 April 2025



© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to

Lilia Bliznashka;
l.bliznashka@cgiar.org

ABSTRACT

Objective Valid and reliable measurement of early childhood development (ECD) is critical for monitoring and evaluating ECD-related policies and programmes. 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. Our objective was to systematically appraise available evidence on the psychometric properties of tools used to measure ECD in LMIC.

Design A systematic review following the Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines.

Data sources MEDLINE, Embase, PubMed, PsycInfo, SciELO and BVS were searched from inception to February 2025.

Eligibility criteria We included studies that examined the reliability, validity, and measurement invariance of tools assessing ECD in children 0–6 years of age living in LMICs.

Data extraction and synthesis Each study was independently screened by two researchers and data extracted by one randomly assigned researcher. Risk of bias was assessed using a checklist developed by the study team assessing bias due to training/administration, selective reporting and missing data. Results were synthesised narratively by country, location, age group at assessment and developmental domain.

Results A total of 160 articles covering 117 tools met inclusion criteria. Most reported psychometric properties were internal consistency reliability (n=117, 64%), concurrent validity (n=81, 45%), convergent validity (n=74, 41%), test–retest reliability (n=73, 40%) and structural validity (n=72, 40%). Measurement invariance was least commonly reported (n=16, 9%). Most articles came from Brazil, China, India and South Africa. Most psychometric evidence was from urban (n=92, 51%) or urban–rural (n=41, 23%) contexts. Study samples focused on children aged 6–17.9 or 48–59.9 months. The most assessed developmental domains were language (n=111, 61%), motor (n=104, 57%) and cognitive (n=82, 45%). Bias due to missing data was most common.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ This systematic review used extensive search filters, made no restrictions on developmental domains and included both single- and multi-domain tools.
- ⇒ The review team was rigorously trained and brought diverse backgrounds in early childhood development measurement in low- and middle-income countries.
- ⇒ We could not conduct full-text review and extraction for articles in Chinese, Farsi and Turkish because these languages were not spoken by the review team.
- ⇒ We did not use a validated tool to assess risk of bias, which may limit the comparability of our findings.

Conclusions Psychometric evidence is fragmented, limited and heterogeneous. More rigorous psychometric analyses, especially on measurement invariance, are needed to establish the quality and accuracy of ECD tools for use in LMICs.

PROSPERO registration number CRD42022372305.

INTRODUCTION

The Sustainable Development Goals (SDGs) recognise the importance of early childhood development (ECD),¹ which sets the foundation for children's later learning and economic outcomes.^{2–4} Rigorous ECD measurement is critical for accurate comparability of children's skills across populations and time, evaluating the effectiveness of ECD interventions and tracking ECD-related policies.

In low- and middle-income countries (LMICs), >140 tools (60% of which originated in high-income countries (HICs)) have been used to assess ECD in children 0–8 years old.⁵ Although ECD tools developed in HICs may be applicable to LMICs, directly applying ECD tools from HICs in LMICs can

be problematic without psychometric evidence for new cultures and contexts. Few studies to date have synthesised the evidence on the psychometric properties of ECD tools.

Evidence pertaining to a tool's reliability, validity and measurement invariance in a given context is critical for selecting an ECD outcome or indicator. Other factors include the purpose of measurement, the population and age range of interest, the developmental domain(s) of interest and administration time and cost.⁶ Evidence of reliability and validity ensures consistent and accurate ECD measurement,⁷ whereas evidence of measurement invariance guarantees assessment of the same construct across countries and subgroups.^{7,8}

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} However, prior work has often focused on individual ECD domains,^{12–14} which has limited use for population level assessment or tracking of SDG-related policies, or a subset of psychometric properties,¹⁰ which is not evidence of reliability and validity as a whole. Furthermore, prior reviews do not disaggregate use of ECD tools by HICs versus LMICs or urban versus rural settings,^{12–14} despite known and persistent disparities in ECD by country income level and urban/rural residence^{15,16} and the fact that young child populations are increasingly diverse due to migration and urbanisation.¹³ These are important distinctions given that psychometric properties can vary by population characteristics.

We reviewed available evidence on 10 psychometric properties of tools used to assess ECD in children 0–6 years old living in LMICs. We summarised the current landscape by ECD tool, developmental domain, country and age group, focusing on 10 types of psychometric evidence. We sought to deepen our understanding and consistently summarise whether psychometric evidence exists for tools used to measure ECD in LMICs. Our findings can assist stakeholders in the selection of ECD tools for their intended use and context and inform what research is needed to improve how we track ECD-related SDGs, programmes and policies in LMICs.

METHODS

Search strategy and selection criteria

We identified articles in any language through MEDLINE, Embase, PubMed, PsycInfo, SciELO and BVS. The search strategy used medical subject headings (MeSH terms), keywords and free-text words along four key elements: population, construct, measurement properties and location (online supplemental table 1). Search terms were combined using Boolean operators. Truncation wildcards were used to include variations of the search terms. The search strategy used a combination of searches through titles, abstracts and keywords. The COnsensus-based Standards for the selection of health status Measurement

INstruments (COSMIN) measurement properties filter was used for the third search element.¹⁷ The search strategy was piloted in PubMed and then adapted for the remaining databases.

Full-text, peer-reviewed articles were included if: (1) the study was conducted in a LMIC, (2) included children 0–6 years old, (3) included at least one ECD domain (cognitive, language, motor, social-emotional, attention/executive function, personal-social and preacademic/academic) defined in online supplemental table 2, (4) developed a new tool or adapted an existing one, and provided primary evidence of at least one of 10 psychometric properties in terms of reliability, validity or measurement invariance (table 1) and (5) were published between 1 January 2007 and 9 March 2023. An updated search was conducted on 28 February 2025 and seven additional studies were identified for inclusion: five through the database searches and two from a Google Scholar alert. The 10 psychometric properties were selected based on prior systematic reviews on psychometric properties of ECD tools,^{9,10,12,18,19} classical test theory²⁰ and reviews of measurement in cross-cultural psychology.^{21,22} We considered all ECD tools regardless of their intended or actual use (diagnosis, screening or surveillance). We included both articles that adapted original tools ('adaptation articles' hereafter) and articles that developed new tools ('development articles' hereafter). Articles were included if at least 50% of the study sample was within the range 0–6 years and the average age in the sample was <6 years or unspecified. If information on the child age range was not available or clear, the article was included. Multi-country articles were included if ≥50% of the countries were LMICs. If an article reported that measurement properties were reported elsewhere (eg, referencing another study and thus providing secondary evidence), the article was ineligible. We included prospective, retrospective, cross-sectional and longitudinal quantitative study designs. Cited references were reviewed for potential inclusion.

We excluded articles where a tool was used to assess an outcome measure (eg, trials reporting impacts on ECD outcomes), but the article did not include measurement objectives.²³ Among articles with an explicit measurement focus, we excluded those reporting only on convergent validity (eg, correlations with sociodemographic variables) which alone provides limited psychometric evidence.²³ We excluded articles studying children with developmental disabilities and disorders (eg, autism spectrum disorder, cerebral palsy), and children with physical disabilities that impair performance on ECD measures (eg, deafness, blindness). Finally, we excluded the following study designs: animal studies, simulation studies, case studies, opinions, letters, preprints, protocols, conference abstracts, ecological

Table 1 Definitions of the psychometric properties used in the selection criteria for included articles

Domain	Measurement property	Example test statistic	Definition
Reliability			The consistency of a test or measurement, that is, how consistently a measure produces similar results with repeated measures over a short period of time or across assessors at the same time point. This can also be thought of as the correlation between observed scores across replications.
	Test–retest reliability	Correlation coefficient	Correlation between scores from the same test from assessments conducted over a short time interval.
	Inter-rater reliability	kappa, Bland-Altman test	The extent to which independent assessors produce similar ratings in judging the same abilities or characteristics in the same target person at the same time.
	Internal consistency reliability	Cronbach's alpha, alpha	Degree of interrelatedness among items on the same tool, that is, how well the items work together to provide information on an underlying construct.
Validity			The degree to which the tool measures what it is supposed to measure, that is, the degree to which the tool reflects the underlying construct.
	Content/face validity		The degree to which the content of the tool is adequate for the construct being measured, that is, assessing the extent to which a tool appears to reflect the underlying construct.
	Concurrent/criterion validity	Correlation coefficient; regression estimate	The degree to which scores on one measurement tool are related to scores obtained at about the same point in time from another tool considered the gold standard.
	Convergent validity	Correlation coefficient; regression estimate	Evidence that scores on a test or measurement are associated with theoretically related measures or variables.
	Predictive	Correlation coefficient; regression estimate	Evidence that a score correlates with a variable that can only be assessed at some point after the test has been administered or the measurement made, for example, evidence that scores now are correlated with scores at a future time point.
	Structural validity (dimensionality)	Exploratory factor analysis: number of factors, eigen values Confirmatory factor analysis: model fit statistics such as Comparative Fit Index, root mean square error of approximation	The degree to which the scores of an assessment are an adequate reflection of the dimensionality of the construct to be measured.
Invariance			The property when a scale or construct provides the same results across different samples, populations, settings or characteristics.
	Measurement invariance over countries	Likelihood ratio χ^2 statistic and p-value from freeing parameters across groups	The degree to which an assessment or construct provides the same results across separate samples in different countries.
	Measurement invariance over other groups	Likelihood ratio χ^2 statistic and p-value from freeing parameters across groups	The degree to which an assessment or construct provides the same results across different groups.
All definitions are based on the APA Dictionary of Psychology. ⁷			

studies, dissertations/theses, reviews or systematic reviews and meta-analyses.

Data extraction

Search results were imported into Covidence, where duplicates were automatically removed. Two reviewers independently screened titles and abstracts for inclusion and reviewed full texts of retained articles. Disagreements in screening or full-text review were resolved through discussion with a third reviewer.

One reviewer developed the data extraction sheet, and two reviewers piloted it. Revisions were made through discussion between the two reviewers. Seven

reviewers were trained on using the data extraction sheet through pilot extractions of included articles. Further revisions were made to the data extraction sheet based on discussions of the piloting process. Data extraction included information on publication details, study meta-data, characteristics of the ECD tool, type of administration, ECD domains assessed and psychometric properties with respect to reliability, validity and measurement invariance. After piloting, each article was extracted by one reviewer. For quality assurance, 20% of articles were randomly selected for independent extraction by a second

reviewer. Any discrepancies regarding data extraction were resolved by a third reviewer. Authors of included articles were not contacted when information was missing or unclear.

Risk of bias assessment

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. One reviewer created a risk of bias checklist by adapting items from the COSMIN²³ and Cochrane's ROBINS-E²⁴ risk of bias tools. The checklist was then refined via discussions with two other reviewers and a psychometrician, and after piloting by four reviewers. The final checklist contained three categories of bias due to: (1) training/administration (not assessed for tools relying on self-assessment), (2) selective reporting (only assessed for studies reporting on convergent or predictive validity) and (3) missing data. Each article was rated separately on each category using the ROBINS-E risk of bias ratings: low risk, some concerns, high risk and very high risk. We also assessed the indirectness of populations²⁵ by assessing whether the sample was limited to a specific setting, the tool covered the entire age range it was intended for, subgroups were generalisable and results were generalisable. Where insufficient information was provided in the article, we rated the study as 'unable to assess'. Seven reviewers were trained and conducted the risk of bias assessment at the article level. Disagreements for articles assessed by two reviewers (20%) were resolved through discussion with a third reviewer.

Data synthesis

Data analysis was conducted at the article-tool level because some articles reported on multiple tools. We created binary indicators for whether evidence on each one of the 10 psychometric properties was reported. In longitudinal studies, we considered the psychometric properties reported at any time point. We then summarised the evidence by country, location, age group at assessment, type of article (development vs adaptation), developmental domain and ECD tool. For all tools, child age at assessment was converted to months based on the information provided in the article. For longitudinal studies, age at first assessment was used. For multi-site/multi-country studies, the full age range at assessment was used across the sites/countries. Because the objective of the paper was to take stock of the available evidence rather than to report the adequacy and relevance of specific psychometric properties, we did not summarise evidence on the psychometric properties themselves. Results were synthesised narratively.

This review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines.²⁶

Patient and public involvement

Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

RESULTS

Study characteristics

Our search strategy identified 6430 records from six databases and two other sources (figure 1). After removing duplicates, 5338 records were excluded during title and abstract screening. After full-text review of the remaining 250 records, 97 records were excluded. After updating the search in February 2025, we included seven more articles (five identified from the six databases and two identified from other sources). We included 160 articles, covering 117 tools. Four articles reported on multiple tools, resulting in 182 article-tool combinations (referred to as articles for brevity).

Psychometric evidence available

Most articles were adaptation articles (n=145, 80%) (figure 2, online supplemental table 3). The most often evaluated psychometric properties were internal consistency reliability (n=117, 64%), concurrent validity (n=81, 45%), convergent validity (n=74, 41%), test-retest reliability (n=73, 40%) and structural validity (n=72, 40%). Measurement invariance was the least frequently reported psychometric property (n=16, 9%) and was primarily evaluated over countries and child sex. The number of articles increased over time (online supplemental figure 1). Between 2007 and 2010, most included articles were development articles; since 2011, most included articles have been adaptation articles.

Psychometric evidence available by country

Psychometric evidence came from 55 countries (figure 3); one-third of articles came from Brazil (n=22, 12%), China (n=18, 10%), India (n=14, 8%) and South Africa (n=10, 5%). For 40% of countries represented, there was only one article reporting psychometric evidence (online supplemental table 4). Psychometric evidence for each ECD tool included was generally limited to 1–2 countries with a few notable exceptions: the Global Scales for Early Development (GSED, 32 countries), the Caregiver Reported Early Development Instrument (CREDI, 18 countries), the International Development and Early Learning Assessment (IDELA, 17 countries), the Bayley Scales of Infant and Toddler Development, Third Edition (BSID-III, 12 countries) and the Ages and Stages Questionnaire-3 (ASQ-3, 11 countries) (online supplemental table 5).

Psychometric evidence available by location

Most psychometric evidence came from urban (n=92, 51%) or urban-rural (n=41, 23%) settings. In South

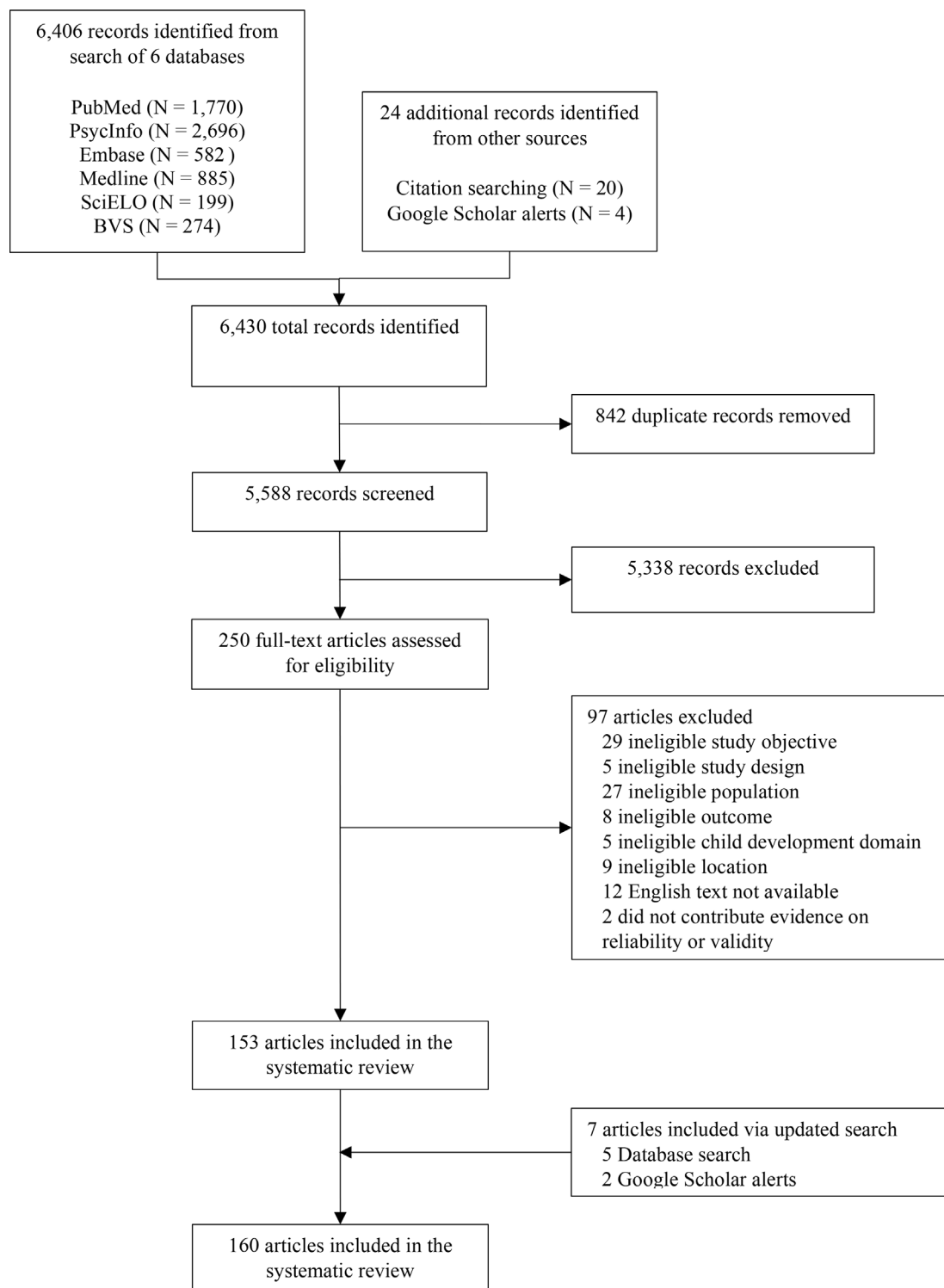


Figure 1 Preferred Reporting Items for Systematic reviews and Meta-analyses (PRISMA) flow diagram of search results and included articles.

Asia and sub-Saharan Africa, a similar number of articles originated from urban, rural and urban–rural settings (online supplemental table 6).

Psychometric evidence available by ECD tool

For development articles, a single article reported psychometric properties for all tools except for the CREDI covered in two articles (online 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 (online 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

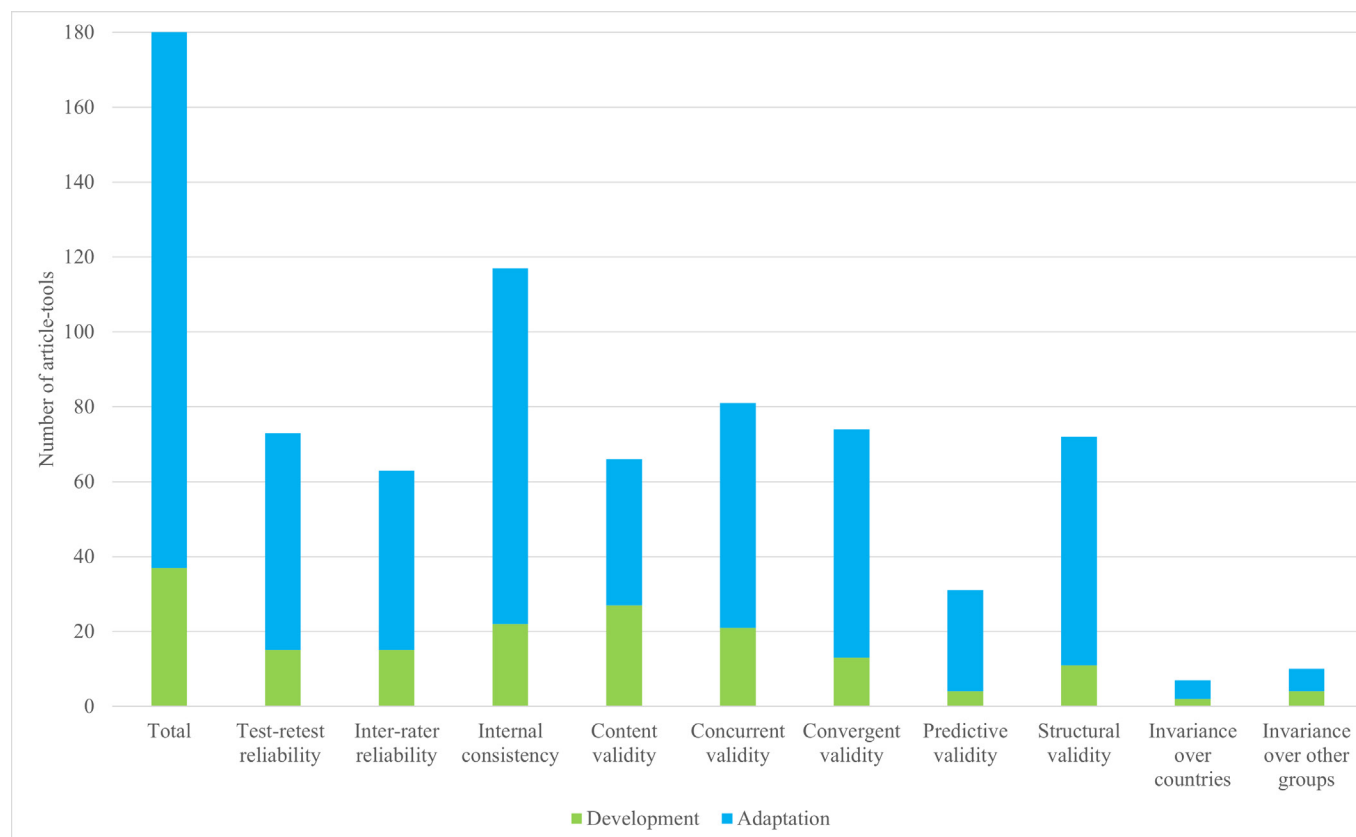


Figure 2 Number of included article-tools and type of psychometric evidence provided.

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) (online supplemental table 5).

Psychometric evidence available by age group

Although included tools targeted 0–71.9 month-old children, most studies focused on children 12–17.9 month (n=86, 47%), 6–11.9 months (n=82, 45%) or

54–59.9 months (n=76, 42%) old. Articles largely did not cover the full age groups defined here, or the full age range the ECD tool can be used for. Some articles included age groups as narrow as 1 month (online supplemental figure 2 and 3).

Psychometric evidence available by developmental domain

Most articles reported on language (n=111, 61%), motor (n=104, 57%) and cognitive (n=82, 45%) development (online supplemental table 9). Academic/

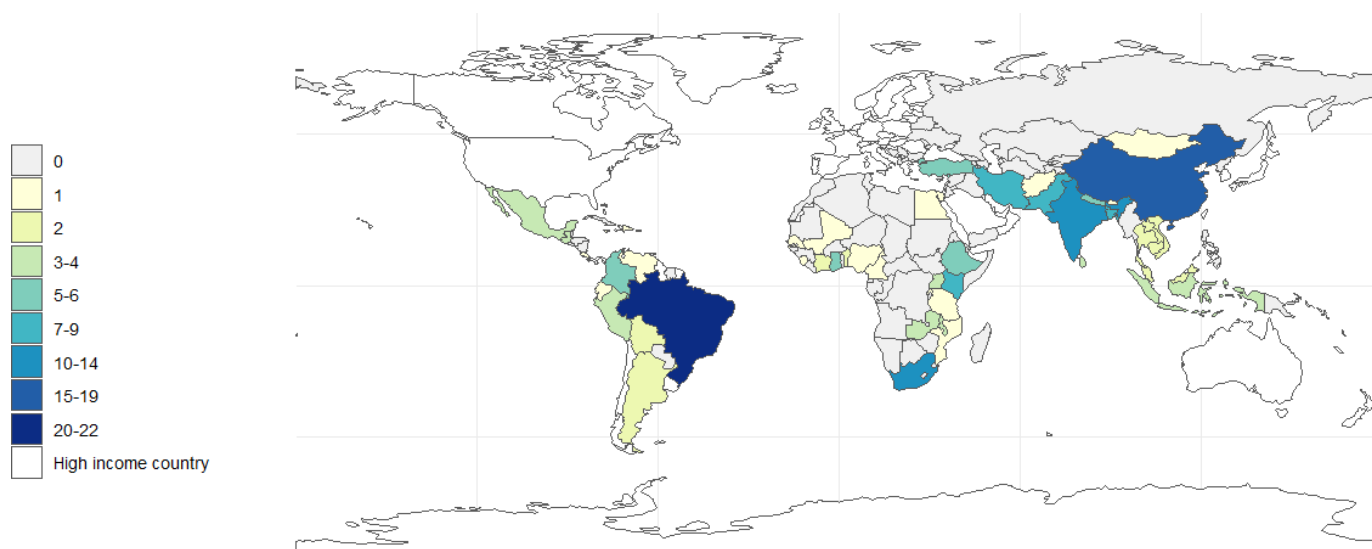


Figure 3 Countries where studies providing evidence on at least one psychometric property were conducted.

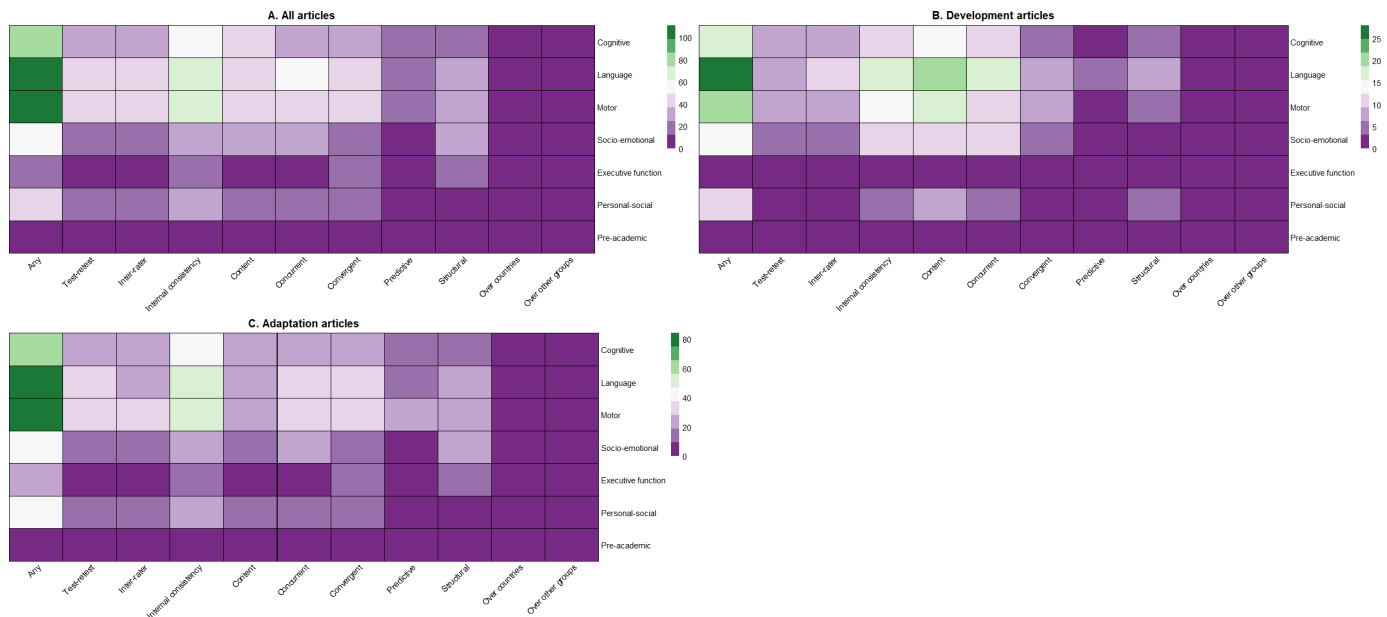


Figure 4 Number of articles providing psychometric evidence by article type and early childhood development domain.

preacademic was the least studied domain ($n=10$, 5%). Within domains, internal consistency reliability was most frequently reported (figure 4).

Risk of bias assessment

Bias due to training/administration was assessed for 59 articles (32%). Of these, 80% had low risk of bias ($n=47$) and 20% some concerns ($n=12$) (online supplemental table 10 and 11). Risk of bias due to selective reporting was assessed for 95 articles (52%). Of these, most articles ($n=86$, 91%) had low risk of bias. Risk of bias due to missing data was assessed for 137 articles (75%). Of these, 48% had some concerns ($n=52$) and 5% high risk ($n=7$). For indirectness, 20 articles (11%) were rated as generalisable and 82 (45%) as probably generalisable.

DISCUSSION

Based on 160 articles, available evidence on 10 psychometric properties of 117 ECD tools for children 0–6 years old in LMICs is fragmented, limited and heterogeneous. The most frequently provided evidence was on internal consistency reliability, test–retest reliability and/or concurrent, convergent and structural validity. Psychometric evidence on measurement invariance was the least commonly available. Although evidence came from 55 LMICs, four countries were most represented. Most evidence came from urban or urban–rural settings. The most studied tools were ASQ-3 and BSID-III.

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 (eg, 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.

Despite efforts to consolidate ECD measurement through tools like CREDI, GSED and IDELA, such tools do not fully meet research, programmatic and policy needs as evidenced by the increase of adaptation and development articles since 2015. This is not surprising given that no ECD tool is suitable for all populations.¹⁸ Our disaggregation by development and adaptation articles permitted a better understanding of the psychometric evidence available and highlighted evidence gaps for both existing and newly developed tools.

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 1 month. This may have limited applicability to diverse age ranges (given that there is a natural variability in child development in the early years¹³ 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. Those developing or adapting ECD tools should consider establishing psychometric properties across the full intended child age range and across both urban and rural settings.

Although this implies longer, more expensive and logistically difficult studies, it would ensure that the tool can be used broadly and in more diverse populations.

A second limitation is the very little evidence on measurement invariance. 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. More work in this domain is needed, particularly for tools widely used for policy making and programme evaluations. Measurement work should be considered relative to competing and more urgent ECD priorities in LMICs.

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 that 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.

Finally, in contrast to prior reviews which assessed the quality of psychometric properties using COSMIN guidelines,^{11 12} we assessed the risk of bias and quality of the underlying studies themselves. We observed a common lack of reporting and transparency in the training of assessors and data management. Risk of bias could not be assessed for many included studies or was considered high. Better reporting standards and guidelines for psychometric studies can help strengthen the field and ensure that more critical evaluation of the evidence is possible.

Nevertheless, some ECD tools had multiple forms of psychometric properties assessed. Using IDELA as an example: one article examined three types of reliability and one type of validity, drawing on pilots from 12 countries,²⁷ followed by another article examining measurement invariance across countries.³⁰ Subsequent articles have examined additional psychometric properties, although in individual countries.^{31 32} Although such examples of consecutively examining psychometric properties should be the norm, they are often financially and logistically infeasible, and the timing does not always align with programmatic and policy agendas. Likewise, ample psychometric evidence was available for ECD tools that have been implemented for longer duration (eg, ASQ and BSID), or have had more available funding (eg, GSED). As a result, the quantity of psychometric evidence available should not be the criterion used to determine the psychometric quality or usability of an ECD tool. When prior psychometric evidence is unavailable and psychometric studies not possible, statistical analysis should be conducted to verify psychometric properties and at a

minimum should include reporting internal consistency reliability and structural validity (where relevant). This can help build evidence across multiple settings and populations and confirm the usefulness of ECD tools with diverse populations. 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 rigour 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 (eg, domains assessed, age range and administration time and cost among others), should be used to determine the most relevant ECD tool for the given context and use.

Several strengths of our review should be noted. We used extensive search filters, made no restrictions on domains, included both single-domain and multi-domain tools, and had a rigorously trained team with diverse backgrounds in ECD measurement in LMICs. Nevertheless, we acknowledge some limitations. We excluded studies published prior to 2007, the search start year when the inaugural *Lancet* Series on Child Development in Developing Countries was published which fundamentally changed the breadth and scale of ECD research in LMICs. Therefore, the search period captured the most important years for the evolution of psychometric research on ECD tools in LMICs. Furthermore, we did not use a validated risk of bias tool, which may limit the comparability of our findings. Finally, although we applied no language restrictions, we were limited to the languages spoken by the review team. We were unable to conduct full-text review and extraction for articles in Chinese, Farsi and Turkish. Consequently, results from China, Iran and Turkey may be underrepresented.

CONCLUSION

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 considerations 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.

Author affiliations

¹Nutrition, Diets, and Health Unit, International Food Policy Research Institute, Washington, District of Columbia, USA

²Global Academy of Agriculture and Food Systems, University of Edinburgh, Edinburgh, UK

³Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

⁴Department of Global Health, Stellenbosch University, Stellenbosch, South Africa

⁵Africa Health Research Institute (AHRI), Somkele, South Africa

⁶International Health Institute, Brown University School of Public Health, Providence, Rhode Island, USA

⁷School of Public Health, University of California Berkeley, Berkeley, California, USA

⁸Centre for Chronic Disease Control, New Delhi, India

⁹Harvard Graduate School of Education, Cambridge, Massachusetts, USA

¹⁰School of Nursing, University of Costa Rica, San Jose, San José, Costa Rica

¹¹Department of Global Health, Emory University Hubert, Atlanta, Georgia, USA

X Jonathan Seiden @jonathanseiden

Contributors LB conceptualised the review, designed the methodology, screened articles, extracted and analysed data, drafted and revised the manuscript. EH extracted data and analysed data, drafted and revised the manuscript. NBA and AR extracted data, and revised the manuscript. XH and KS-C screened articles, extracted data and revised the manuscript. SEN and JS extracted data and revised the manuscript. DO revised the manuscript. HOP screened articles and revised the manuscript. AT curated the data and revised the manuscript. JJ conceptualised the review, designed the methodology, screened articles and revised the manuscript. All authors read and approved the final manuscript. LB is the guarantor.

Funding This work was supported by the CGIAR Initiative on Resilient Cities. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund: <https://www.cgiar.org/funders>. SEN's time was supported by the National Institute of Mental Health (grant number T32 MH078788).

Map disclaimer The inclusion of any map (including the depiction of any boundaries there), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

Competing interests JS has contributed and continues to contribute to the development of the CREDI, GSED, IDELA and AIM-ECD. The remaining authors declare no conflicts of interest.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Lilia Bliznashka <http://orcid.org/0000-0003-2084-1141>

Helen O Pritchik <http://orcid.org/0000-0002-5665-0884>

Joshua Jeong <http://orcid.org/0000-0002-4130-468X>

REFERENCES

- United Nations. Transforming our world: the 2030 agenda for sustainable development. 2015. Available: https://sdgs.un.org/sites/default/files/publications/21252030_Agenda_for_Sustainable_Development_web.pdf
- Walker SP, Chang SM, Vera-Hernández M, et al. Early childhood stimulation benefits adult competence and reduces violent behavior. *Pediatrics* 2011;127:849–57.
- Gertler P, Heckman J, Pinto R, et al. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science* 2014;344:998–1001.
- Campbell F, Conti G, Heckman JJ, et al. Early childhood investments substantially boost adult health. *Science* 2014;343:1478–85.
- Fernald LCH, Prado EL, Kariger PK, et al. A toolkit for measuring early childhood development in low- and middle-income countries. World Bank; 2017. Available: <https://openknowledge.worldbank.org/entities/publication/deb106bb-7361-55c3-9c3d-edb33986a1e6>
- Pushparatnam A, Seiden JM, Luna Bazaldua DA. Publication: guiding questions for choosing the right tools to measure early childhood outcomes: why, what, who, and how. World Bank; 2022. Available: <https://openknowledge.worldbank.org/handle/10986/37030>
- American Psychological Association. APA dictionary of psychology. Available: <https://dictionary.apa.org/> [Accessed 26 2023].
- Putnick DL, Bornstein MH. Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Dev Rev* 2016;41:71–90.
- Semrud-Clikeman M, Romero RAA, Prado EL, et al. [Formula: see text]Selecting measures for the neurodevelopmental assessment of children in low- and middle-income countries. *Child Neuropsychol* 2017;23:761–802.
- Munoz-Chereau B, Ang L, Dockrell J, et al. Measuring early child development across low and middle-income countries: A systematic review. *J Early Child Res* 2021;19:443–70.
- Boggs D, Milner KM, Chandna J, et al. Rating early child development outcome measurement tools for routine health programme use. *Arch Dis Child* 2019;104:S22–33.
- Griffiths A, Toovey R, Morgan PE, et al. Psychometric properties of gross motor assessment tools for children: a systematic review. *BMJ Open* 2018;8:e021734.
- Halle TG, Darling-Churchill KE. Review of measures of social and emotional development. *J Appl Dev Psychol* 2016;45:8–18.
- Humphrey N, Kalambouka A, Wigelsworth M, et al. Measures of Social and Emotional Skills for Children and Young People. *Educ Psychol Meas* 2011;71:617–37.
- Lu C, Cuartas J, Fink G, et al. Inequalities in early childhood care and development in low/middle-income countries: 2010–2018. *BMJ Glob Health* 2020;5:e002314.
- McCoy DC, Peet ED, Ezzati M, et al. Early Childhood Developmental Status in Low- and Middle-Income Countries: National, Regional, and Global Prevalence Estimates Using Predictive Modeling. *PLoS Med* 2016;13:e1002034.
- Terwee CB, Jansma EP, Riphagen II, et al. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;18:1115–23.
- Sabanathan S, Wills B, Gladstone M. Child development assessment tools in low-income and middle-income countries: how can we use them more appropriately? *Arch Dis Child* 2015;100:482–8.
- Gridley N, Blower S, Dunn A, et al. Psychometric Properties of Child (0–5 Years) Outcome Measures as used in Randomized Controlled Trials of Parent Programs: A Systematic Review. *Clin Child Fam Psychol Rev* 2019;22:388–405.
- Crocker LM, Algina J. *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston, 1986.
- Hui CH, Triandis HC. Measurement in Cross-Cultural Psychology. *J Cross Cult Psychol* 1985;16:131–52.
- Peña ED. Lost in translation: methodological considerations in cross-cultural research. *Child Dev* 2007;78:1255–64.
- Mokkink LB, Prinsen CA, Patrick DL, et al. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs). 2018;1–78.
- ROBINS-E development group. Risk of bias in non-randomized studies - of exposure (robins-e). Launch Version; 2022. Available: <https://www.riskofbias.info/welcome/robins-e-tool>



- 25 GRADE Working Group. Handbook for grading the quality of evidence and the strength of recommendations using the grade approach. 2013. Available: <https://gdt.gradeapro.org/app/handbook/handbook.html>
- 26 Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
- 27 Pisani L, Borisova I, Dowd AJ. Developing and validating the International Development and Early Learning Assessment (IDELA). *Int J Educ Res* 2018;91:1–15.
- 28 Kersten P, Czuba K, McPherson K, *et al.* A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *Int J Behav Dev* 2016;40:64–75.
- 29 Velikonja T, Edbrooke-Childs J, Calderon A, *et al.* The psychometric properties of the Ages & Stages Questionnaires for ages 2-2.5: a systematic review. *Child Care Health Dev* 2017;43:1–17.
- 30 Halpin PF, Wolf S, Yoshikawa H, *et al.* Measuring early learning and development across cultures: Invariance of the IDELA across five countries. *Dev Psychol* 2019;55:23–37.
- 31 Wolf S, Halpin P, Yoshikawa H, *et al.* Measuring school readiness globally: assessing the construct validity and measurement invariance of the International Development and Early Learning Assessment (IDELA) in Ethiopia. *Early Child Res Q* 2017;41:21–36.
- 32 Pisani L, Seiden J, Wolf S. Longitudinal evidence on the predictive validity of the International Development and Early Learning Assessment (IDELA). *Educ Asse Eval Acc* 2022;34:173–94.