Educational Psychology Review (2024) 36:110
https://doi.org/10.1007/s10648-024-09950-6

REVIEW ARTICLE

Measuring Mathematical Skills in Early Childhood: a Systematic Review of the Psychometric Properties of Early Maths Assessments and Screeners

Laura A. Outhwaite[1] · Pirjo Aunio[2] · Jaimie Ka Yu Leung[3] · Jo Van Herwegen[1,3]

Accepted: 5 September 2024

Abstract

Successful early mathematical development is vital to children's later education, employment, and wellbeing outcomes. However, established measurement tools are infrequently used to (i) assess children's mathematical skills and (ii) identify children with or at-risk of mathematical learning difficulties. In response, this pre-registered systematic review aimed to provide an overview of measurement tools that have been evaluated for their psychometric properties for measuring the mathematical skills of children aged 0–8 years. The reliability and validity evidence reported for the identified measurement tools were then synthesised, including in relation to common acceptability thresholds. Overall, 41 mathematical assessments and 25 screeners were identified. Our study revealed five main findings. Firstly, most measurement tools were categorised as child-direct measures delivered individually with a trained assessor in a paper-based format. Secondly, the majority of the identified measurement tools have not been evaluated for aspects of reliability and validity most relevant to education measures, and only 15 measurement tools met the common acceptability thresholds for more than two areas of psychometric evidence. Thirdly, only four screeners demonstrated an acceptable ability to distinguish between typically developing children and those with or at-risk of mathematical learning difficulties. Fourthly, only one mathematical assessment and one screener met the common acceptability threshold for predictive validity. Finally, only 11 mathematical assessments and one screener were found to concurrently align with other validated measurement tools. Building on this current evidence and improving measurement quality is vital for raising methodological standards in mathematical learning and development research.

Keywords Mathematics · Early Childhood · Assessment · Screener · Measurement

Extended author information available on the last page of the article

Published online: 17 September 2024

Introduction

Successful early mathematical development is vital to children's later education, employment, and wellbeing outcomes (Bailey et al., 2020; Crawford & Cribb, 2013; Davis-Kean et al., 2022; Reyna et al., 2009). However, 55% of school-aged children worldwide do not have the level of mathematical skills needed for education and everyday life (UNESCO, 2017). Gaps between low and high attaining children also emerge early in childhood and persist throughout education (Aubrey et al., 2006). Many children also struggle to learn mathematics with estimates suggesting that between 5 and 14% of children aged 6 years and older have mathematical learning difficulties (MLD) (Morsanyi et al., 2018; Muñez et al., 2023).

To address some of these issues, research on mathematical learning and development has grown substantially in recent years. This includes knowledge advances on how typically and atypically developing children acquire mathematical skills (e.g. Gilmore, 2023; Nelson & Powell, 2018; Van Herwegen & Simms, 2020), and how cognitive development and the home and school learning environments impact these processes (e.g. Hornburg et al., 2021; Nogues & Dorneles, 2021; Turan & De Smedt, 2022), as well as how children's mathematical development can be supported through effective interventions (e.g. Ramani et al., 2012; Sella et al., 2021; Van Herwegen et al., 2018). However, recent synthesises highlight the infrequent use of established measurement tools to (i) assess children's mathematical skills (Outhwaite et al., 2022; Simms et al., 2019), and the inconsistent criteria used to (ii) identify children with or at-risk of MLD (Lewis & Fisher, 2016).

In response, the current review aimed to provide an overview of measurement tools that have been evaluated for their psychometric properties for measuring the mathematical skills of children aged 0–8 years. Specifically, the current review focused on the reliability and validity evidence most relevant to education measurements for assessing mathematical skills and identifying children with or at-risk of MLD.

Defining Mathematical Assessments and Screeners

For the purposes of the current study, measurement tools have been conceptualised as an umbrella term, which includes mathematical assessments and screeners. Mathematical assessments, in general, are designed to measure mathematical development over time and/or in response to intervention (e.g. pre- to post-test). When mathematical assessments include a standardised, norm-referenced sample, they can also be used to identify children with or at-risk of MLD based on percentile rank scores.

In contrast, screeners are measurement tools that are typically used as an efficient means to identify children with or at-risk of MLD, including those who may need additional educational support. In some cases, screeners can also be used

to monitor children's mathematical progress, particularly when they are aligned with the curriculum and are administered at more than one time point (see Nelson et al., 2023 for review on curriculum-based measures). However, as many screeners often include a small number of items and incorporate a relatively concentrated set of maths skills/concepts, caution should be taken when using screeners for this wider purpose. As such, the current review focuses on screeners for the purpose identifying children with or at-risk of MLD.

Defining Mathematical Development

It is widely acknowledged that mathematical development is a complex, multicomponent process with many skills that children need to learn from early childhood onwards (Gilmore, 2023). Early childhood is defined here as 0–8 years (UNESCO, 2023). Several models attempt to summarise the structure of early maths (Devlin et al., 2022), and thus propose the skills that should be included in mathematical assessments for this age group. For example, various models highlight the importance of number skills, such as children's knowledge of the rules and processes of numbers (e.g. the counting sequence and cardinality) and how they relate to each other (e.g. ordinality and symbolic comparison) (e.g. Aunio & Räsänen, 2016; Clements & Sarama, 2009; Purpura & Lonigan, 2015).

In addition, these models of mathematical development (Aunio & Räsänen, 2016; Clements & Sarama, 2009; Purpura & Lonigan, 2015), also include arithmetic skills, such as addition and subtraction presented in both single and multi-digit operations, as well as word problems. Alongside these number and arithmetic skills, other models of mathematical development propose a broader conceptualisation of early maths, which includes patterning (e.g. recreating repeated patterns of objects), measurement (e.g. comparing objects based on size or weight), and geometry skills (e.g. shape recognition) (e.g. Braeuning et al., 2020; Milburn et al., 2019). Some of these models (e.g. Clements & Sarama, 2009) describe a broad range of mathematical skills developing in early childhood, and others (e.g. Aunio & Räsänen, 2016) focus on mathematical skills considered essential for later mathematical development and predicting MLD.

Previous reviews have summarised some measurement tools for children's mathematical skills, but up to age 6 years with standardisations to the UK population only (Dockrell et al., 2017). Other reviews have taken a more global perspective but have focused on curriculum-based measures (Nelson et al., 2023) and teacher-implemented assessments for older children, aged 9–12 years (Hakkarainen et al., 2023). As such, it is currently unclear which mathematical assessments have been developed and validated to produce reliable indications of children's skills in early childhood.

Defining Mathematical Learning Difficulties

MLD is an umbrella term used to describe persistent problems with learning and applying mathematical facts and procedures (SASC, 2019). It includes children who fit the diagnosis for dyscalculia, mathematical disorder, or mathematical disabilities.

As definitions and diagnosis criteria differ significantly between countries and researchers (Szűcs & Goswami, 2013), the term MLD will be used in the current study to refer to children who persistently struggle with mathematics.

Children with MLD often experience persistent difficulties with reading and writing numerals, understanding how numbers relate to each other or what numbers mean, as well as remembering number facts, calculation, or mathematical reasoning (Butterworth, 2005; Vanbinst et al., 2014). Some propose that MLD is caused by a single core deficit to magnitude processing or approximate number sense (ANS) (Butterworth, 2005; Mazzocco et al., 2011), which is commonly measured using non-symbolic (i.e. dots) magnitude comparison tasks (Nosworthy et al., 2013). In contrast, others have argued that symbolic magnitude processing is a critical correlate of children's mathematical learning and that difficulties with these skills are a better predictor for MLD than other skills, such as phonological processing or working memory (De Smedt, 2022). However, it is also possible that different children with MLD struggle for different reasons and that sub-groups might be present (Bartelet et al., 2014; Costa et al., 2018).

Due to the different definitions for MLD and the varying views of its causes concerning non-symbolic and symbolic magnitude processing, measurement tools that aim to identify children with or at-risk of MLD differ widely in terms of the mathematical abilities covered. For example, while some screeners are short and only assess non-symbolic (i.e. dots) and symbolic (i.e. digits) magnitude processing (e.g. Nosworthy et al., 2013), other screeners include a wider range of mathematical abilities (e.g. Butterworth, 2003). However, it is currently unclear which measurement tools have been developed and validated to produce reliable identifications of children with or at-risk of MLD.

Indicators of Reliability and Validity for Measurement Tools

The Standards for Educational and Psychological Measurements (AERA et al., 2014) and Consensus Based Standards for the Selection of Health Status Measurement Instruments (COSMIN) guidelines (Mokkink et al., 2016; Prinsen et al., 2018) provide frameworks for appraising the psychometric properties (i.e. reliability and validity evidence) of measurement tools in education and health research. The current review focuses on the reliability and validity evidence most relevant to education measurements for assessing mathematical skills and identifying children with or at-risk of MLD. Common acceptability thresholds for these reliability and validity indicators in the context of educational research are summarised in Table 1.

Content Validity

Reporting measurement development and content validity is highly important for understanding what the measured construct is and its theoretical background, as well as what the measure is designed for, what is the target population, and the context of use. It is essential to consider if the measurement is relevant and comprehensible

Table 1 Summary of psychometric property indicators and the associated common acceptability thresholds

Psychometric Evidence	Example Analysis Methods	Common Acceptability Thresholds
Content validity	Expert panel of experts and users	Agreement across experts, with adjustments made to items when required
Structural validity	Confirmatory factor analysis (CFA)	RMSEA < .06; CFI > .95; TLI > .95 (Hu & Bentler, 1999)
	Rasch model	0.5–1.5 (Linacre, 2017)
Internal consistency	Cronbach's alpha; Kuder-Richardson (KR-20) coefficient; split-half reliability correlations	≥ .70 (Prinsen et al., 2018)
Reliability	Correlations for test–retest and inter-rater reliability	≥ . 70 (NCII, 2019)
Criterion validity	Diagnostic accuracy	Sensitivity ≥ .90; Specificity ≥ .70 (Jenkins et al., 2007; Kilgus et al., 2014)
	Concurrent, divergent, and predictive correlations between the evaluated measurement tool and 'Gold Standard' measurement tools	≥ .60 (NCII, 2019)

for users and how well it covers the phenomena assessed (i.e. comprehensiveness). In reporting articles, this evidence can be seen, for instance, in the theoretical framework explaining the theoretical background of the construct and the focus population. The evidence related to relevance, comprehensibility, and comprehensiveness are commonly gathered by using panels of experts and users, in addition to conducting pilot studies.

Structural Validity and Internal Consistency

When there is empirical data collected with the measurement tool, it is possible to report evidence of structural validity and internal consistency. Evaluations of structural validity focus on examining whether the assessment tool works as assumed, based on theory as a unidimensional or multidimensional measure. This is typically evaluated using factor analysis methods.

Evidence of internal consistency is also related to the structure of the measurement tool and refers to the degree to which included items are interrelated. It is commonly measured using Cronbach alpha for continuous data and Kuder-Richardson 20 (KR-20) coefficient for dichotomously scored data. It can also be measured using split-half reliability, which refers to the extent to which all parts of the assessment tool contribute equally to the overall measurement indicator. Ideally, internal consistencies should be reported for each of the measurement dimensions identified in the structural validity evaluation.

Reliability

The evidence of reliability includes indicators of test–retest and/or inter-rater reliability. The assumption related to test–retest reliability is that the scores of children should remain consistent across multiple measurements, often within a minimum two-week timeframe. Inter-rater reliability evidence is relevant for observational tools and refers to the consistency in scores across at least two observers.

Criterion Validity

Criterion validity produces evidence related to the relationship between the measurement tool under development and theoretically aligned measurement tools and/or external criteria. For example, when making comparisons between the measurement tool under development and other theoretically aligned measurement tools, criterion validity can be measured as concurrent (i.e. a similar measurement tool administered during the same testing period), divergent (i.e. a measurement tool measuring a different skill domain in the same testing period), and predictive validity (e.g. a similar measurement tool administered at a delayed time point). It is recommended that ‘Gold Standard’ measurement tools are used as the reference for criterion validity evaluations. This is because ‘Gold Standard’ measurement tools typically have undergone extensive development, including the establishment of various types of reliability and validity evidence, and are widely accepted as the best measurement

tools currently available. When 'Gold Standard' measurement tools are used as a reference to the criterion validity of a new measurement tool, it is expected that both tools measure the same concept(s). However, in the field of mathematical learning and development, these 'Gold Standards' are infrequently available in many countries and cultures (Hakkarainen et al., 2023).

In the case of accurately identifying children with or at-risk of MLD, evidence of criterion validity, in the form of predictive validity and/or diagnostic accuracy, is especially relevant. Predictive validity evidence of a measurement tool includes the assumption that the same children will be identified as having learning difficulties over time. To be able to produce predictive evidence, longitudinal data are needed, preferably at least six months between the measurements to give enough time for learning and development.

In terms of diagnostic accuracy, measurement tools need to be sensitive (e.g. identify true cases of children with or at-risk of MLD) and specific (e.g. identify true cases of children who do not have MLD) enough in the identification of target groups. To reduce the risk of missing children who are genuinely at risk of learning difficulties (i.e. false negatives), indicators of sensitivity are commonly prioritised, at a cost of reduced specificity in measurement tools for screening purposes (Jenkins et al., 2007; Klingbeil et al., 2019).

Cultural and Language Considerations

Overall, it is also recommended that the psychometric properties of the measurement tool are invariant across different groups of children, such as those from different countries and language groups. This ensures that children from different cultural and linguistic backgrounds are not inherently disadvantaged when using the measurement tool. It also affords the development of broader theoretical understandings of children's mathematical learning and development (Pitchford & Outhwaite, 2016), which have traditionally been focused on Western, Educated, Industrialised, Rich, and Democratic (abbreviated as WEIRD) societies (Beller & Jordan, 2018) in the Minority World (e.g. North America and Western Europe) (Draper et al., 2022).

Current Review

To support research in mathematical learning and development, this systematic review aimed to provide an overview of measurement tools that have been evaluated for their psychometric properties for measuring the mathematical skills of children aged 0–8 years. Specifically, the current review focused on the reliability and validity evidence most relevant to education measurements for assessing mathematical skills and identifying children with or at-risk of MLD. The reliability and validity evidence reported for the identified measurement tools were then synthesised, including in relation to common acceptability thresholds. Based on this evidence, measurement tools with the most promising psychometric properties were then identified. Such synthesises are important for supporting researchers, educators, and

other stakeholders to select measurement tools that are most suitable for assessing children's mathematical skills over time, including in response to interventions, and for identifying children with or at-risk of MLD (Hakkarainen et al., 2023).

Methods

The protocol for this systematic review was pre-registered on the Open Science Framework (blinded for review) with ethical approval granted by (blinded for review). The PRISMA protocol was used to secure the quality of reporting in the current review (Page et al., 2021).

Search Strategy

The systematic literature search was conducted across seven scholarly databases and two grey literature sources (see Figs. 1 and 2) with the following search string: "Primary school" OR "elementary school" OR kindergart* OR preschool* OR "early years" OR child* OR toddler OR "child development" AND "assessment measure" OR screen* OR "parent report" OR "teacher report" OR "caregiver report" OR observation OR test* OR checklist AND math* OR "number sense" OR numeracy OR symbolic OR "non symbolic" OR counting OR arithmetic* OR geomet* OR shape AND Psychometric* OR "Psychometric Properties" OR reliability OR validity OR sensitivity OR "internal consistency". A backward citation of included studies ($n=57$) was also conducted, including the test manuals of the measurement tools most frequently used when establishing criterion validity. This search strategy was completed in March 2021 (from January 1990–present) and was updated in June 2023 (from January 2021–present). An additional forward citation search of included studies ($n=71$) was conducted in May 2024 to ensure the latest and most comprehensive data were used in the current review.

Inclusion and Exclusion Criteria

To be included in the current review, studies needed to meet the following pre-registered inclusion and exclusion criteria.

Population

Studies needed to focus on mathematical measurement tools for children aged 0–8 years. If studies reported a measurement tool that was suitable for children extending beyond the specified age range (e.g. 5–11 years), this tool was eligible for inclusion. No restriction was placed on whether the measurement tool was designed for typically developing children or for identifying those with or at-risk of MLD. The first author categorised the purpose of each measurement tool

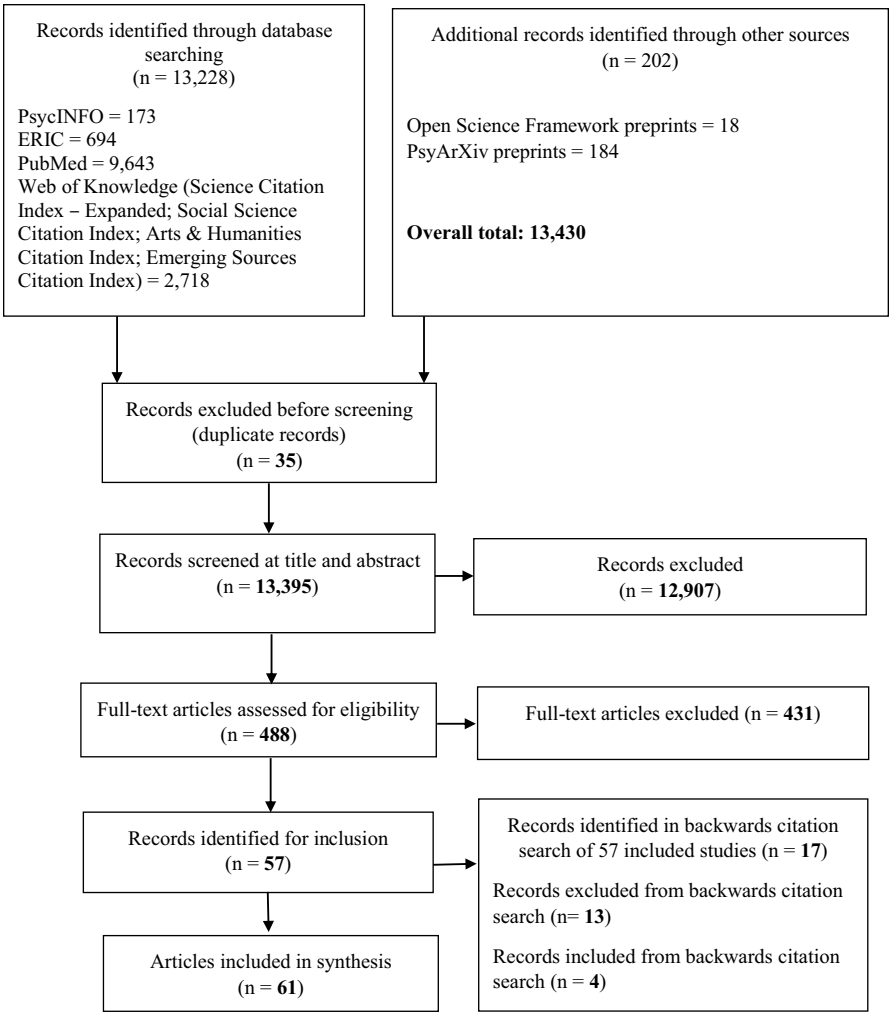

Fig. 1 PRISMA flow diagram of studies through the systematic review (original search, March 2021)

(i.e. assessment or screener) based on how it was presented in the included psychometric studies. Twenty percent of measurement tools were also second-coded by the last author with 100% agreement.

Measurement Tool

Included studies needed to report the psychometric properties of a named measurement tool, which measures any area of mathematics, including number, arithmetic, and shape, space, and measure. Measurement tools that assessed children's mathematics anxiety, language, or vocabulary, as well as teachers/caregivers' perceptions

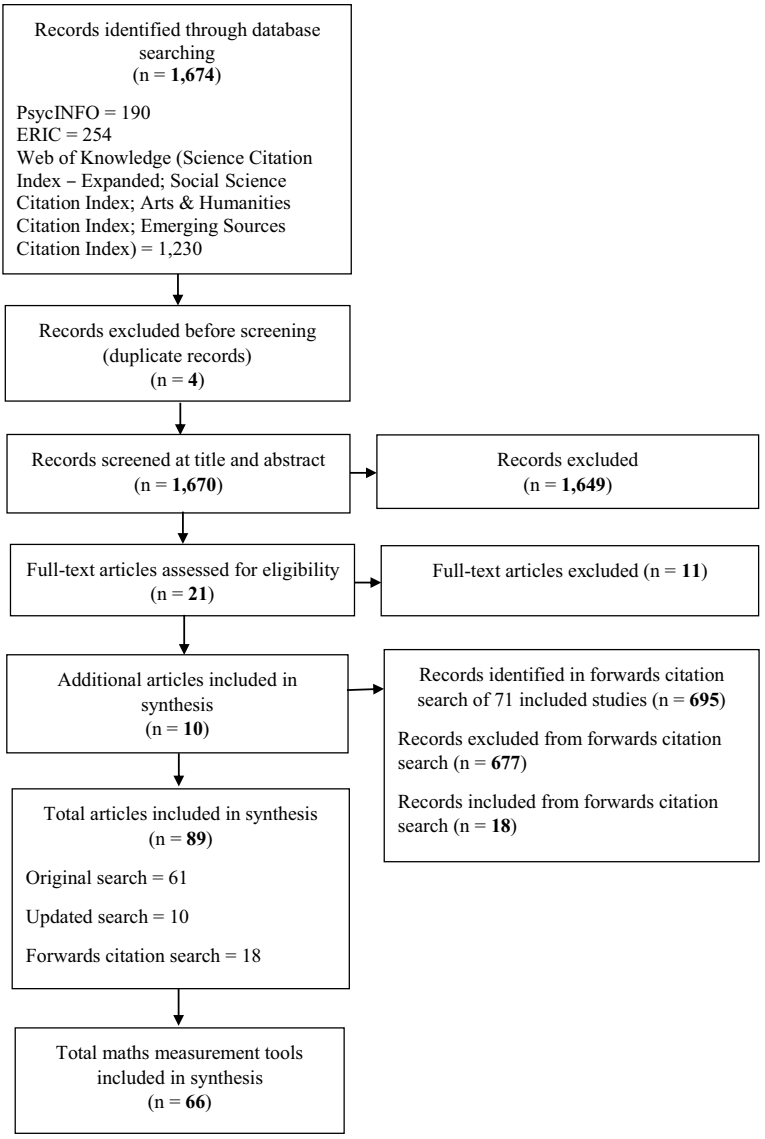

Fig. 2 PRISMA flow diagram of studies through the systematic review (updated search, June 2023 and May 2024)

of the importance of mathematics, were not eligible for inclusion. International large-scale tests (e.g. PISA) or national government statutory assessments were also beyond the scope of the current review and were not eligible for inclusion. No restriction was placed on whether the measurement tool was a direct measure of a child's mathematical skills or teacher/caregiver report of children's maths skills.

Psychometric Properties

Studies also needed to describe the psychometric properties (e.g. reliability and validity evidence), of the named measurement tool (see Table 1). If some details were missing, these were labelled as 'not reported' in the study synthesis.

Other Criteria

No restriction was placed on the geographical location or the language of the measurement tool. However, the full-text records needed to be accessible to download and available in English. Studies also needed to be published since January 1990 and report original data; commentary or position papers were not eligible for inclusion.

Record Screening

As outlined in the PRISMA Flow Diagram (Page et al., 2021; see Fig. 1), the initial searches in March 2021 identified 61 eligible studies. One reviewer (first author) was responsible for screening all records at both levels. A random 20% sample of records was screened by an additional reviewer (see acknowledgements) to ensure high levels of agreement ($\kappa = 0.84$). An updated search strategy was completed in June 2023 (see Fig. 2) and identified an additional 10 eligible studies ($n = 71$). Consistent with the initial search, one reviewer (third author) was responsible for screening all records at both levels. A random 20% sample of records was also screened by an additional reviewer (first author) to ensure high levels of agreement ($\kappa = 0.93$). The forward citation search completed in May 2024 identified a further 18 studies. In total, 89 studies were included in the current review.

Coding Framework

To establish an overview of each of the measurement tools identified in the 89 eligible studies, information was extracted based on the age range covered, country(s) and language(s) in which the tools were developed, and the measurement type (e.g. child-direct) and format (e.g. paper-based), as well as the measurement mode (e.g. individual) and administrator (e.g. researcher/ training assessor). Information relating to the number of items and the mathematical concepts assessed were also extracted, based directly on the terminology used in the eligible studies. Although there were inconsistencies in the terminologies used for different mathematical concepts (e.g. ANS, non-symbolic magnitude, dot comparison), the assessment tasks were broadly categorised as number (N), arithmetic (A), and shape, space, and measure (SSM). These 'areas of maths' categories were based on widely recognised models of mathematical development (e.g. Aunio & Räsänen, 2016; Clements & Sarama, 2009; Milburn et al., 2019; Purpura & Lonigan, 2015).

Data related to the psychometric properties (i.e. reliability and validity evidence) were also extracted for each of the measurement tools in the study synthesis. These

data were then rated based on the common acceptability thresholds in education research (see Table 1). If the relevant psychometric property evidence fully met the outlined thresholds, the measurement tool was rated as 'Acceptable'. If a range of results were reported, which were both above and below the thresholds, it was rated as 'Mixed'. If the evidence did not meet these thresholds, it was rated as 'Low'. In cases where acceptability thresholds were not widely available within the literature, conventional thresholds for Pearson's correlations were used (<0.30 = low; 0.3–0.5 = medium; >0.5 = high/acceptable) or were rated as 'not applicable' (NA), if other forms of analysis were used.

Results

Overview of Measurement Tools

In total, 66 measurement tools were identified across 89 included studies. This included 41 mathematical assessments designed for children aged 1–14 years and 25 screeners suitable for children aged 3–14 years. As summarised in Table 2, most measurement tools were child-direct measures ($n=57$) administered individually ($n=58$) with a trained assessor ($n=54$) in a paper-based format ($n=47$). Most measurement tools targeted number ($n=60$) and/or arithmetic skills ($n=51$), with less than half of the identified assessments and screeners measuring shape, space, and measure skills ($n=26$).

Although the identified measurement tools were evaluated in over 55 countries and 31 languages, over half of the assessments and screeners were developed in WEIRD societies in minority countries and/or in English ($n=36$). Only ten assessments and three screeners were evaluated in different countries, cultures, and/or language groups (see Table 2). For most of these measurement tools, the different language groups were considered within the same study. However, as the evaluations of the English and Spanish versions of the Birthday Party assessment (Lee, 2016), the English and Turkish Versions of the NSS (Jordan et al., 2010, 2012), and the English and Greek versions of the PENS-B screener (Purpura et al., 2015) were conducted separately, the synthesis of psychometric properties henceforth refers to 42 assessments and 27 screeners.

Content Validity

Content validity in the form of expert opinion on the suitability and adaptation of test items was only reported for eight mathematical assessments (AAT, Ralston et al., 2018; EMAT, Ceylan & Aslan, 2023; IDELA, Save the Children, 2019; Numeracy-Caregiver report questionnaire, Pushparatnam et al., 2021; Numeracy-Child direct assessment, Pushparatnam et al., 2021; REMA, Clements et al., 2008; Dong et al., 2023; TENA, Bojorque et al., 2015; ENT Test, Aunio et al., 2006) and four screeners (BNPT, Olkun et al., 2016; Dyscalculia Test, Eteng-Uket, 2023; EM-CBM, Clarke et al., 2023; NSS, Jordan et al., 2012). All were rated as acceptable.

Table 2 Overview of the measurement tools identified through the systematic review (*N*, number; *A*, arithmetic; *SSM*, shape, space, and measure)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Mathematical assessments								
Academic Rating Scale (ARS)-adapted (Kilday et al., 2012)	3–5	USA (English)	Observation; paper-based	Individual or group; teacher	**12 items**: number sense; numerical operations; geometry; measurement	Y	Y	Y
Ani Banani Test (ten Braak & Størksen, 2021)	4–7	Norway (Norwegian)	Child-direct; tablet-based (child uses)	Individual; researcher/trained assessor	**18 items**: numeracy; geometry; problem solving	Y	Y	Y
Arithmetic Calculation Efficiency Test (TECA) (Singer & Cuadro, 2014)	6–11	Uruguay (Spanish)	Child-direct; paper-based	Group; researcher/trained assessor	**144 items**: addition; subtraction; multiplication; division	-	Y	-
Assessment of Algebraic Thinking (AAT) (Ralston et al., 2018)	6–11	USA (English)	Child-direct; paper-based	Group; teacher	**25 items**: open number sentences; equivalence; work with variables; efficient numerical calculation; generalisation; numerical patterns; figural patterns; generalising figural patterns	Y	Y	Y
Birthday Party- Long Version (Ginsburg & Pappas, 2016)	3–5	USA (English; Spanish)	Child-direct; computer-based (child uses)	Individual; teacher	**30–36 items**: number and operations; shape; space; pattern	Y	Y	Y
CIRCLE Progress Monitoring (CPM) Math Subtest (Assel et al., 2020)	4–5	USA (English)	Child-direct; paper-based	Individual; researcher/trained assessor	**27 items**: rote counting; Shape naming; number naming; shape discrimination; counting; simple addition and subtraction word problems	Y	Y	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Cognitive Diagnostic Test (Li et al., 2020)	3–6	China (English)	Interview; paper-based	Individual; researcher/trained assessor	**38 items**: cardinality concept; set comparison; addition and subtraction within 10; combine; result-unknown change; change-unknown change; consistent language comparison; inconsistent language comparison; addition and subtraction inverse reasoning; additive composition reasoning; one-to-many correspondence reasoning	Y	Y	-
Comprehensive Learning Test-Mathematics (CLT-M) (Lee et al., 2017)	5–14	South Korea (Korean)	Child-direct; computer-based (child uses)	Individual; researcher/trained assessor	**5 subtests** (number of items NR): whole number computation; numeral comparing/magnitude; numeral comparing/distance; enumeration of dot group; number line estimation	Y	-	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Comprehensive Research-Based Early Math Ability Test (CREMAT) (Clements et al., 2022)	6–8	USA (English)	Child-direct; computer-based (child uses)	Individual; teacher	**42 items**: measurement; length; area	-	-	Y
Curriculum Based Measures for Kindergarten- Grade 3 (Lee & Lembke, 2016)	5–9	USA (English)	Child-direct; tablet-based (child uses)	Individual; researcher/ trained assessor	**8 tasks** (number of items NR): counting; missing number; quantity discrimination; next number; number identification; computation; concepts; number facts	Y	Y	-
DIFER School Readiness Test Battery Counting and Basic Numeracy (Csapó et al., 2014)	6–7	Hungary (Hungarian)	Child-direct; computer-based (child uses)	Individual; researcher/ trained assessor	**13 items**: number; number relations; basic mathematical thinking	Y	Y	-
Early Arithmetic, Reading and Learning Indicators (EARLI) – Numeracy measures (DiPerna et al., 2007)	3–4	USA (English)	Child-direct; paper-based	Individual; researcher/ trained assessor	**58 items**: number recognition; shape recognition; measurement concepts	Y	-	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Early Grade Mathematics Assessment (EGMA) (RTI International, 2014)	6–10	14 LMICs (Various)[a]	Child-direct; paper-based	Individual; researcher/trained assessor	**77 items**: number identification; quantity discrimination; missing number; addition; subtraction; multiplication; division; shape recognition	Y	Y	Y
Early Learning Outcomes Measure (ELOM) (Snelling et al., 2019)	4–5	South Africa (Afrikaans, English, Setswana, isiZulu; isiXhosa)	Child-direct; paper-based	Individual; researcher/trained assessor	**8 items**: counting; addition and subtraction; sorting and classification; spatial vocabulary; measurement vocabulary	Y	Y	Y
Early Measurement Assessment Tool (EMAT) (Ceylan & Aslan, 2023)	4–8	Türkiye (Turkish)	Child-direct; paper-based	Individual; researcher/trained assessor	**38 items**: length; area; volume; angle	-	-	Y
Early Years Toolbox-Early Numeracy (Howard et al., 2022)	3–5	Australia (English)	Child-direct; tablet-based (child uses);	Individual; teacher	**79 items**: number sense; cardinality and counting; numerical operations; spatial and measurement constructs; patterning	Y	Y	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Evaluación Neuropsicológica Infantil-Preescolar (ENI-P)-Numerical Abilities Test (Beltrán-Navarro et al., 2018)	2–4	Mexico (Spanish)	Child-direct; paper-based;	Individual; researcher/trained assessor	**26 items**: magnitude comparison; counting; subitising; basic calculation	Y	Y	-
Galileo Early Maths-Revised (Kowalski et al., 2018)	3–5	USA (English)	Observation; computer-based	Individual or group; teacher	**22 items**: number writing; shape recognition; counting; sorting; measurement comparison; ordinality; more and less; addition; sharing; division; pattern recognition	Y	Y	Y
Heidelberger Rechen Test 1–4 (Hassler Hallstedt & Ghaderi, 2018)	6–10	Sweden (Swedish)	Child-direct; tablet-based (child uses)	Individual; researcher/trained assessor	**121 items**: addition; subtraction; missing term; count amount; tap rate	Y	Y	-
KeyMath-Revised (Connolly, 1988)	7–10	USA (English)	Child-direct; paper-based	Individual; researcher/trained assessor	**258 items**: numeration; geometry; addition; subtraction; measurement; time and money; rational numbers; multiplication; division; mental computation; estimation; interpreting data; problem solving	Y	Y	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Kieler Kindergartentest Mathematik (KiKi) (Van Hoogmoed et al., 2022)	4–6	Germany; Netherlands (German; Dutch)	Child-direct; paper-based	Individual; researcher/trained assessor	**31–32 items:** sets, numbers, and operations; measurement; space and shape; change and relationships; data and chance	Y	Y	Y
Mathematical and Arithmetic Competence Diagnostic (MARKO-D) (Ricken et al., 2013)	6–7	Germany; South Africa (German; English; Afrikaans; isiZulu; Sesotho)	Child-direct; paper-based	Individual; researcher/trained assessor	**55 items**: counting; ordinality; cardinality; part-part-whole; relationality	Y	-	-
Mathematical Profile (MathPro) Test (Karagiannakis & Noël, 2020)	6–12	Belgium (Dutch)	Child-direct; computer-based (child uses)	Individual; researcher/trained assessor	**212–339 items**: dot magnitude comparison; single and multidigit number magnitude comparison; number dictation; next number; previous number; subitising; enumeration; addition facts retrieval; multiplication facts retrieval; mental calculations; number lines 0–100; number lines 0–1000; squares; building blocks; word problems; calculation principles; numerical patterns	Y	Y	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Mathematical Reasoning Test (Nunes et al., 2015)	7–9	UK (English)	Child-direct; paper-based	Group; teacher	**17 items**: additive composition; additive reasoning; multiplicative reasoning	-	Y	-
mCLASS: Math (Lee et al., 2010)	5–9	USA (English)	Interview; computer-based (child uses)	Individual; researcher/trained assessor	**5 domains** (number of items NR): counting; addition; subtraction; multiplication; written numbers	Y	Y	-
Neuropsychological Test Battery for Number Processing and Calculation in Children – Revised (NUCALC- R)[b] (von Aster et al., 2006)	7–12	Brazil; Germany; Greece (Brazilian Portuguese; German; Greek)	Child-direct; paper-based	Individual; researcher/trained assessor	**100 items**: counting dots; counting backwards; dictation of numbers; positioning numbers; oral comparison; perceptive estimation; contextual estimation; written comparison; mental calculation; problem solving	Y	Y	-
Number Sense Test (Malofeeva et al., 2004)	3–5	USA (English)	Child-direct; paper-based	Individual; researcher/trained assessor	**165 items**: counting; number identification; number-object correspondence; ordinality; comparison; addition; subtraction	Y	Y	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Numeracy-Caregiver report questionnaire (Pushparatnam et al., 2021)	4–6	8 LMICs (Various)[c]	Interview; paper-based	Individual; researcher/trained assessor with parent	**24 items**: verbal counting; set production' mental addition; numeral identification; spatial sense; measurement vocabulary	Y	Y	Y
Numeracy-Child direct assessment (Pushparatnam et al., 2021)	4–6	10 LMICs (Various)[d]	Child-direct; paper-based	Individual; researcher/trained assessor	**42 items**: verbal counting; set production; mental addition; numeral identification; spatial sense; measurement vocabulary	Y	Y	Y
Observing and Analysing Children's Mathematical Development (OAMD) (Bunck et al., 2017)	5–9	Netherlands (Dutch)	Observation & interview; paper-based	Individual; researcher/trained assessor	**18 items**: counting; numbers; addition/subtraction; multiplication/division	Y	Y	-
Parent ratings of numeracy skills (Lin et al., 2021)	3–5	USA (English)	Observation; paper-based;	Individual or group; Parent	**11 items:** Verbal counting; Simple arithmetic; Numeral identification	Y	Y	-
Quantitative Reasoning Test (Nunes et al., 2015)	5–6	UK (English)	Child-direct; paper-based	Individual; researcher/trained assessor	**32 items:** Additive composition; Inverse relations; Additive reasoning; Multiplicative reasoning	-	Y	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Research-Based Early Maths Assessment (REMA) (Clements et al., 2008)	4–5	USA (English)	Child-direct; paper-based	Individual; researcher/ trained assessor	**199 items**: comparing and ordering; counting; arithmetic; recognition of number and subitising; composing number; geometry; comparing shape; identifying shape; turns; representing shape; composing shape; measuring; patterning	Y	Y	Y
Research-Based Early Maths Assessment-Short Form (REMA-SF)[e] (Weiland et al., 2012)	4–5	USA (English)	Child-direct; paper-based	Individual; researcher/ trained assessor	**19 items**: counting; comparing number and sequencing; recognition of number and subitising; numerals; composition of number; arithmetic; shape; patterning shape; compose shape	Y	Y	Y
School Achievement Test- 2nd Edition-Arithmetic Subtest (Viapiana et al., 2016)	6–14	Brazil (Brazilian Portuguese)	Child-direct; paper-based	Individual; researcher/ trained assessor	**202 items**: number recognition, composition, and writing; counting; sequencing; arithmetic; decimals; fractions	Y	-	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Teacher Rating Scale – Early Numeracy (TRS-EN) (Vessonen et al., 2023)	3–4	Finland (Finnish)	Observation; paper-based	Individual or group; teacher	**22 items**: counting; numerical relations; basic arithmetic skills	Y	Y	-
Teaching Strategies GOLD (Heroman et al., 2010)	1–4	USA (English)	Observation; paper-based	Individual; teacher	**7 items**: number concepts and operations; spatial relationships and shapes; measurement and comparison; pattern knowledge	Y	Y	Y
Test of Early Number and Arithmetic (TENA) (Bojorque et al., 2015)	4–5	Ecuador (Spanish)	Child-direct; paper-based	Individual; researcher/ trained assessor	**54 items**: quantifiers; one-to-one correspondence; order relations more than/less than; counting; quantity identification and association with numerals; ordering; reading and writing numerals; addition; subtraction	Y	Y	-
The International Development and Early Learning Assessment (IDELA)-Emergent Numeracy (Save the Children, 2019)	3–6	78 countries[f]	Child direct; paper-based	Individual; researcher/ trained assessor	**7 domains** (number of items NR): size and length comparison; shape identification; number identification; numerical knowledge; addition and subtraction; puzzle	Y	Y	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
The Utrecht Test of Early Numeracy (ENT) Test (Van Luit et al., 1994; Van de Rijt et al., 2003)	4–8	8 European countries (Various)[g]	Child-direct; paper-based	Individual; researcher/ trained assessor	**40 items**: comparison; classification; making correspondence; seriation; using number words; synchronous and shortened counting; resultative counting; general knowledge of numbers	Y	Y	-
Tools for Early Assessment in Math Danish Version (DK-TEAM) (Sjoe et al., 2019)	3–6	Denmark (Danish)	Child-direct; tablet-based (assessor uses)	Individual; researcher/ trained assessor	**19 items**: patterns and pre-algebraic thinking; recognising shapes; comparing shapes; counting; comparing and ordering numbers; numerals; composing numbers	Y	Y	Y
Screeners								
Arabic number-writing task (Moura et al., 2015)	6–10	Brazil (Brazilian Portuguese)	Child-direct; paper-based	Individual; researcher/ trained assessor	**28 items**: Write dictated 1–4-digit numbers	Y	-	-
Assessing Student Proficiency of Early Number Sense (ASPENS) (Clarke et al., 2011)	5–7	USA (English)	Child-direct; paper-based	Individual; researcher/ trained assessor	**4 tasks** (number of items NR): numeral identification; Magnitude comparison; Missing numbers; Basic arithmetic facts and base 10	Y	Y	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Basic Number Processing Test (BNPT) (Olkun et al., 2016)	6–9	Türkiye (Turkish)	Child-direct; tablet-based (child uses)	Individual; researcher/trained assessor	**71 items**: canonic dot counting; symbolic number comparison; mental number line	Y	-	-
Birthday Party- Short Version (Ginsburg & Pappas, 2016)	3–5	USA (English)	Child-direct; computer-based (child uses)	Individual; teacher	**13–21 items**: number and operations; shape; space; pattern	Y	Y	Y
Dyscalculia screener (Butterworth, 2003)	6–14	UK (English)	Child-direct; computer-based (child uses)	Individual; researcher/trained assessor	**4 domains** (number of items NR): dot enumeration; number comparison; addition; multiplication	Y	Y	-
Dyscalculia Test (Eteng-Uket, 2023)	7–13	Nigeria (Nigerian)	Child-direct; paper-based	Individual; researcher/trained assessor	**85 items**: number sense; arithmetic operations; computation; working memory	Y	Y	-
Early Measurement Curriculum-Based Measures (EM-CBM) (Clarke et al., 2023)	6–7	USA (English)	Child-direct; paper and tablet-based (assessor uses)	Individual; researcher/trained assessor	**120 items**: length-comparison; length-measurement; iteration-application; iteration-conceptual	-	-	Y
Early Numeracy (EN)-Test (Koponen et al., 2011)	5–8	Finland (Finnish, Swedish)	Child-direct; paper-based	Group; researcher/trained assessor	**48- 64 items**: symbolic and non-symbolic number knowledge; understanding mathematical relations; counting; basic arithmetic	Y	Y	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Early Numeracy Screener (Lopez-Pedersen et al., 2021)	6–7	Norway (Norwegian)	Child-direct; paper-based	Group; teacher	**52 items**: numerical relational skills; counting skills; arithmetic skills	Y	Y	-
Early Numeracy Skill Indicators (Methe et al., 2008)	5–6	USA (English)	Child-direct; paper-based	Individual; researcher/trained assessor	**4 tasks** (number of items NR): counting on fluency; match quantity fluency; number recognition fluency; ordinal position fluency	Y	-	-
House of Numbers (HoN) (Chatzaki et al., 2024)	5–6	Germany (German)	Child-direct; paper-based	Individual; researcher/trained assessor	**24 items**: counting; mental number line; cardinality and decomposition; class inclusion and embeddedness; relationality	Y	-	-
Indicators of Basic Early Math Skills (IPAM) (Jiménez & de León, 2019)	6–7	Spain (Spanish)	Child-direct; Paper-based	Group; researcher/trained assessor	**5 tasks** (number of items NR): quantity discrimination; missing number; single-digit computation; multi-digit computation; place value	Y	Y	-
Math Essential Skill Screener- Elementary Version (MESS-E) (Erford et al., 1998)	6–8	USA (English)	Child-direct; paper-based	Individual or group; researcher/trained assessor	**27 items**: writing numerals; addition; subtraction; time; money; fractions; word problems (addition and subtraction)	Y	Y	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Mathematical School Readiness (MSR) (Mejias et al., 2019)	6–7	Belgium (French)	Child-direct; paper-based	Individual; researcher/trained assessor	**3 items**: number writing; Number comparison; Arithmetic problem solving	Y	Y	-
Number line assessment (Clarke et al., 2020)	5–7	USA (English)	Child-direct; tablet-based (child uses)	Individual; researcher/trained assessor	**26 items**: number line estimation 0–20 and 0–100	Y	-	-
Number Line Assessment 0–100 (Sutherland et al., 2021)	5–6	USA (English)	Child-direct; tablet-based (child uses)	Individual; researcher/trained assessor	**26 items**: number line estimation 0–100	Y	-	-
Number Sense Screener (NSS) (Jordan et al., 2010, 2012; Kiziltepe, 2019)	5–6	USA; Türkiye (English; Turkish)	Child-direct; paper-based	Individual; researcher/trained assessor	**26–33 items**: counting knowledge and principles; number recognition; number knowledge; non-verbal addition and subtraction; addition and subtraction story problems; addition and subtraction number combinations	Y	Y	-
Number Sets Test (Geary et al., 2009)	5–6	USA (English)	Child-direct; paper-based	Individual; Researcher/trained assessor	**84 items**: comparing set sizes to 5 and 9	Y	-	-
Numeracy Screener (Nosworthy et al., 2013)	5–9	Canada (English)	Child-direct; paper-based	Individual; Researcher/trained assessor	**112 items**: symbolic magnitude comparison; non-symbolic magnitude comparison	Y	-	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Preschool Early Numeracy Skills Screener- Brief Version (PENS-B) (Purpura et al., 2015)	3–5	USA; Greece (English; Greek)	Child-direct; paper-based	Individual; researcher/ trained assessor	**24 items**: counting; set comparison; numeral identification; set to numerals; number order; relative size; story problems; number comparison; number combinations; ordinality	Y	Y	-
Preschool Numeracy Indicators (Floyd et al., 2006)	3–6	USA (English)	Child-direct; paper-based	Individual; researcher/ trained assessor	**5 tasks** (number of items NR): one-to-one correspondence; counting fluency; oral counting fluency; number naming fluency; quantity comparison fluency	Y	-	-
Primary Math Assessment Diagnostic (PMA-D) (Brendefur et al., 2018)	5–8	USA (English)	Child-direct; computer-based (child uses)	Individual; researcher/ trained assessor	**64 items**: number identification; number recognition; number sequence; quantity discrimination; fact fluency addition; fact fluency subtraction; number sentences; bar model; join; separate; part-whole; transitivity; composing; decomposing; rotation	Y	Y	Y

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Primary Math Assessment Screener (PMA-S) (Brendefur et al., 2018)	5–8	USA (English)	Child-direct; computer-based (child uses)	Individual; researcher/trained assessor	**6 items**: number sequencing; number facts; relational thinking; context; measurement; spatial reasoning	Y	Y	Y
Symbolic Magnitude Processing (SYMP) Test (Brankaer et al., 2017)	6–11	Belgium (Dutch)	Child-direct; paper-based	Group; researcher/trained assessor	**2 tasks**: one-digit symbolic comparison (digits between 1 and 9); two-digit symbolic comparison (digits between 11 and 99)	Y	-	-
Universal Screening of Math Skills/ Despiste Universal de Competências Matemáticas (DUCMa) (Cruz et al., 2024)	4–6	Portugal (Portuguese)	Child-direct; paper-based	Group or individual; researcher/trained assessor	**8 tasks**: cardinality; number recognition; number writing; counting; subitising; quantity discrimination; addition; subtraction	Y	Y	-
Measurement tools widely used to assess criterion validity								
Test of Early Mathematics Abilities-3rd Version (TEMA-3; Ginsburg & Baroody, 2003)	3–8	USA; Singapore; China; Spain; Türkiye (English; Mandarin; Spanish; Turkish)	Child-direct; paper-based	Individual; researcher/trained assessor	**72 items**: informal mathematics-numbering; number comparisons; calculations; concepts. formal mathematics-numeral literacy; number facts; calculation; concepts	Y	Y	-

Table 2 (continued)

Name of Measurement Tool (Original Authors)	Age Range (Years)	Country (Languages)	Measure Type; Format	Measure Mode; Administrator	Number of Items: Measurement Tasks	Areas of Maths Development		
						N	A	SSM
Woodcock-Johnson III Achievement Tests Maths (WJ-III-ACH Maths) (Schrank et al., 2001; Woodcock et al., 2001)	5–19	USA; Spain (English; Spanish)	Child-direct; paper-based	Individual; researcher/trained assessor	**4 sub-tests**: calculation (C); maths fluency (MF); applied problems (AP); quantitative concepts (QC) Four sub-tests grouped into three clusters: Broad Math (C; MF; AP); Math Calculation Skills (C; MF); Math Reasoning (AP; QC)	Y	Y	Y

[a]14 low- and middle-income countries (LMICs): Democratic Republic of Congo, Dominican Republic, Ghana, Iraq, Jordan, Kenya, Liberia, Malawi, Mali, Morocco, Nicaragua, Nigeria, Rwanda, and Zambia (available in English and various local languages)

[b]Also known as Neuropsychologische Testbatterie fur Zahlenarbeitung und Retchen bei Kindern (ZAKERI) in German

[c]8 LMICs: Ethiopia, Laos, Lesotho, Madagascar, Nigeria, Pakistan, and two anonymous Central and South American countries (available in various local languages)

[d]10 LMICs: Ethiopia, Kenya, Laos, Lesotho, Nigeria, Pakistan, Sudan, Tanzania, and two anonymous Central and South American countries (available in various local languages)

[e]Also known as US- Tools for Early Assessment (TEAM)- Short (see Sjoe et al., 2019; Weiland et al., 2012)

[f]Evaluations conducted in Afghanistan, Bangladesh, Bhutan, Brazil, Bolivia, Egypt, Ethiopia, Ghana, Indonesia, Malawi, Mali, Mozambique, Pakistan, Rwanda, Uganda, Vietnam, Zambia (Halpin et al., 2019; Pisani et al., 2018; Pisani et al., 2022; Shavitt et al., 2022; Wolf et al., 2017). For full list, see https://resourcecentre.savethechildren.net/document/idela-the-international-development-and-early-learning-assessment/

[g]8 European countries: Belgium, Finland, Germany, Greece, Romania; Netherlands, UK, Slovenia (Finnish; German; Greek; Romanian; Dutch; English; Slovenian)

Concurrent validity was also evaluated with 11 screeners, with comparisons commonly made with Woodcock-Johnson Math subtests ($n=3$) and TEMA-3 ($n=3$). However, only one screener met the acceptability thresholds (see Table 4). Divergent validity with standardised reading measurement tools was considered in three screeners, but only one was rated as acceptable. Predictive validity was considered in nine screeners, over periods ranging from 10 weeks to 3 years. However, only one screener had acceptable predictive validity (NSS; Jordan et al., 2012). All other screeners were rated as either mixed ($n=1$) or low ($n=7$) on the acceptability thresholds (see Table 4). Diagnostic accuracy was also considered in ten screeners. However, there were large variations in the reported sensitivity and specificity, with only four screeners meeting the acceptability thresholds (see Table 4).

Overall, the Woodcock-Johnson III Math subtests and TEMA-3 were the measurement tools most widely used to assess criterion validity. As such, an overview of these measures is reported in Table 2 with the psychometric properties included in Tables 3 and 4.

Measurement Tools with Promising Evidence

Table 5 summarises the nine mathematical assessments and six screeners with the most promising psychometric evidence identified within the current review.

Discussion

This study reports the first pre-registered systematic review of the psychometric properties of mathematical assessments and screeners in early childhood. This review aimed to provide an overview of measurement tools that have been evaluated for their psychometric properties for measuring mathematical skills in children aged 0–8 years. Specifically, the current review focused on the psychometric (i.e. reliability and validity) evidence most relevant to education measurements for assessing mathematical skills and identifying children with or at-risk of MLD (AERA et al., 2014; Mokkink et al., 2016; Prinsen et al., 2018). Eighty-nine individual studies relating to 66 measurement tools were identified, of which 41 were mathematical assessments and 25 were screeners. The psychometric properties of these measurement tools were then synthesised and appraised in line with common acceptability thresholds for the five indicators of reliability and validity (content validity, structural validity, internal consistency, reliability, and criterion validity).

The current review revealed five main findings. Firstly, most measurement tools were categorised as child-direct measures delivered individually with a trained assessor in a paper-based format. Secondly, the majority of the identified measurement tools have not been evaluated for aspects of reliability and validity most

Table 3 Structural validity and external validity for identified mathematical assessments and screeners

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Mathematical assessments						
Academic Rating Scale (ARS)-adapted (Kilday et al., 2012)	NR	NR	NA	NR	NR	NA
Ani Banani Test (ten Braak & Størksen, 2021)	CFA	1 factor model RMSEA ≤ .03, CFI ≥ .96, TLI ≥ .95	Acceptable	NR	NR	NA
Arithmetic Calculation Efficiency Test (TECA) (Singer & Cuadro, 2014)	CFA	1 factor model Tanaka index > .98	Acceptable	NR	NR	NA
Assessment of Algebraic Thinking (AAT) (Ralston et al., 2018)	IRT	Most item total correlations > .25	Low	Cronbach alpha	α = > .70	Acceptable
Birthday Party-Long Version (English) (Lee, 2016)	CFA	4 factor model RMSEA ≥ .05, CFI ≥ .96, TLI ≥ .91	Mixed	Cronbach alpha	α = .76–.94	Acceptable
Birthday Party-Long Version (Spanish) (Lee, 2016)	NR	NR	NA	Cronbach alpha	α = .34–.86	Mixed
CIRCLE Progress Monitoring (CPM) Math Subtest (Assel et al., 2020)	CFA	5 factor model RMSEA = .05, CFI = .98	Acceptable	Cronbach alpha	α = .94	Acceptable
Cognitive Diagnostic Test (Li et al., 2020)	NR	NR	NA	NR	NR	NA

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Comprehensive Learning Test-Mathematics (CLT-M) (Lee et al., 2017)	PCA	4 factor model that explained 66.4% of the cumulative variance	NA	NR	NR	NA
Comprehensive Research-Based Early Math Ability Test (CREMAT) (Clements et al., 2022)	NR	NR	NA	NR	NR	NA
Curriculum Based Measures for Kindergarten- Grade 3 (Lee & Lembke, 2016)	CFA	8 factor model RMSEA = .00–.05, SRMR = .003–023, CFI = .99–1.00	Acceptable	Cronbach alpha	α = .69–.97	Mixed
DIFER School Readiness Test Battery Counting and Basic Numeracy (Csapó et al., 2014; Józsa et al., 2023)	Rasch Model	Infit MNSQ = .97–.98 Outfit MNSQ = 1.20–1.99	Acceptable Mixed	Cronbach alpha	α = .74–.94	Acceptable
Early Arithmetic, Reading and Learning Indicators (EARLI) – Numeracy measures (Cheng et al., 2017; Lei et al., 2009)	Unidimensionality	Evidence of unidimensionality for 5 groups of items	NA	Cronbach alpha	α = .82–.98	Acceptable

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Early Grade Mathematics Assessment (EGMA) (Ketterlin-Geller et al., 2018; Perry, 2020)	EFA	1 factor model, RMSEA = .07–.10, TLI = .70–.85	Low	Cronbach alpha	α = .74–.88	Acceptable
Early Learning Outcomes Measure (ELOM) (Anderson et al., 2021; Snelling et al., 2019)	CFA	1 factor model, RMSEA = .01, CFI = 1.00, SRMR = .01	Acceptable	IRT	Person reliability = .63–.75	Acceptable
Early Measurement Assessment Tool (EMAT) (Ceylan & Aslan, 2023)	MIRT	1 dimension, 2 parameter model, RMSEA = .03, SRMR = .05, CFI = .99	Acceptable	Cronbach alpha	α = .82–.95	Acceptable
Early Years Toolbox-Early Numeracy (Howard et al., 2022)	Unidimensionality	Evidence of unidimensionality for 70 items	NA	NR	NR	NA
Evaluación Neuropsicológica Infantil-Preescolar (ENI-P)-Numerical Abilities Test (Beltrán-Navarro et al., 2018)	NR	NR	NA	Cronbach alpha	α = .48–.96	Mixed
Galileo Early Maths-Revised (Kowalski et al., 2018)	NR	NR	NA	NR	NR	NA

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Heidelberger Rechen Test (HRT) 1–4 (Hassler Hallstedt & Ghaderi, 2018)	NR	NR	NA	NR	NR	NA
KeyMath-Revised (Rhodes et al., 2015)	CFA	1 factor model, RMSEA = .03, CFI, .91	Mixed	Split half reliability	r = .56–75	Mixed
Kieler Kindergartentest Mathematik (KiKi) (Van Hoogmoed et al., 2022)	IRT	3 factor model, BIC = 8704, CAIC = 8762	NA	NR	NR	NA
Mathematical and Arithmetic Competence Diagnostic (MARKO-D) (Bezuidenhout, 2018; Fritz et al., 2014; Henning et al., 2021)	Rasch model	All items within acceptable MNSQ values	NA	IRT	Person reliability = .87–.91	Acceptable
Mathematical Profile (MathPro) Test (Karagiannakis & Noël, 2020)	NR	NR	NA	Cronbach alpha	α = .42–.95	Mixed
Mathematical Reasoning Test (Nunes et al., 2015)	PCA	1 factor model that explained 73.5% of the variance	NA	Cronbach alpha	α = .75	Acceptable

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
mCLASS: Math (Ginsburg et al., 2016)	NR	NR	NA	NR	NR	NA
Neuropsychological Test Battery for Number Processing and Calculation in Children – Revised (NUCALC- R) (Dos Santos et al., 2012; Koumoula et al., 2004)	NR	NR	NA	NR	NR	NA
Number Sense Test (Malofeeva et al., 2004)	NR	NR	NA	Cronbach alpha	α = .93–.97	Acceptable
Numeracy-Caregiver report questionnaire (Pushparatnam et al., 2021)	NR	NR	NA	NR	NR	NA
Numeracy-Child direct assessment (Pushparatnam et al., 2021)	NR	NR	NA	NR	NR	NA
Observing and Analysing Children's Mathematical Development (OAMD) (Bunck et al., 2017)	NR	NR	NA	Cronbach alpha	α = .74–.86	Acceptable

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Parent ratings of numeracy skills (Lin et al., 2021)	CFA	1 factor model, RMSEA = .00, CFI = 1.00, TLI = 1.00	Acceptable	Cronbach alpha	α = .93	Acceptable
Quantitative Reasoning Test (Nunes et al., 2015)	PCA	1 factor model that explained 73.5% of the variance	NA	Cronbach alpha	α = .69–.90	Mixed
Research-Based Early Maths Assessment (REMA) (Clements et al., 2008; Dong et al., 2023)	Rasch model vs. error variance	Separation index = 6.66	NA	Cronbach alpha IRT	α = .71–.89 Person reliability = .93	Acceptable Acceptable
Research-Based Early Maths Assessment-Short Form (REMA-SF) (Weiland et al., 2012)	Rasch model	Infit MNSQ = .73–1.46 Outfit MNSQ = .57–1.46	Acceptable Acceptable	Cronbach alpha IRT	α = .71–.79 Person reliability = .68–.76	Acceptable Mixed
School Achievement Test- 2nd Edition-Arithmetic Subtest (Viapiana et al., 2016)	CFA	2 factor model TLI = .99, RMSEA = .04, SMRS = .04	Acceptable	Cronbach alpha	α = .95	Acceptable
Teacher Rating Scale – Early Numeracy (TRS-EN) (Vessonen et al., 2023)	CFA	3 factor model, RMSEA = .05–.06, CFI = .95–.99, TLI = .94–.99 Family socio-economic status factors not related to teachers' ratings	Acceptable	Cronbach alpha	α = .92–.95	Acceptable

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Teaching Strategies GOLD (Burts & Kim, 2014; Lambert et al., 2014, 2015)	CFA	6 factor model, SRMR = .03, CFI = .93, RMSEA = .07	Low	Cronbach alpha	α = .94–.97	Acceptable
Test of Early Number and Arithmetic (TENA) (Bojorque et al., 2015)	NR	NR	NA	Cronbach alpha	α = .89–.91	Acceptable
The International Development and Early Learning Assessment (IDELA)- Emergent Numeracy (Pisani et al., 2018, 2022; Shavitt et al., 2022; Wolf et al., 2017)	CFA	1 factor model, RMSEA = .02, TLI = .99	Acceptable	Cronbach alpha	α = .52–.92	Mixed
The Utrecht Test of Early Numeracy (ENT) Test (Aunio et al., 2006; David et al., 2015)	NR	NR	NA	Cronbach alpha	α = .79–.90	Acceptable
Tools for Early Assessment in Math Danish Version (DK-TEAM) (Sjoe et al., 2019)	NR	NR	NA	NR	NR	NA
Screeners						

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Arabic number-writing task (Moura et al., 2015)	NR	NR	NA	KR-20 coefficient Split half reliability	KR-20 = .91 *r* = .94	Acceptable Acceptable
Assessing Student Proficiency of Early Number Sense (ASPENS) (Brafford et al., 2023; Sutherland et al., 2021)	NR	NR	NA	NR	NR	NA
Basic Number Processing Test (BNPT) (Olkun et al., 2016)	NR	NR	NA	Cronbach alpha KR-20 coefficient	α = .72–.96 KR-20 = .69–.72	Acceptable Mixed
Birthday Party- Short Version (Lee, 2016)	NR	NR	NA	NR	NR	NA
Dyscalculia screener (Butterworth, 2003)	NR	NR	NA	NR	NR	NA
Dyscalculia Test (Eteng-Uket, 2023)	Unidimensionality	Evidence of unidimensionality for 3 factors	Acceptable	KR-20 coefficient Split half reliability	KR-20 = .93 *r* = .92–.89	Acceptable Acceptable
Early Measurement Curriculum-Based Measures (EM-CBM) (Clarke et al., 2023)	NR	NR	NA	NR	NR	NA

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Early Numeracy (EN)-Test (Hellstrand et al., 2020)	CFA	4 factor model, RMSEA = .03, CFI = .89- .94, TLI = .88- .97 Consistent across language groups	Mixed	Cronbach alpha	α = .91–.95	Acceptable
Early Numeracy Screener (Lopez-Pedersen et al., 2021)	CFA	3 factor model RMSEA = .05, CFI = .94, TLI = .93	Mixed	Cronbach alpha	α = .79–.94	Acceptable
Early Numeracy Skill Indicators (Methe et al., 2008)	NR	NR	NA	KR-20 coefficient	KR-20 = .53–.83	Mixed
House of Numbers (HoN) (Chatzaki et al., 2024)	Rasch model	Infit = .82–1.20 Outfit = .62–1.57	Acceptable Mixed	IRT	Person reliability = .82	Acceptable
Indicators of Basic Early Math Skills (IPAM) (de León et al., 2021; 2022)	CFA	1 factor model RMSEA = .00-.05, CFI = 1.00, SRMR = .01- .02	Acceptable	NR	NR	NA
Math Essential Skill Screener- Elementary Version (MESS-E) (Erford et al., 1998)	Exploratory PCA	1 factor model Eigenvalue = 10.80, % of variance = 37.2%	NA	KR-20 coefficient	KR-20 = .92	Acceptable

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Mathematical School Readiness (MSR) (Mejias et al., 2019)	Unidimensionality	Evidence of unidimensionality for 3 tasks, $r = .28–.49$	Mixed	Cronbach alpha	$\alpha = .63–.95$	Mixed
Number Line Assessment 0–20, 0–100 (Clarke et al., 2020)	NR	NR	NA	Cronbach alpha	$\alpha = .83–.93$	Acceptable
Number Line Assessment 0–100 (Sutherland et al., 2021)	NR	NR	NA	NR	NR	NA
Number Sense Screener (NSS) (English Version) (Jordan et al., 2010, 2012)	Rasch model	Infit = .79–1.31 Outfit = .42–2.17	AcceptableLow	Cronbach alpha	$\alpha > .80$	Acceptable
Number Sense Screener (NSS) (Turkish Version) (Aktulun, 2019; Kiziltepe, 2019)	Rasch model	Infit = .81–1.35 Outfit = .60–1.39	AcceptableAcceptable	Cronbach alpha	$\alpha = .83–.88$	Acceptable
Number Sets Test (Geary et al., 2009)	NR	NR	NA	NR	NR	NA
Numeracy Screener (Bugden et al., 2021; Hawes et al., 2019; Nosworthy et al., 2013)	NR	NR	NA	NR	NR	NA

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Preschool Early Numeracy Skills Screener- Brief Version (PENS-B) (English Version) (Purpura et al., 2015)	Correlation with latent factor score	$r = .94$	Acceptable	Cronbach alpha Split half reliability	$\alpha = .93$ $r = .90$	Acceptable Acceptable
Preschool Early Numeracy Skills Screener- Brief Version (PENS-B) (Greek Version) (Tsigilis et al., 2023)	CFA	2 factor model, RMSEA = .04, CFI = .99	Acceptable	NR	NR	NA
Preschool Numeracy Indicators (Floyd et al., 2006)	NR	NR	NA	NR	NR	NA
Primary Math Assessment Diagnostic (PMA-D) (Brendefur et al., 2018)	NR	NR	NA	Cronbach alpha	$\alpha = .82$- .93	Acceptable
Primary Math Assessment Screener (PMA-S) (Brendefur et al., 2018)	NR	NR	NA	NR	NR	NA
Symbolic Magnitude Processing (SYMP) Test (Brankaer et al., 2017)	NR	NR	NA	NR	NR	NA

Table 3 (continued)

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Universal Screening of Math Skills/ Despiste Universal de Competências Matemáticas (DUCMa) (Cruz et al., 2024)	CFA	1 factor model per subscale, RMSEA = .00–.10, CFI = .97–1.00, TLI = .96–1.00	Mixed	KR-20 coefficient	KR-20 = .80–.90	Acceptable
Measurement tools widely used to assess criterion validity						
Test of Early Mathematics Abilities- 3rd Version (TEMA-3; Ginsburg & Baroody, 2003)	Discrimination Index	r = .45–.68	Acceptable	Cronbach alpha	α = .92–.96	Acceptable
Woodcock-Johnson III Achievement Tests Maths (WJ-III-ACH Maths) (Schrank et al., 2001)	CFA	9 factor model with 3 maths-related clusters (Broad Math; Math Calculation; Math Reasoning) Correlations between achievement clusters, r = .50–.70	Acceptable	Split half reliability	r = .86–.93	Acceptable

BIC Bayesian information criterion, *CAIC* consistent Akaike's information criterion, *CFA* confirmatory factor analysis, *CFI* confirmatory factor index, *IRT* item response theory, *MIRT* multidimensional item response theory, *NA* not applicable, *NR* not reported, *PCA* principal component analysis, *RMSEA* root mean square error approximation, *SRMR* standardised root mean square residual, *TLI* Tucker-Lewis Index

Structural Validity

Twenty-five mathematical assessments included a measure of structural validity, of which confirmatory factor analysis (CFA) was the most frequent approach ($n = 12$). However, only 11 assessments met the common acceptability thresholds and were deemed to have good model fit (see Table 3). Twelve screeners also included a measure of structural validity, of which CFA was also the most common method ($n = 5$) and five screeners met the acceptable threshold criteria (see Table 3).

Internal Consistency

Over half of the mathematical assessments reported internal consistency ($n = 27$) and most reached the acceptable threshold ($n = 20$) (see Table 3). However, of the 20 mathematical assessments with acceptable internal consistency, only two assessments reported disaggregated internal consistency results for the multiple dimensions identified in the structural validity evaluation (Birthday Party- Long Version-English, Lee, 2016; TRS-EN, Vessonen et al., 2023).

Over half of the identified screeners also reported internal consistency ($n = 15$) with 13 meeting the acceptable thresholds. Within those that demonstrated acceptable internal consistency, only three screeners reported internal consistency for the different factors identified in the structural validity evaluation (EN- Test, Hellstrand et al., 2020; Early Numeracy Screener, Lopez-Pedersen et al., 2021; Dyscalculia Test, Eteng-Uket, 2023).

Reliability

Fifteen mathematical assessments included indicators of test–retest reliability (controlled for age) with intervals ranging from 3–7 days to 2–6 months, and nine were rated as acceptable. Eight assessments reported inter-rater reliability, of which seven met the acceptable threshold (see Table 4). Twelve of the identified screeners also included indicators of test–retest reliability (controlled for age) with time intervals ranging from 26.5 days to 17 months. However, only four screeners were rated as having acceptable reliability using these methods (see Table 4).

Criterion Validity

Concurrent validity was evaluated with 24 mathematical assessments, with comparisons most frequently made with the Woodcock-Johnson Math subtests ($n = 7$). However, only 11 mathematical assessments met acceptability thresholds (see Table 4). Divergent validity with standardised language, reading, and non-verbal reasoning measurement tools was considered in seven mathematical assessments, but only two were rated as acceptable. Predictive validity was also considered in seven mathematical assessments, typically over 1–2 years. However, only one was rated as acceptable on the common threshold criteria (TRS-EN, Vessonen et al., 2023).

Table 4 Reliability and Criterion Validity for Identified Mathematical Assessments and Screeners

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Mathematical assessments						
Academic Rating Scale (ARS)-adapted (Kilday et al., 2012)	NR	NR	NA	Concurrent (TEMA-3; M-TEAM)	β = .44–.52	Low
Ani Banani Test (ten Braak & Størksen, 2021)	NR	NR	NA	Concurrent (PENS-B) Predictive (PENS-B; NSMA; end of year)	r = .53 r = .59–.65	Low Mixed
Arithmetic Calculation Efficiency Test (TECA) (Singer & Cuadro, 2014)	Test–retest (interval NR)	r = .85–.94	Acceptable	NR	NR	NA
Assessment of Algebraic Thinking (AAT) (Ralston et al., 2018)	Inter-rater	94%	Acceptable	NR	NR	NA
Birthday Party- Long Version (English) (Lee, 2016)	Test–retest (2 weeks) Inter-rater	r = .24–.82 k = .71–1.00	Mixed Acceptable	Concurrent (YCAT) Predictive (YCAT; 1–2 years)	r = .32–.75 r = .28–.66	Mixed Mixed
Birthday Party- Long Version (Spanish) (Lee, 2016)	NR	NR	NA	Concurrent (YCAT)	r = .19–.69	Mixed
CIRCLE Progress Monitoring (CPM) Math Subtest (Assel et al., 2020)	Test–retest (beginning to middle of school year)	r = .78	Acceptable	Concurrent (WJ-III-AP; TEMA-3) Predictive (WJ-III-AP; TEMA-3; 1–2 years) Divergent (EOWPVT; WJ-III-LW; WJ-III-PC)	r = .65 r = .55 r = .42–.61	Acceptable Low Mixed
Cognitive Diagnostic Test (Li et al., 2020)	NR	NR	NA	Concurrent (WJ-IV-AP; WJ-IV-C)	r = .62–.77	Acceptable

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Comprehensive Learning Test-Mathematics (CLT-M) (Lee et al., 2017)	Test–retest (2 weeks)	$r = .87$	Acceptable	NR	NR	NA
Comprehensive Research-Based Early Math Ability Test (CREMAT) (Clements et al., 2022)	NR	NR	NA	NR	NR	NA
Curriculum Based Measures for Kindergarten- Grade 3 (Lee & Lembke, 2016)	Test–retest (2 weeks)	$r = .36–.86$	Mixed	Concurrent (WJ-III-Math)	$r = .14- .58$	Low
DIFER School Readiness Test Battery Counting and Basic Numeracy (Csapó et al., 2014; Józsa et al., 2023)	NR	NR	NA	Divergent (other DIFER domains)	$r = .69–.77$	Acceptable
Early Arithmetic, Reading and Learning Indicators (EARLI) – Numeracy measures (Cheng et al., 2017; Lei et al., 2009)	NR	NR	NA	Concurrent (WJ-III-AP; WJ-III-QC)	$r = .32–.83$	Mixed

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Early Grade Mathematics Assessment (EGMA) (Ketterlin-Geller et al., 2018; Perry, 2020)	NR	NR	NA	NR	NR	NA
Early Learning Outcomes Measure (ELOM) (Anderson et al., 2021; Snelling et al., 2019)	Test–retest (1 week) Inter-rater	$r = .90$ $k = .68–.92$	Acceptable Mixed	Concurrent (WPPSI-IV)	$r = .64$	Acceptable
Early Measurement Assessment Tool (EMAT) (Ceylan & Aslan, 2023)	Test–retest (6 weeks)	$r = .91$	Acceptable	Concurrent (TEMA-3)	$r = .76–.84$	Acceptable
Early Years Toolbox-Early Numeracy (Howard et al., 2022)	Test–retest (1 week)	$r = .89$	Acceptable	Concurrent (DAS; PENS)	$r = .74–.80$	Acceptable
Evaluación Neuropsicológica Infantil-Preescolar (ENI-P)-Numerical Abilities Test (Beltrán-Navarro et al., 2018)	Test–retest (15 days)	$r = .30–.84$	Mixed	NR	NR	NA
Galileo Early Maths-Revised (Kowalski et al., 2018)	NR	NR	NA	Concurrent (Child-direct tasks)	$r = .47$	Low

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Heidelberger Rechen Test (HRT) 1–4 (Hassler Hallstedt & Ghaderi, 2018)	Test–retest (3–7 days)	$r = .29–.82$	Mixed	Concurrent (Math Battery)	$r = .67–.82$	Acceptable
KeyMath-Revised (Rhodes et al., 2015)	NR	NR	NA	NR	NR	NA
Kieler Kindergartentest Mathematik (KiKi) (Van Hoogmoed et al., 2022)	NR	NR	NA	Concurrent (ENT-R)	$r = .72$	Acceptable
Mathematical and Arithmetic Competence Diagnostic (MARKO-D) (Bezuidenhout, 2018; Fritz et al., 2014; Henning et al., 2021)	NR	NR	NA	NR	NR	NA
Mathematical Profile (MathPro) Test (Karagiannakis & Noël, 2020)	NR	NR	NA	Concurrent (Standardized maths test)	$r = .47–.64$	Mixed
Mathematical Reasoning Test (Nunes et al., 2015)	NR	NR	NA	NR	NR	NA
mCLASS: Math (Ginsburg et al., 2016)	Inter-rater	$k = .76- .95$	Acceptable	Concurrent (WJ-III-Math)	$r = .50–.61$	Mixed

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Neuropsychological Test Battery for Number Processing and Calculation in Children – Revised (NUCALC-R) (Dos Santos et al., 2012; Koumoula et al., 2004)	NR	NR	NA	Concurrent (WISC-III-A) Divergent (ATHENA; WISC-III- DS)	$r = .41–.64$ $r = .45–.52$	Mixed Low
Number Sense Test (Malofeeva et al., 2004)	NR	NR	NA	Divergent (WPPSI-Vocabulary)	$r = .33–.54$	Low
Numeracy-Caregiver report questionnaire (Pushparatnam et al., 2021)	NR	NR	NA	NR	NR	NA
Numeracy-Child direct assessment (Pushparatnam et al., 2021)	NR	NR	NA	NR	NR	NA
Observing and Analysing Children's Mathematical Development (OAMD) (Bunck et al., 2017)	Test–retest (2–6 months)	$r = .47$	Low	Concurrent (CMT; CKT)	$r = .39–.50$	Low
Parent ratings of numeracy skills (Lin et al., 2021)	NR	NR	NA	Concurrent (Child direct tasks; PENS-B)	$r = .31–.56$	Low

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Quantitative Reasoning Test (Nunes et al., 2015)	Test–retest (4.5 months)	$r = .78$	Acceptable	Predictive (Math Reasoning Test; interval NR)	$\Delta R^2 = .06–.24$	Low
Research-Based Early Maths Assessment (REMA) (Clements et al., 2008; Dong et al., 2023)	Inter-rater	98%	Acceptable	NR	NR	NA
Research-Based Early Maths Assessment-Short Form (REMA-SF) (Dong et al., 2021; Weiland et al., 2012)	Test–retest (1 term)	$r = .90$	Acceptable	Concurrent (REMA; WJ-III-AP) Divergent (PPVT-III; WJ-III-LW)	$r = 71–.74$ $r = .64$	Acceptable Acceptable
School Achievement Test-2nd Edition-Arithmetic Subtest (Viapiana et al., 2016)	NR	NR	NA	NR	NR	NA
Teacher Rating Scale – Early Numeracy (TRS-EN) (Vessonen et al., 2023)	NR	NR	NA	Concurrent (ENT) Predictive (ENT; 1 year)	$r = .60$ $r = .63–.69$	Acceptable Acceptable
Teaching Strategies GOLD (Lambert et al., 2014, 2015; Russo et al., 2019; Vitiello & Williford, 2021)	Inter-rater	$k = .92$	Acceptable	Concurrent (BBCS-R; WJ-III) Predictive (WJ-III; 1 term) Divergent (other TS Gold domains)	$r = .27–.74$ $r = .39$ $r = .13–.44$	Mixed Low Low

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Test of Early Number and Arithmetic (TENA) (Bojorque et al., 2015)	Inter-rater	$k = .92$	Acceptable	Concurrent (ENT)	$r = .80–.89$	Acceptable
The International Development and Early Learning Assessment (IDELA)- Emergent Numeracy (Pisani et al., 2018, 2022; Shavitt et al., 2022)	Inter-rater Test–retest (3 weeks)	$k = .88–.99$ ICC = .66	AcceptableLow	Concurrent (EGMA) Predictive (EGMA; 2 years)	$r = .63$ $R^2 = .55$	AcceptableLow
The Utrecht Test of Early Numeracy (ENT) Test (Aunio et al., 2006; David et al., 2015)	Test–retest (3 months)	$r = .75$	Acceptable	Divergent (Raven's progressive matrices)	$r = .47$	Low
Tools for Early Assessment in Math Danish Version (DK-TEAM) (Sjoe et al., 2019)	Test–retest (4.5 months)	$r = .43–93$	Mixed	NR	NR	NA
Screeners						
Arabic number-writing task (Moura et al., 2015)	NR	NR	NA	Diagnostic accuracy	Sensitivity = .58–.85 Specificity = .28–.88	Low Mixed

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Assessing Student Proficiency of Early Number Sense (ASPENS) (Brafford et al., 2023; Sutherland et al., 2021)	Test–retest (interval NR)	$r = .71–.87$	Acceptable	Diagnostic accuracy Concurrent (TerraNova) Predictive (TerraNova; end of year)	Sensitivity = .91 Specificity = .83 $r = .56$ $r = .50–.53$	Acceptable Acceptable Low Low
Basic Number Processing Test (BNPT) (Olkun et al., 2016)	NR	NR	NA	Concurrent (MAT; CPT)	$r = -.64–-.20$	Mixed
Birthday Party- Short Version (Lee, 2016)	NR	NR	NA	NR	NR	NA
Dyscalculia screener (Butterworth, 2003)	NR	NR	NA	NR	NR	NA
Dyscalculia Test (Eteng-Uket, 2023)	NR	NR	NA	NR	NR	NA
Early Measurement Curriculum-Based Measures (EM-CBM) (Clarke et al., 2023)	Test–retest (10 weeks)	$r = .33–.48$	Low	Concurrent (ASPENS) Predictive (EasyCBM; 10 weeks)	$r = .25–.43$ $r = -.07–.44$	Low Low
Early Numeracy (EN)-Test (Hellstrand et al., 2020)	NR	NR	NA	NR	NR	NA
Early Numeracy Screener (Lopez-Pedersen et al., 2021)	NR	NR	NA	Predictive (Norwegian national test scores; 6 months)	$r = .19–.25$	Low

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Early Numeracy Skill Indicators (Methe et al., 2008)	Test–retest (13 weeks)	$r = .68–.98$	Mixed	Diagnostic accuracy Concurrent (TEMA-3) Predictive (TEMA-3; end of year)	58–84% correct classification $r = .20–.72$ $r = .41–.70$	Low Mixed Mixed
House of Numbers (HoN) (Chatzaki et al., 2024)	NR	NR	NA	Diagnostic accuracy	Sensitivity = .91–1.00 Specificity = .85–.93	Acceptable Acceptable
Indicators of Basic Early Math Skills (IPAM) (de León et al., 2021; 2022)	Test–retest (3 months)	$r = .43–.67$	Mixed	Concurrent (Sn-BADyG) Predictive (Sn-BADyG; end of year)	$r = .48–.60$ $r = .36–.58$	Mixed Low
Math Essential Skill Screener- Elementary Version (MESS-E) (Erford et al., 1998)	Test–retest (30 days)	$r = .86$	Acceptable	Diagnostic accuracy Concurrent (WJ-R; WRAT-R; KeyMath-R)	Sensitivity = .98 Specificity = .88 $r = .49–.80$	Acceptable Acceptable Mixed
Mathematical School Readiness (MSR) (Mejias et al., 2019)	NR	NR	NA	Concurrent (TTR; KRT-R)	$r = .56$	Low
Number Line Assessment 0–20, 0–100 (Clarke et al., 2020)	Test–retest (interval NR)	$r = .70–.72$	Acceptable	NR	NR	NA
Number Line Assessment 0–100 (Sutherland et al., 2021)	Test–retest (8 months)	$r = .58$	Low	Diagnostic accuracy	Sensitivity = .69–.91 Specificity = .39–.81	Mixed Mixed
Number Sense Screener (NSS) (English Version) (Jordan et al., 2010, 2012)	Test–retest (12 months) Test–retest (17 months)	$r = .70–86$ $r = .61–.69$	Acceptable Low	Diagnostic accuracy Predictive (WJ-III-Math; 2 years) Divergent (DIBELS; WASI; Digit Span)	Sensitivity = .70–.86 Specificity = .35–.85 $r = .70–.72$ $r = .12–.56$	Mixed Mixed Acceptable Low

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Number Sense Screener (NSS) (Turkish Version) (Aktulun, 2019; Kiziltepe, 2019)	NR	NR	NA	NR	NR	NA
Number Sets Test (Geary et al., 2009)	NR	NR	NA	Diagnostic accuracy Predictive (Third Grade Maths Achievement; 3 years)	Sensitivity = .69 Specificity = .67 Estimate = 5.01	Low Low NA
Numeracy Screener (Bugden et al., 2021; Hawes et al., 2019; Nosworthy et al., 2013)	Test–retest (M = 89.55 days)	r = .61–.72	Mixed	Diagnostic accuracy Concurrent (WJ-III-Maths) Divergent (WJ-III-Reading) Predictive (school maths grades; 1 year)	Sensitivity = .62 Specificity = .87 r = .22–.25 r = .15–.19 r = .23–.31	Low Acceptable Low Low Low
Preschool Early Numeracy Skills Screener-Brief Version (PENS-B) (English Version) (Purpura et al., 2015)	NR	NR	NA	Concurrent (TEMA-3) Divergent (GRTR; EOWPVT)	r = .73 r = .60–.63	AcceptableAcceptable
Preschool Early Numeracy Skills Screener- Brief Version (PENS-B) (Greek Version) (Tsigilis et al., 2023)	Marginal reliability index	.79–.82	Acceptable	NR	NR	NA
Preschool Numeracy Indicators (Floyd et al., 2006)	Test–retest (M = 26.5 days)	r = .32–.92	Mixed	Concurrent (BBCS-R; WJ-III-AP; TEMA-3)	r = .29–70	Mixed

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Primary Math Assessment Diagnostic (PMA-D) (Brendefur et al., 2018)	NR	NR	NA	NR	NR	NA
Primary Math Assessment Screener (PMA-S) (Brendefur et al., 2018)	NR	NR	NA	NR	NR	NA
Symbolic Magnitude Processing (SYMP) Test (Brankaer et al., 2017)	Test–retest (interval NR)	$r = .62–.77$	Mixed	Diagnostic accuracy Concurrent (standardized maths test)	TD children (≥ 35th percentile on standardized maths test) consistently outperformed MLD children (≤ the 10th percentile), except 10–11 years $r = .16–.40$	Acceptable Low
Universal Screening of Math Skills/Despiste Universal de Competências Matemáticas (DUCMa) (Cruz et al., 2024)	Test–retest (start and end of school year)	$r = .39–.72$	Mixed	Predictive (school maths grades; 1 year)	$r = .27–.45$	Low
Measurement tools widely used to assess criterion validity						

Table 4 (continued)

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Test of Early Mathematics Abilities-3rd Version (TEMA-3; Ginsburg & Baroody, 2003)	Test–retest (2 weeks)	$r = .82–.93$	Acceptable	Concurrent (KeyMath-R/NU; WJ-III-Math; DAB-3; YCAT)	$r = .54–.91$	Mixed
Woodcock-Johnson III Achievement Tests Maths (WJ-III-ACH Maths) (Schrank et al., 2001)	Test–retest (1 day)	$r = .91–.95$	Acceptable	Concurrent (Kaufman Test of Educational Achievement maths sub-tests; WIAT maths sub-tests)	$r = .29–.70$	Mixed

MLD mathematical learning difficulties, *NA* not applicable, *NR* not reported, *TD* typically developing; tests used to establish criterion validity: *BBCS-R* Bracken Basic Concept Scale-Revised (Panter & Bracken, 2009), *CKT* Dutch Cito Mathematics Test for Kindergarten (Koerhuis, 2010), *CMT* Dutch Cito Mathematics Test for Grades 1–3 (Janssen et al., 2005a, b, 2006), *CPT* Calculation Performance Test (Olkun et al., 2013), *DAB-3* diagnostic achievement battery – Third Edition (Newcomer, 2001), *DAS* differential ability scales, Early Number Concepts Scale (Elliott et al., 2007), *DIBELS* dynamic indicators of basic early literacy skills–Sixth Edition (Good et al., 2002), *ENT-R* early numeracy test-revised (Van Luit & Van de Rijt, 2009), *EOWPVT* expressive one-word picture vocabulary test (Martin & Brownell, 2011), *GRTR* get ready to read (Lonigan & Wilson, 2008), *KeyMath-R/NU* KeyMath-revised: a diagnostic inventory of essential mathematics-normative update (Connolly, 1988), *KRT-R* Kortrijkse Rekentest-Revisie (Baudonck et al., 2006), *MAT* math achievement test (Fidan, 2013), *M-TEAM* math battery (Fuchs et al., 2003) modified tools for early assessment in mathematics (based on Clements et al., 2011), *NSMA* national school maths assessment, *PENS* preschool early numeracy skills screener (Purpura & Lonigan, 2015), *PENS-B* preschool early numeracy skills screener-brief version (Purpura et al., 2015), *PPVT-III* Peabody picture vocabulary test III (Dunn & Dunn, 1997), *Sn-BADyG* numerical computation measure of the Battery of Differential and General Skills (Yuste-Hernanz, 2002), *TEMA-3* test of early mathematics abilities-3rd version (Ginsburg & Baroody, 2003), *TTR* tempo test Rekenen (De Vos, 1992), *WASI* Wechsler abbreviated scale of intelligence (Wechsler, 1999), *WIAT* Wechsler individual achievement test (Wechsler, 1992), *WISC-III-A* Wechsler intelligence scale for children–third edition-arithmetic (Wechsler, 1997), *WISC-III-DS* Wechsler intelligence scale for children–third edition-digit span (Wechsler, 1997), *WJ-III-AP* Woodcock-Johnson III tests of achievement, applied problems subtest (McGrew et al., 2007; Woodcock et al., 2001), *WJ-III-LW* Woodcock-Johnson III tests of achievement, letter-word identification subtest (McGrew et al., 2007), *WJ-III-PC* Woodcock-Johnson III tests of achievement, passage comprehension task (McGrew et al., 2007), *WJ-III-QC* Woodcock-Johnson III tests of achievement, quantitative concepts subtest (Woodcock et al., 2001), *WJ-IV-AP* Woodcock-Johnson IV tests of achievement, applied problems subtest (Schrank et al., 2014), *WJ-IV-C* Woodcock-Johnson IV tests of achievement, calculation subtest (Schrank et al., 2014), *WJ-R* Woodcock-Johnson tests of achievement-revised mathematics cluster (Woodcock & Johnson, 1989), *WPPSI* Wechsler preschool and primary intelligence scales (Wechsler, 1967), *WRAT-R* wide-range achievement test-revised level 1 arithmetic subtest (Jastak & Wilkinson, 1984), *YCAT* young children's achievement test (Hresko et al., 2000)

relevant to education measures. Only 15 measurement tools met the common acceptability thresholds for more than two areas of psychometric evidence. Thirdly, only four screeners demonstrated an acceptable ability to distinguish between typically developing children and those with or at-risk of MLD. Fourthly, only one mathematical assessment and one screener met the common acceptability threshold for predictive validity. Finally, only 11 mathematical assessments and one screener were found to concurrently align with other validated measurement tools. Directions for future research based on these five main findings will be discussed. Overall, this study is relevant to researchers, practitioners, and other stakeholders interested in the effective use of measurement tools to assess young children's mathematical skills over time, in response to interventions, and/or to reliably identify children with or at-risk of MLD.

Overview of Measurement Tools

Firstly, the current review showed that most measurement tools were categorised as child-direct measures delivered individually with a trained assessor in a paper-based format. Most measurement tools targeted number and/or arithmetic skills, with fewer tools measuring shape, space, and measure skills. Although the identified measurement tools were evaluated in 55 countries and 31 languages, most assessments and screeners were developed in WEIRD societies in minority countries and/or in English. Only ten assessments and three screeners were evaluated in more than one country (see Table 2). This poses an ongoing challenge for the field of mathematical learning and development as the underrepresentation of multilingual majority countries (i.e. non-WEIRD societies) in test development leads to publication bias and a lack of scientific evidence related to children's learning in various countries (Draper et al., 2022).

Psychometric Evaluations of the Identified Measurement Tools

Secondly, the majority of the identified measurement tools have not been evaluated for aspects of reliability and validity most relevant to education measures, and few tools met the common acceptability thresholds for these indicators. For example, only nine assessments (DIFER, Csapó et al., 2014; Early Years Toolbox, Howard et al., 2022; ELOM, Snelling et al., 2019; EMAT, Ceylan & Aslan, 2023; ENT, Van Luit et al., 1994; Van de Rijt et al., 2003; IDELA, Save the Children, 2019; Parent Ratings of Numeracy Skills, Lin et al., 2021; REMA-SF, Weiland et al., 2012; TRS-EN, Vessonen et al., 2023) and six screeners (ASPENS, Clarke et al., 2011; Dyscalculia Test, Eteng-Uket, 2023; HoN, Chatzaki et al., 2024; MESS-E, Erford et al., 1998; NSS [English version], Jordan et al., 2010; PENS-B, Purpura et al., 2015) were identified to meet the common acceptability thresholds for more than two areas of psychometric evidence (see Table 5). These findings suggest that these 15 measurement tools currently have the most promising psychometric evidence to assess young children's mathematical skills and/or to reliably identify children with or at-risk of MLD.

Table 5 Mathematical assessments and screeners identified in the current review to have multiple dimensions of acceptable psychometric evidence

Measurement Tool	Acceptable Psychometric Evidence					
	Co.V	SV	IC	R	Cr.V	Total
Mathematical assessments						
Early Measurement Assessment Tool (EMAT)	✓	✓	✓	✓	✓	5
Early Learning Outcomes Measure (ELOM)	-	✓	✓	✓	✓	4
The International Development and Early Learning Assessment (IDELA)-Emergent Numeracy	✓	✓	-	✓	✓	4
Research-Based Early Mathematics Assessment- Short Form (REMA-SF)	-	✓	✓	✓	✓	4
Teacher Rating Scale – Early Numeracy (TRS-EN)	-	✓	✓	-	✓	3
The Utrecht Test of Early Numeracy (ENT) Test	✓	-	✓	✓	-	3
DIFER School Readiness Test Battery Counting and Basic Numeracy	-	-	✓	-	✓	2
Early Years Toolbox-Early Numeracy	-	-	-	✓	✓	2
Parent ratings of numeracy skills	-	✓	✓	-	-	2
Screeners						
Number Sense Screener (NSS) (English Version)	✓	-	✓	✓	✓	4
Dyscalculia Test	✓	✓	✓	-	-	3
Math Essential Skill Screener-Elementary Version (MESS-E)	-	-	✓	✓	✓	3
Preschool Early Numeracy Skills Screener-Brief Version (PENS-B) (English Version)	-	✓	✓	-	✓	3
Assessing Student Proficiency of Early Number Sense (ASPENS)	-	-	-	✓	✓	2
House of Numbers (HoN)	-	-	✓	-	✓	2
Measurement tools widely used to assess criterion validity						
Test of Early Mathematical Ability-3rd Edition (TEMA-3)	✓	✓	✓	✓	-	4
Woodcock-Johnson III Achievement Tests Maths (WJ-III-ACH Maths)	✓	✓	✓	✓	-	4

Co.V content validity, *Cr.V* criterion validity, *IC* internal consistency, *R* reliability, *SV* structural validity

Although it would be preferable for more measurement tools to meet these criteria, the current findings, combined with the practical information summarised in Table 2, offer a useful starting point for other researchers to decide which early maths measurement tool to use in their work. For example, the broad skill focus included in the Early Years Toolbox-Early Numeracy (Howard et al., 2022) may be suitable for consideration in a maths intervention study with English-speaking children aged 3–4 years (e.g. Scerif et al., 2023). Whereas the ease with which children's early maths skills can be indicated by parent reports in the parent ratings of numeracy skills assessment (Lin et al., 2021) may be better suited for large-scale, survey studies (e.g. Cosso et al., 2024 adapted this measure for use with Latine families in the USA).

Identifying Children with or At-Risk of MLD

Thirdly, in terms of diagnostic validity for identifying children with or at-risk of MLD, only the ASPENS (Clarke et al., 2011), HoN (Chatzaki et al., 2024), and MESS-E (Erford et al., 1998) screeners were found to have acceptable sensitivity and specificity. In addition, the SYMP Test (Brankaer et al., 2017) also demonstrated an acceptable ability to distinguish between typically developing children and those with MLD. Although the Numeracy Screener (Nosworthy et al., 2013) demonstrated specificity greater than 0.70, the sensitivity results were below the common acceptability threshold of 0.90. Establishing strong sensitivity in measurement tools is important for accurately identifying true cases of children with or at-risk of MLD and reducing the risk of missing those most in need (Jenkins et al., 2007; Klingbeil et al., 2019).

Fourthly, predictive validity can also be used to evaluate the suitability of measurement tools for detecting children with or at-risk of MLD over time. This study found that only seven mathematical assessments and nine screeners included evaluations of predictive validity, and only two measures met the common acceptability threshold (NSS; Jordan et al., 2012; TRS-EN, Vessonen et al., 2023). However, these results may, in part, be due to issues relating to consistencies with the external measurement tool or criteria. For example, the Early Numeracy Screener showed low predictive validity with the Norwegian national test scores measured 6 months later (Lopez-Pedersen et al., 2021). In explaining these results, the authors highlighted inconsistencies in the types of items across the two measurement tools; while the Early Numeracy Screener includes untimed items and emphasises accuracy, the national test has timed items and focuses on fluency.

Criterion Validity with Other Validated Measures

Finally, this study found that only 11 of the mathematical assessments and one of the screeners concurrently aligned with other validated measures of early mathematical skills (see Table 4). While establishing the criterion validity of assessments and screeners with other validated measures is considered an important component of the measurement development process (AERA et al., 2014; Mokkink et al., 2016;

Prinsen et al., 2018), it remains an ongoing challenge within the field of mathematical learning and development. For example, the credibility of the criterion validity evaluation relies on the relevance, reliability, and validity of the other measures used as the basis for the concurrent comparison. In particular, the two measurement tools must be conceptually aligned (AERA et al., 2014). In the current review, the identified measurement tools that did show acceptable levels of concurrent validity were compared to a broad range of measures (see Table 4), of which the Test of Early Mathematics Abilities-3rd Version (TEMA-3; Ginsburg & Baroody, 2003) and the Woodcock-Johnson III Math subtests (Schrank et al., 2001; Woodcock & Johnson, 1989; Woodcock et al., 2001) were the most widely used.

Most of the identified measures that demonstrated acceptable (concurrent) criterion validity when compared to the TEMA-3 or Woodcock-Johnson Math subtests, broadly speaking, measured similar areas of mathematical development. For example, the PENS-B (Purpura et al., 2015) and the TEMA-3 focused on number and arithmetic skills (Ginsburg & Baroody, 2003), while the REMA-SF (Weiland et al., 2012), CPM (Assel et al., 2020), and the Woodcock-Johnson III Applied Problems Math subtest also included shape, space, and measure items (Schrank et al., 2001; Woodcock et al., 2001). Similarly, most identified measurement tools that did not demonstrate acceptable (concurrent) criterion validity did not conceptually align with the TEMA-3 ($n=4$) or Woodcock-Johnson Math subtests ($n=6$) (see Tables 2 and 4). Issues relating to the limited conceptual alignment between measurement tools may be further exacerbated by the lack of consensus relating to the complex structure of early maths (Devlin et al., 2022; Gilmore, 2023) and the inconsistencies in the terminology used to describe the mathematical skills children need to acquire in early childhood.

Furthermore, although the Woodcock-Johnson Math subtests are also available in Spanish (Muñoz-Sandoval et al., 2009) and the TEMA-3 is translated into Mandarin, Spanish, and Dutch (e.g. Paik et al., 2011; Huang et al., 2022) with psychometric evaluations conducted in China, Singapore, and Spain (Ginsburg & Baroody, 2007; Kang et al., 2014; Yao et al., 2017), these tools are not widely available in a range of different languages and cultures. They also require a trained assessor for administration, as well as substantial costs to purchase the necessary materials, which may limit their usability.

To address some of these challenges, other measurement tools, such as the Early Grade Mathematics Assessment (EGMA; RTI International, 2014), have recently been adapted for self-administration (i.e. does not require a trained assessor). The SA-EGMA is a child-direct assessment administered on solar-powered, touch-screen tablet devices and requires minimal adult supervision (Pitchford & Outhwaite, 2016). It has been piloted in Ghana, Sierra Leone, and Liberia (all English-speaking) with forthcoming adaptations for Malawi and French-speaking countries (Ryan, 2023). However, the psychometric properties of the SA-EGMA are yet to be reported.

Directions for Future Research

Based on these five main findings, there are four recommendations for future research to improve the psychometric evidence and availability of measurement tools for

mathematics in early childhood. Firstly, future research should focus on developing and reporting the reliability and validity evidence of a broad range of existing measurement tools. This research should aim to establish a set of 'Gold Standard' measurements in the field of mathematical learning and development. These measurement tools should span across different ages, mathematical skill areas, and different measurement types (i.e. child-direct and parent/teacher-report), which can be used for different study design purposes (e.g. large-scale longitudinal designs and intervention studies). Overall, this will contribute to improving the methodological rigour of this field.

Secondly, the development of these measurement tools should aim to be inclusive of different languages, countries, and cultures. The current study highlights successful examples where measurement tools have been adapted and/or translated for use in different educational contexts (e.g. Pushparatnam et al., 2021; Save the Children, 2019; Van Luit et al., 1994; Van de Rijt et al., 2003). For example, these studies highlight the value of collaborations with country-specific teams to ensure the measurement tool is contextually relevant, adaptable (e.g. translation-back translation procedures), feasible with assessors, and appropriate for use with children across different countries and/or cultural contexts (Pisani et al., 2018). Future research in this area should also work towards open-access measurement tools that practitioners can use (Hakkarainen et al., 2023) and other researchers in low-resource contexts (Pitchford & Outhwaite, 2016). This will help facilitate greater representation of multilingual, majority countries (i.e. non-WEIRD societies) in mathematical learning and development research (Draper et al., 2022).

Thirdly, while enhancing existing measures should be prioritised, future research should also seek to develop new measurement tools that utilise innovative technologies. For example, the current study highlights that technology-based, self-administered measurement tools can increase access and participation of marginalised and hard-to-reach groups in research (Ryan, 2023). Future research should advance these recent developments and evaluate whether digital measurement tools are reliable and valid in early childhood, particularly with very young children. These new approaches to measurement tools will require interdisciplinary collaborations, including psychologists, education professionals, and software engineers. It will also require co-production with end-users, such as researchers, teachers, and parents (Duraiappah et al., 2022).

Finally, to support the development of existing and new measurement tools, future research should also work towards a commonly accepted definition of the structure of early mathematics (Devlin et al., 2022; Gilmore, 2023). This will elucidate which skills should be included in these measurement tools. Furthermore, an understanding of the maths skills included within measurement tools, using common terminology, can support theoretical insights into the processes and mechanisms involved in early mathematical development.

Limitations of the Current Review

Although this study conducted a systematic search of the literature to identify measurement tools for early mathematical skills, not every available measure was included in

the current synthesis. This was because the search strategy was designed to identify studies that had evaluated the psychometric properties of measurement tools, rather than identifying measurement tools based on their use in intervention, longitudinal, or other studies. Future synthesises should incorporate this broader search strategy, as well as qualitative methods with the mathematical learning and development research and practitioner communities to establish which measurement tools are most widely used in the field, and why. This will provide a more in-depth understanding of the best practices and challenges when measuring mathematical skills in early childhood.

Similarly, the current review was affected by publication bias as the search strategy only incorporated full-text studies that were available in English. Although the current review identified measurement tools that are available in 31 languages, some measures, such as the Tempo Test Rekenen (TTR; De Vos, 1992), were excluded from the current review. This was because the studies and/or test manuals, which reported the psychometric properties of these measures, were only available in other languages, such as in French in the case of the TTR (Lafay et al., 2020). To address this bias, future studies should seek to systematically review measurement tools that are specifically available in languages other than English. This will contribute to efforts to increase diverse representation in child development research. The current review also does not include a quality assessment of the included studies (e.g. sample size and characteristics, analytical methods justified and appropriate). This should also be incorporated into future research, alongside the quality assessments of other identified measurement tools (e.g. using the COSMIN taxonomy).

Conclusion

This pre-registered systematic review is the first study to provide an overview of mathematical measurement tools for children aged 0–8 years and a synthesis of the reported reliability and validity evidence, including in relation to common acceptability thresholds. Although a relatively large number of assessments ($n=41$) and screeners ($n=25$) were identified in the current review, significant gaps remain in the appraisal of these measurement tools. Building on this evidence and improving measurement quality is vital to raising methodological standards in mathematical learning and development research.

Acknowledgements We would like to thank Marcella Lam, Victoria Levy, and Camilla Mendizabal for their assistance in record screening and data extraction.

Data Availability Data is available by reasonable request to the first author.

Declarations

Ethical Approval Ethical approval for this study was granted by the IOE ethics committee (REC1689).

Conflict of Interest The authors declare no competing interests.

References

*89 studies identified through systematic review

AERA, APA, & NCME.(2014). *Standards for educational and psychological testing.* American Psychological Association.

*Aktulun, O. U. (2019). Validity and reliability study of Turkish version of number sense screener for children aged 72-83 months. *Journal of Education and Training Studies, 7*(2), 64-75.

*Anderson, K. J., Henning, T. J., Moonsamy. J. R., Scott, M., du Plooy, C., & Dawes, A. R. L. (2021). Test-retest reliability and concurrent validity of the South African Early Learning Outcomes Measure (ELOM).*South African Journal of Childhood Education, 11*(1), 1-9.

*Assel, M. A., Montroy, J. J., Williams, J. M., Foster, M., Landry, S. H., Zucker, T., Crawford, A., Hyatt, H., & Bhavsar, V. (2020). Initial validation of a math progress monitoring measure for prekindergarten students. *Journal of Psychoeducational Assessment, 38*(8), 1014-1032.

Aubrey, C., Godfrey, R., & Dahl, S. (2006). Early mathematics development and later achievement: Further evidence. *Mathematics Education Research Journal, 18*, 27–46.

*Aunio, P., Hautamäki, J., Heiskari, P. & Van Luit, J. E. H. (2006). The early numeracy test in Finnish: Children's norms. *Scandinavian Journal of Psychology, 47,* 369-378.

Aunio, P., & Räsänen, P. (2016). Core numerical skills for learning mathematics in children aged five to eight years–A working model for educators. *European Early Childhood Education Research Journal, 24*(5), 684–704.

Bailey, D. H., Oh, Y., Farkas, G., Morgan, P., & Hillemeier, M. (2020). Reciprocal effects of reading and mathematics? Beyond the cross-lagged panel model. *Developmental Psychology, 56*(5), 912.

Bartelet, D., Ansari, D., Vaessen, A., & Blomert, L. (2014). Cognitive subtypes of mathematics learning difficulties in primary education. *Research in Developmental Disabilities, 35*(3), 657–670.

Baudonck, M., Debusschere, A., Dewulf, B., Samyn, F., Vercaemst, V., & Desoete, A. (2006). Kortrijkse Rekentest-Revisie [Revised Kortrijk Arithmetic Test]. *Kortrijk, Belgium: Revalidatiecentrum Overleie*.

Beller, S., & Jordan, F. (2018). The cultural challenge in mathematical cognition. *Journal of Numerical Cognition, 4*(2), 448–463.

*Beltrán-Navarro, B., Abreu-Mendoza, R. A., Matute, E., & Rosselli, M. (2018). Development of early numerical abilities of Spanish-speaking Mexican preschoolers: A new assessment tool. *Applied Neuropsychology: Child, 7*(2), 117-128.

*Bezuidenhout, H. S. (2018). Diagnostic test for number concept development during early childhood. *South African Journal of Childhood Education, 8*(1), 1-10.

*Bojorque, G., Torbeyns, J., Moscoso, J., Van Nijlen, D., & Verschaffel, L. (2015). Early number and arithmetic performance of Ecuadorian 4-5-year-olds. *Educational Studies, 41*(5), 565-586.

Braeuning, D., Ribner, A., Moeller, K., & Blair, C. (2020). The multifactorial nature of early numeracy and its stability. *Frontiers in Psychology, 11*, 518981.

*Brafford, T., Clarke, B., Gersten, R. M., Smolkowski, K., Sutherland, M., Dimino, J., & Fainstein, D. (2023). Exploring an early numeracy screening measure for English learners in primary grades. *Early Childhood Research Quarterly*, *63*, 278-287.

*Brankaer, C., Ghesquière, P., & De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test). *Behavior Research Methods*, *49*, 1361-1373.

*Brendefur, J. L., Johnson, E. S., Thiede, K. W., Strother, S., & Severson, H. H. (2018). Developing a multi-dimensional early elementary mathematics screener and diagnostic tool: The primary mathematics assessment. *Early Childhood Education Journal*, *46*, 153-157.

*Bugden, S., Peters, L., Nosworthy, N., Archibald, L., & Ansari, D. (2021). Identifying children with persistent developmental dyscalculia from a 2-min test of symbolic and nonsymbolic numerical magnitude processing. *Mind, Brain, and Education*, *15*(1), 88-102.

*Bunck, M. J. A., Terlien, E., van Groenestijn, M., Toll, S. W. M., & Van Luit, J. E. H. (2017). Observing and analyzing children's mathematical development, based on action theory. *Educational Studies in Mathematics*, *96*, 289-304.

*Burts, D. C., & Kim, D. H. (2014). The teaching strategies GOLD assessment system: Measurement properties and use. *HS Dialog: The Research to Practice Journal for the Early Childhood Field*, *17*(3). https://doi.org/10.55370/hsdialog.v17i3.170

*Butterworth, B. (2003). *Dyscalculia screener*. NferNelson Pub.

Butterworth, B. (2005). Developmental dyscalculia. In *The Handbook of Mathematical Cognition* (pp. 455–467). Psychology Press.

*Ceylan, M., & Aslan, D. (2023). Length, area, volume, and angle and turn measurement in early childhood: Validating the early measurement assessment tool. *International Journal of Early Years Education*, 1–16. https://doi.org/10.1080/09669760.2023.2269978

*Chatzaki, M. A., Skillen, J., Ricken, G., & Seitz-Stein, K. (2024, March). Exploring the potential of a game-based preschool assessment of mathematical competencies. In *Frontiers in Education* (Vol. 9, p. 1337716). Frontiers Media SA.

Cheng, W., Lei, P. W., & DiPerna, J. C. (2017). An examination of construct validity for the EARLI numeracy skill measures. *The Journal of Experimental Education, 85*(1), 54–72.

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency of number sense (ASPENS)*. Sopris Learning: Cambium Learning Group.

*Clarke, B., Strand Cary, M. G., Shanley, L., & Sutherland, M. (2020). Exploring the promise of a number line assessment to help identify students at-risk in mathematics. *Assessment for Effective Intervention*, *45*(2), 151-160.

*Clarke, B., Sutherland, M., Doabler, C. T., Lesner, T., Fainstein, D., Nolan, K., Landis, B., & Kosty, D. (2023). Developing and investigating the promise of early measurement screeners. *School Psychology Review*, *52*(6), 696–708.

*Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early maths assessment. *Educational Psychology*, *28*(4), 457-482.

*Clements, D. H., Sarama, J., Tatsuoka, C., Banse, H., & Tatsuoka, K. (2022). Evaluating a model for developing cognitively diagnostic adaptive assessments: The case of young children's length measurement. *Journal of Research in Childhood Education*, *36*(1), 143-158.

Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math*. Routledge.

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *T.E.A.M.-Tools for early assessment in mathematics*. SRA/McGraw-Hill.

Connolly, A. (1988). *KeyMath-revised: A diagnostic inventory of essential mathematics examiner manual*. American Guidance Service.

Cosso, J., Purpura, D. J., & Yoshikawa, H. (2024). The home numeracy environment of Latine families: A mixed methods measurement development study. *Journal of Educational Psychology, 11*(6), 853–870.

Costa, H. M., Nicholson, B., Donlan, C., & Van Herwegen, J. (2018). Low performance on mathematical tasks in preschoolers: The importance of domain-general and domain-specific abilities. *Journal of Intellectual Disability Research, 62*(4), 292–302.

Crawford, C., & Cribb, J. (2013). *Reading and maths skills at age 10 and earnings in later life: a brief analysis using the British Cohort Study*. (CAYT Impact Study REP03). Institute for Fiscal Studies and CAYT.

*Cruz, J., Alves, D., Carvalho, M., Mendes, S. A., Rodrigues, B., & Cadime, I. (2024). Assessment of math abilities before school entry: a tool development. In *Frontiers in Education* (Vol. 8, p. 1347143). Frontiers Media SA.

*Csapó, B., Molnár, G. & Nagy, J. (2014).Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology, 106*(3), 639–650.

*David, C., Dobrean, A., & Hans Van Luit, J. E. (2015). Psychometric properties of early numeracy test in Romanian language preliminary data. *Transylvanian Journal of Psychology/Erdélyi Pszichológiai Szemle*, *16*(1), 57.

Davis-Kean, P. E., Domina, T., Kuhfeld, M., Ellis, A., & Gershoff, E. T. (2022). It matters how you start: Early numeracy mastery predicts high school math course-taking and college attendance. *Infant and Child Development, 31*(2), e2281.

*de León, S. C., Jiménez, J. E., García, E., Gutiérrez, N., & Gil, V. (2021). Universal screening in mathematics for Spanish students in first grade. *Learning Disability Quarterly*, *44*(2), 123-135.

*de León, S. C., Jiménez, J. E., & Hernández-Cabrera, J. A. (2022). Confirmatory factor analysis of the indicators of basic early math skills. *Current Psychology*, *41*, 585-596.

De Smedt, B. (2022). Individual differences in mathematical cognition: A Bert's eye view. *Current Opinion in Behavioral Sciences, 46*, 101175.

De Vos, T. (1992). *Tempo test rekenen (TTR)*. Berkhout.

Devlin, D., Moeller, K., & Sella, F. (2022). The structure of early numeracy: Evidence from multi-factorial models. *Trends in Neuroscience and Education, 26*, 100171.

DiPerna, J. C., Morgan, P. L., & Lei, P. (2007). Development of early arithmetic, reading, and learning indicators for head start (EARLI Project). *Semi-annual performance report to the U.S. Department of Health and Human Services Administration for Children and Families.* University Park: Pennsylvania State University, College of Education.

Dockrell, J., Hurry, J., Cowan, R., Flouri, E., & Dawson, A. (2017). *Review of assessment measures in the early years: Language and literacy, numeracy and social emotional development and mental health*. Education Endowment Foundation.

*Dong, Y., Clements, D. H., Day-Hess, C. A., Sarama, J., & Dumas, D. (2021). Measuring early childhood mathematical cognition: Validating and equating two forms of the Research-based Early Mathematics Assessment. *Journal of Psychoeducational Assessment*, *39*(8), 983-998.

*Dong, Y., Dumas, D., Clements, D. H., Day-Hess, C. A., & Sarama, J. (2023).Evaluating the consequential validity of the research-based early mathematics assessment. *Journal of Psychoeducational Assessment*, *41*(5), 575-582.

*Dos Santos, F. H., Da Silva, P. A., Ribeiro, F. S., Dias, A. L. R. P., Frigerio, M. C., Dellatolas, G., & von Aster, M. (2012). Number processing and calculation in brazilian children aged 7-12 years. *The Spanish Journal of Psychology*, *15*(2), 513-525.

Draper, C. E., Barnett, L. M., Cook, C. J., Cuartas, J. A., Howard, S. J., McCoy, D. C., Merkley, R., Molano, A., Maldonado-Carreño, C., Obradović, J., Scerif, G., Valentini, N. C., Venetsanou, F., & Yousafzai, A. K. (2022). Publishing child development research from around the world: An unfair playing field resulting in most of the world's child population underrepresented in research. *Infant and Child Development, 32*(6), e2375.

Dunn, L. M., & Dunn, L. M. (1997). Peabody picture vocabulary test (3rd ed.). American Guidance Service.

Duraiappah, A.K., Atteveldt, N.M., Buil, J.M., Singh, K. and Wu, R., 2022. *Summary for decision makers,* reimagining *education: The international science and evidence based education assessment.* UNESCO MGIEP.

Elliott, C. D., Salerno, J. D., Dumont, R., & Willis, J. O. (2007). *Differential ability scales Second edition*. Harcourt Assessment.

*Erford, B. T., Bagley, D. L., Hopper, J. A., Lee, R. M., Panagopulos, K. A., & Preller, D. B. (1998). Reliability and validity of the Math Essential Skill Screener—Elementary Version (MESS-E). *Psychology in the Schools*, *35*(2), 127-135.

*Eteng-Uket, S. (2023). The Development, validation, and standardization of a new tool: The Dyscalculia Test. *Numeracy*, *16*(2), 1.

Fidan, E. (2013). *İlkokul Öğrencileri İçin Matematik Dersi Sayılar Öğrenme Alanında Başarı Testi Geliştirilmesi.* (Yayımlanmamış Yüksek Lisans Tezi), Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü.

*Floyd, R. G., Hojnoski, R., & Key, J. (2006). Preliminary evidence of the technical adequacy of the preschool numeracy indicators. *School Psychology Review*, *35*(4), 627-644.

*Fritz, A., Balzer, L., Ehlert, A., Herholdt, R., & Ragpot, L. (2014). A mathematics competence test for Grade 1 children migrates from Germany to South Africa. *South African Journal of Childhood Education*, *4*(2), 114-133.

Fuchs, L. S., Hamlett, C. L., & Powell, S. R. (2003). Grade 3 Math Battery (pp. 37203). Nashville, TN: Department of Special Education. Available from L. S. Fuchs, 228 Peabody, Vanderbilt University.

*Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. *Journal of Psychoeducational Assessment*, *27*(3), 265-279.

Gilmore, C. (2023). Understanding the complexities of mathematical cognition: A multi-level framework. *Quarterly Journal of Experimental Psychology, 76*(9), 1953–1972.

Ginsburg, H., & Baroody, A. J. (2003). *TEMA-3: Test of early mathematics ability*. Pro-ed.

Ginsburg, H., & Baroody, A. (2007). Test of early mathematics ability [Spanish language version, adapted by MC Núñez y I. Lozano]. *TEA Ediciones: Madrid, Spain*.

*Ginsburg, H. P., Lee, Y. S., & Pappas, S. (2016). A research-inspired and computer-guided clinical interview for mathematics assessment: Introduction, reliability and validity. *ZDM Mathematics Education*, *48*, 1003-1018.

Ginsburg, H. P., & Pappas, S. (2016). Invitation to the birthday party: Rationale and description. *ZDM, 48*, 947–960.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using dynamic indicators of basic early literacy skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 699–720). National Association of School Psychologists.

Hakkarainen, A., Cordier, R., Parsons, L., Yoon, S., Laine, A., Aunio, P., & Speyer, R. (2023). A systematic review of functional numeracy measures for 9–12-year-olds: Validity and reliability evidence. *International Journal of Educational Research, 119*, 102172.

*Halpin, P. F., Wolf, S., Yoshikawa, H., Rojas, N., Kabay, S., Pisani, L., & Dowd, A. J. (2019). Measuring early learning and development across cultures: Invariance of the IDELA across five countries. *Developmental Psychology*, *55*(1), 23.

*Hassler Hallstedt, M., & Ghaderi, A. (2018).Tablets instead of paper-based tests for young children? Comparability between paper and tablet versions of the mathematical Heidelberger Rechen Test 1-4. *Educational Assessment*, *23*(3), 195-210.

*Hawes, Z., Nosworthy, N., Archibald, L., & Ansari, D. (2019). Kindergarten children's symbolic number comparison skills relates to 1st grade mathematics achievement: Evidence from a two-minute paper-and-pencil test. *Learning and Instruction*, *59*, 21-33.

*Hellstrand, H., Korhonen, J., Räsänen, P., Linnanmäki, K., & Aunio, P. (2020). Reliability and validity evidence of the early numeracy test for identifying children at risk for mathematical learning difficulties. *International Journal of Educational Research*, *102*, 101580.

*Henning, E., Balzer, L., Ehlert, A., & Fritz, A. (2021). Development of an instrument to assess early number concept development in four South African languages. *South African Journal of Education*, *41*(4), 1-12.

Heroman, C., Burts, D. C., Berke, K.-L., & Bickart, T. S. (2010). *Teaching strategies GOLD® objectives for development & learning*. Teaching Strategies LLC.

Hornburg, C. B., Borriello, G. A., Kung, M., Lin, J., Litkowski, E., Cosso, J., ... & Purpura, D. J. (2021). Next directions in measurement of the home mathematics environment: An international and interdisciplinary perspective. *Journal of Numerical Cognition*, *7*(2), 195.

*Howard, S. J., Neilsen-Hewett, C., de Rosnay, M., Melhuish, E. C., & Buckley-Walker, K. (2022). Validity, reliability and viability of pre-school educators' use of early years toolbox early numeracy. *Australasian Journal of Early Childhood*, *47*(2), 92-106.

Hresko, W. P., Peak, P. K., Herron, S. R., & Hicks, D. L. (2000). *Young children's achievement test: YCAT-2*. Pro-ed.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Huang, Q., Sun, J., Lau, E. Y. H., & Zhou, Y. L. (2022). Linking Chinese mothers' and fathers' scaffolding with children's initiative and mathematics performance: A moderated mediation model. *Early Childhood Research Quarterly, 59*, 74–83.

Janssen, J., Scheltens, F., & Kraemer, J. M. (2005a). *Leerling- en onderwijsvolgsysteem rekenenwiskunde groep 3* [Student and education monitoring system mathematics grade 1]. Cito.

Janssen, J., Scheltens, F., & Kraemer, J. M. (2005b). *Leerling- en onderwijsvolgsysteem rekenenwiskunde groep 4* [Student and education monitoring system mathematics grade 2]. Cito.

Janssen, J., Scheltens, F., & Kraemer, J. M. (2006). *Leerling- en onderwijsvolgsysteem rekenen-wiskunde groep 5* [Student and education monitoring system mathematics grade 3]. Cito.

Jastak, S., & Wilkinson, G. S. (1984). *The wide range achievement test-revised*. Jastak Associates.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582–600.

Jiménez, J. E., & de León, S. C. (2019).*Indicadores de progreso de aprendizaje en matemáticas (IPAM)-2º curso de educación primaria [Indicators of basic early math skills (IPAM)- 2nd grade of primary school].* In J. E. Jimenez (Ed.), Modelo de respuesta a la intervención. Un enfoque preventivo para el abordaje de las dificultades específicas de aprendizaje [Response to intervention model. A preventive approach for learning disabilities]. Madrid: Pirámide.

*Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010).Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review*, *39*(2), 181-195.

*Jordan, N. C., Glutting, J. J., & Dyson, N. (2012). *Number Sense Screener™ (NSS™) User's Guide, K-1, Research Edition.* Paul H. Brookes Publishing Co. Available from: https://brookespublishing.com/wp-content/uploads/2021/06/NSS-technical-report.pdf. Accessed May 2024.

*Józsa, K., Oo, T. Z., Borbélyová, D., & Zentai, G. (2023). Exploring the accuracy and consistency of a school readiness assessment tool for preschoolers: Reliability, validity and measurement invariance analysis. *Journal of Intelligence*, *11*(10), 189.

Kang, D., Zhou, X., Tian, L. L., Li, Z. Q., & Xu, J. J. (2014). On the applicability of test of early child mathematics ability (Chinese edition) among 5–6 year-old children in Shanghai. *Early Childhood Education (Educational Sciences)*, *6*(6), 39–45.

*Karagiannakis, G., & Noël, M. P. (2020). Mathematical profile test: A preliminary evaluation of an online assessment for mathematics skills of children in grades 1–6. *Behavioral Sciences*, *10*(8), 126.

*Ketterlin-Geller, L. R., Perry, L., Platas, L. M., & Sitbakhan, Y. (2018). Aligning test scoring procedures with test uses of the early grade mathematics assessment: A balancing act. *Global Education Review, 5*(3), 143-164.

*Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, *30*(2), 148-159.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (RCBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52*(4), 377–405.

*Kiziltepe, G. I. (2019). Validity and reliability study for the Turkish version of number sense screener for 60-71 months-old children. *Journal of Education and Training Studies*, *7*(2), 24-35.

Klingbeil, D. A., Maurice, S. A., Van Norman, E. R., Nelson, P. M., Birr, C., Hanrahan, A. R., ... & Lopez, A. L. (2019). Improving mathematics screening in middle school. *School Psychology Review*, *48*(4), 383–398.

Koerhuis, I. (2010). *Rekenen voor kleuters [Mathematics for kindergarten].* Cito.

*Koponen, T., Salminen, J., Aunio, P., Polet, J., & Hellstrand, H. (2011). *LukiMat - Bedömning av* läran-det*: Identifiering av stödbehov i matematik i förskola. Handbok* [LukiMat – Assessment for learning: Identifying children in need of support in mathematics in kindergarten. Handbook].

*Koumoula, A., Tsironi, V., Stamouli, V., Bardani, I., Siapati, S., Graham, A., Kafantaris, I., Charalambidou, I., Dellatolas, G & Von Aster, M. (2004). An epidemiological study of number processing and mental calculation in Greek schoolchildren. *Journal of Learning Disabilities*, *37*(5), 377-388.

*Kowalski, K., Brown, R. D., Pretti-Frontczak, K., Uchida, C., & Sacks, D. F. (2018). The accuracy of teachers' judgments for assessing young children's emerging literacy and math skills. *Psychology in the Schools*, *55*(9), 997-1012.

Lafay, A., Osana, H. P., Michaud, S., & Nosworthy, N. (2020). Dépistage des difficultés mathématiques: Validation et normalisation franco-québécoise du Tempo Test Rekenen et du Numeracy Screener version française. *Glossa, 127*, 32–57.

*Lambert, R. G., Kim, D. H., & Burts, D. C. (2014). Using teacher ratings to track the growth and development of young children using the Teaching Strategies GOLD® assessment system. *Journal of Psychoeducational Assessment*, *32*(1), 27-39.

*Lambert, R. G., Kim, D. H., & Burts, D. C. (2015). The measurement properties of the Teaching Strategies GOLD® assessment system. *Early Childhood Research Quarterly*, *33*, 49-63.

*Lee, Y. S., & Lembke, E. (2016). Developing and evaluating a kindergarten to third grade CBM mathematics assessment. *ZDM Mathematics Education*, *48*, 1019-1030.

Lee, Y. S., Pappas, S., Chiong, C., & Ginsburg, H. P. (2010). *mCLASS®: MATH—technical Manual*. Wireless Generation Inc.

*Lee, E. K., Jung, J., Kang, S. H., Park, E. H., Choi, I., Park, S., & Yoo, H. K. (2017). Development of the computerized mathematics test in Korean children and adolescents. *Journal of the Korean Academy of Child and Adolescent Psychiatry*, *28*(3), 174-182.

*Lee, Y. S. (2016). Psychometric analyses of the Birthday Party. *ZDM Mathematics Education*, *48*, 961-975.

*Lei, P. W., Wu, Q., DiPerna, J. C., & Morgan, P. L. (2009). Developing short forms of the EARLI numeracy measures: Comparison of item selection methods. *Educational and Psychological Measurement*, *69*(5), 825-842.

Lewis, K. E., & Fisher, M. B. (2016). Taking stock of 40 years of research on mathematical learning disability: Methodological issues and future directions. *Journal for Research in Mathematics Education, 47*(4), 338–371.

*Li, L., Zhou, X., Huang, J., Tu, D., Gao, X., Yang, Z., & Li, M. (2020). Assessing kindergarteners' mathematics problem solving: The development of a cognitive diagnostic test. *Studies in Educational Evaluation*, *66*, 100879.

*Lin, J., Napoli, A. R., Schmitt, S. A., & Purpura, D. J. (2021).The relation between parent ratings and direct assessments of preschoolers' numeracy skills. *Learning and Instruction*, *71*, 101375.

Linacre, J. M. (2017). Teaching Rasch measurement. *Rasch Measurement Transactions, 31*(2), 1630–1631.

Lonigan, C. J., & Wilson, S. B. (2008). *Report on the revised Get Ready to Read! screening tool: Psychometrics and normative information* [Technical report]. National Center for Learning Disabilities.

*Lopez-Pedersen, A., Mononen, R., Korhonen, J., Aunio, P., & Melby-Lervåg, M. (2021). Validation of an early numeracy screener for first graders. *Scandinavian Journal of Educational Research*, *65*(3), 404-424.

*Malofeeva, E., Day, J., Saco, X., Young, L., & Ciancio, D. (2004). Construction and evaluation of a number sense test with head start children. *Journal of Educational Psychology*, *96*(4), 648-659.

Martin, N. A., & Brownell, R. (2011). *Expressive oneword picture vocabulary test manual* (4th ed.). Academic Therapy Publications.

Mazzocco, M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82*(4), 1224–1237.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Technical manual, Woodcock–Johnson III normative update*. Riverside.

*Mejias, S., Muller, C., & Schiltz, C. (2019). Assessing mathematical school readiness. *Frontiers in Psychology*, *10*, 1173.

*Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review*, *37*(3), 359-373.

Milburn, T. F., Lonigan, C. J., DeFlorio, L., & Klein, A. (2019). Dimensionality of preschoolers' informal mathematical abilities. *Early Childhood Research Quarterly, 47*, 487–495.

Mokkink, L. B., Prinsen, C. A., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Brazilian Journal of Physical Therapy, 20*, 105–113.

Morsanyi, K., van Bers, B. M., McCormack, T., & McGourty, J. (2018). The prevalence of specific learning disorder in mathematics and comorbidity with other developmental disorders in primary school-age children. *British Journal of Psychology, 109*(4), 917–940.

*Moura, R., Lopes-Silva, J. B., Vieira, L. R., Paiva, G. M., Prado, A. C. D. A., Wood, G., & Haase, V. G. (2015). From "five" to 5 for 5 minutes: Arabic number transcoding as a short, specific, and sensitive screening tool for mathematics learning difficulties. *Archives of Clinical Neuropsychology*, *30*(1), 88-98.

Muñez, D., Bull, R., Lee, K., & Ruiz, C. (2023). Heterogeneity in children at risk of math learning difficulties. *Child Development, 94*(4), 1033–1048.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., Mather, N., & Ardoino, G. (2009). Batería III Woodcock-Muñoz. *Ciencias Psicológicas, 3*(2), 245–246.

NCII. (2019). *Academic progress monitoring tools chart rating rubric.* National Centre on Intensive Intervention.

Nelson, G., Kiss, A. J., Codding, R. S., McKevett, N. M., Schmitt, J. F., Park, S., ... & Hwang, J. (2023). Review of curriculum-based measurement in mathematics: An update and extension of the literature. *Journal of School Psychology*, *97*, 1–42.

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities, 51*(6), 523–539.

Newcomer, P. L. (2001). *DAB-3: Diagnostic Achievement Battery*. Pro-Ed.

Nogues, C. P., & Dorneles, B. V. (2021). Systematic review on the precursors of initial mathematical performance. *International Journal of Educational Research Open, 2*, 100035.

*Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A two-minute paper-and-pencil test of symbolic and nonsymbolic numerical magnitude processing explains variability in primary school children's arithmetic competence. *PloS One*, *8*(7), e67918.

*Nunes, T., Bryant, P., Evans, D., & Barros, R. (2015).Assessing quantitative reasoning in young children. *Mathematical Thinking and Learning*, *17*(2-3), 178-196.

Olkun, S., Can, D., & Yeşilpınar, M. (2013). *Hesaplama Performansı Testi: Geçerlilik Ve Güvenilirlik Çalışması*. Paper presented at the USOS 2013 Ulusal Sınıf Öğretmenliği Sempozyumu.

*Olkun, S., Altun, A., Şahin, S. G., & Kaya, G. (2016). Psychometric properties of a screening tool for elementary school student's math learning disorder risk. *International Journal of Learning, Teaching and Educational Research*, *15*(12), 48-66.

Outhwaite, L., Early, E., Herodotou, C., & Van Herwegen, J. (2022). Can Maths apps add value to young children's learning? A systematic review and content analysis. *Nuffield Foundation.*

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Bmj*, *372*, n71.

Paik, J. H., van Gelderen, L., Gonzales, M., de Jong, P. F., & Hayes, M. (2011). Cultural differences in early math skills among US, Taiwanese, Dutch, and Peruvian preschoolers. *International Journal of Early Years Education, 19*(2), 133–143.

Panter, J. E., & Bracken, B. A. (2009). Validity of the Bracken School Readiness Assessment for predicting first grade readiness. *Psychology in the Schools, 46*(5), 397–409.

*Perry, L. (2020). Development of an early grade relational reasoning subtask: Collecting validity evidence on technical adequacy and reliability. *International Journal of Science and Mathematics Education*, *18*(3), 589-609.

*Pisani, L., Borisova, I., & Dowd, A. J. (2018). Developing and validating the international development and early learning assessment (IDELA). *International Journal of Educational Research*, *91*, 1-15.

*Pisani, L., Seiden, J., & Wolf, S. (2022). Longitudinal evidence on the predictive validity of the International Development and Early Learning Assessment (IDELA). *Educational Assessment, Evaluation and Accountability*, *34*(2), 173-194.

Pitchford, N. J., & Outhwaite, L. A. (2016). Can touch screen tablets be used to assess cognitive and motor skills in early years primary school children? A cross-cultural study. *Frontiers in Psychology, 7*, 217270.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*, 1147–1157.

*Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015).Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review*, *44*(1), 41–59.

Purpura, D. J., & Lonigan, C. J. (2015). Early numeracy assessment: The development of the preschool early numeracy scales. *Early Education and Development, 26*(2), 286–313.

*Pushparatnam, A., Luna Bazaldua, D. A., Holla, A., Azevedo, J. P., Clarke, M., & Devercelli, A. (2021).Measuring early childhood development among 4–6 year olds: The identification of psychometrically robust items across diverse contexts. *Frontiers in Public Health*, *9*, 1-11.

*Ralston, N. C., Li, M., & Taylor, C. (2018). The development and initial validation of an assessment of algebraic thinking for students in the elementary grades. *Educational Assessment*, *23*(3), 211-227.

Ramani, G. B., Siegler, R. S., & Hitti, A. (2012). Taking it to the classroom: Number board games as a small group learning activity. *Journal of Educational Psychology, 104*(3), 661.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943.

*Rhodes, K. T., Branum-Martin, L., Morris, R. D., Romski, M., & Sevcik, R. A. (2015). Testing math or testing language? The construct validity of the KeyMath-Revised for children with intellectual disability and language difficulties. *American Journal on Intellectual and Developmental Disabilities*, *120*(6), 542-568.

Ricken, G., Fritz, A., & Balzer, L. (2013). *MARKO-D – Mathematics und Rechnen – Test zur Erfassung von Konzepten im Vorschulalter [MARKO-D: Mathematics and arithmetic – Test for assessing concepts in pre-school age]*. Hogrefe.

RTI International. (2014). *Early Grade Mathematics Assessment (EGMA) Toolkit.* RTI International.

*Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly*, *48*, 14-25.

Ryan, J. (2023). *Introducing the Self-Administered EGRA and EGMA (SA-EGRA/SA-EGMA).* Available from: https://shared.rti.org/content/introducing-self-administered-egra-and-egma-sa-egra-sa-egma-0. Accessed May 2024.

SASC. (2019). *SASC Guidance on assessment of dyscalculia and maths difficulties within other specific learning difficulties*. SASC: SpLD Assessment Standards Committee.

*Save the Children. (2019). IDELA – The International Development and Early Learning Assessment. Adaptation and Administration Guide. Available from: https://resourcecentre.savethechildren.net/document/idela-the-international-development-and-early-learning-assessment/. Accessed May 2024.

Scerif, G., Gattas, S., Godfrey, A., Hawes, Z., Howard, S., Merkley, R., O'Connor, R., & Sučević, J. (2023). *Orchestrating numeracy and the executive - "The ONE" Programme.* Nuffield Foundation. Available from: https://www.nuffieldfoundation.org/wp-content/uploads/2021/01/Scerif-Final-report-Fostering-resilience-by-injecting-executive-challenge-into-early-maths.pdf. Accessed May 2024.

Schrank, F.A., McGrew, K.S., & Woodcock, R.W. (2001). *Woodcock-Johnson III assessment service bulletin number 2 technical abstract.* Riverside Publishing Company.

Schrank, F. A., McGrew, K. S., Mather, N., Wendling, B. J., & LaForte, E. M. (2014). *Woodcock-Johnson IV tests of achievement: Form C*. Riverside Publishing Company.

Sella, F., Onnivello, S., Lunardon, M., Lanfranchi, S., & Zorzi, M. (2021). Training basic numerical skills in children with Down syndrome using the computerized game "The Number Race." *Scientific Reports, 11*(1), 2087.

*Shavitt, I., de Araujo Scatollin, M. A., Rossi, A. S. U., Mercadante, M. P., Gamez, L., Resegue, R. M., ... & do Rosário, M. C. (2022). Transcultural adaptation and psychometric properties of the International Development and Early Learning Assessment (IDELA) in Brazilian pre-school children. *International Journal of Educational Research Open*, *3*, 100138.

Simms, V., McKeaveney, C., Sloan, S., & Gilmore, C. (2019). *Interventions to improve mathematical achievement in primary school-aged children*. Nuffield Foundation.

*Singer, V., & Cuadro, A. (2014). Psychometric properties of an experimental test for the assessment of basic arithmetic calculation efficiency/Propiedades psicométricas de una prueba experimental para la evaluación de la eficacia del cálculo aritmético básico. *Estudios de Psicología*, *35*(1), 183-192.

*Sjoe, N. M., Bleses, D., Dybdal, L., Tideman, E., Kirkeby, H., Sehested, K. K., Nielsen, H., Kreiner, S. & Jensen, P. (2019). Short Danish version of the Tools for Early Assessment in Math (TEAM) for 3–6-year-olds. *Early Education and Development*, *30*(2), 238-258.

Snelling, M., Dawes, A., Biersteker, L., Girdwood, E., & Tredoux, C. (2019). The development of a South African Early Learning Outcomes Measure: A South African instrument for measuring early learning program outcomes. *Child Care, Health and Development, 45*(2), 257–270.

*Sutherland, M., Clarke, B., Nese, J. F., Cary, M. S., Shanley, L., Furjanic, D., & Durán, L. (2021). Investigating the utility of a kindergarten number line assessment compared to an early numeracy screening battery. *Early Childhood Research Quarterly*, *55*, 119-128.

Szűcs, D., & Goswami, U. (2013). Developmental dyscalculia: Fresh perspectives. *Trends in Neuroscience and Education, 2*(2), 33–37.

*ten Braak, D., & Størksen, I. (2021). Psychometric properties of the Ani Banani Math Test. *European Journal of Developmental Psychology*, *18*(4), 610-628.

*Tsigilis, N., Krousorati, K., Gregoriadis, A., & Grammatikopoulos, V. (2023). Psychometric evaluation of the preschool early numeracy skills test–brief version within the item response theory framework. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12536

Turan, E., & De Smedt, B. (2022). Mathematical language and mathematical abilities in preschool: A systematic literature review. *Educational Research Review, 36*, 100457.

UNESCO. (2017). *More than one-half of children and adolescents are not learning worldwide*. UIS Fact Sheet No. 46.

UNESCO. (2023). *Early childhood care and education. An investment in wellbeing, gender equality, social cohesion, and lifelong learning.* Available from: https://www.unesco.org/en/early-childhood-education. Accessed May 2024.

*Van de Rijt, B., Godfrey, R., Aubrey, C., van Luit, J. E., Ghesquière, P., Torbeyns, J., Hasemann, K., Tancig, S., Kavkler, M., Magajna, L., & Tzouriadou, M. (2003). The development of early numeracy in Europe. *Journal of Early Childhood Research*, *1*(2), 155-180.

Van Herwegen, J., & Simms, V. (2020). Mathematical development in Williams syndrome: A systematic review. *Research in Developmental Disabilities, 100*, 103609.

Van Herwegen, J., Costa, H. M., Nicholson, B., & Donlan, C. (2018). Improving number abilities in low achieving preschoolers: Symbolic versus non-symbolic training programs. *Research in Developmental Disabilities, 77*, 1–11.

*Van Hoogmoed, A. H., Van den Ham, A., Jordan, A., Duchhardt, C., Kroesbergen, E. H., & Heinze, A. (2022). Exploring the reliability, validity, and dimensionality of the 'Kieler kindergarten test for mathematics'. *Pedagogische Studiën*, *99*(4), 304-324.

Van Luit, J. E. H., & Van de Rijt, B. A. M. (2009). *Utrechtse getalbegrip toets-Revised [Early numeracy test-Revised]*. Graviant.

Van Luit, J. E. H., Van de Rijt, B. A. M. & Pennings, A. H. (1994). *Utrechtse Getalbegrip Toets [Utrecht Test of Number Sense].* Graviant.

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2014). Arithmetic strategy development and its domain-specific and domain-general cognitive correlates: A longitudinal study in children with persistent mathematical learning difficulties. *Research in Developmental Disabilities, 35*(11), 3001–3013.

*Vessonen, T., Widlund, A., Hakkarainen, A., & Aunio, P. (2023). Validating the early numeracy teacher rating scale for preschoolers (TRS–EN). *European Early Childhood Education Research Journal*, *31*(2), 205-224.

*Viapiana, V. F., Mendonça Filho, E. J. D., Fonseca, R. P., Giacomoni, C. H., & Stein, L. M. (2016). Development of the arithmetic subtest of the school achievement test-Second Edition. *Psicologia: Reflexão e Crítica*, *29*(39), 1–10.

*Vitiello, V. E., & Williford, A. P. (2021). Alignment of teacher ratings and child direct assessments in preschool: A closer look at teaching strategies GOLD. *Early Childhood Research Quarterly*, *56*, 114-123.

von Aster, M. G., Weinhold Zulauf, M., & Horn, R. (2006). *Zareki-R Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern [Neuropsychological Test Battery for Number Processing and Calculation in Children*]. Harcourt Test Services.

Wechsler, D. (1967). *Manual WPPSI: Wechsler Pre-school and Primary Intelligence Scale*. Psychological Corp.

Wechsler, D. (1992). *Wechsler Individual Achievement Test*. Psychological Corporation.

Wechsler, D. (1997), *Wechsler Intelligence Scale for Children–Third edition* [Greek version]. Hellinika Grammata.

*Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology*, *32*(3), 311-333.

*Wolf, S., Halpin, P., Yoshikawa, H., Dowd, A. J., Pisani, L., & Borisova, I. (2017). Measuring school readiness globally: Assessing the construct validity and measurement invariance of the International Development and Early Learning Assessment (IDELA) in Ethiopia. *Early Childhood Research Quarterly*, *41*, 21-36.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson: Tests of achievement–revised*. DLM.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III NU complete*. Riverside Publishing.

Yao, S. Y., Muñez, D., Bull, R., Lee, K., Khng, K. H., & Poon, K. (2017). Rasch modeling of the test of early mathematics ability–third edition with a sample of K1 children in Singapore. *Journal of Psychoeducational Assessment, 35*(6), 615–627.

Yuste-Hernanz, C. (2002). *BADyG-E1:Batería de aptitudes diferencialesy generales [The battery of differential and general abilities]* (2nd ed). Ciencias de la Educación Preescolar y Especial, CEPE.

Authors and Affiliations

Laura A. Outhwaite[1] · Pirjo Aunio[2] · Jaimie Ka Yu Leung[3] · Jo Van Herwegen[1,3]

✉ Laura A. Outhwaite
l.outhwaite@ucl.ac.uk

[1] Centre for Education Policy and Equalising Opportunities, IOE, UCL's Faculty of Education and Society, London, UK

[2] Department of Education, Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland

[3] Department of Psychology and Human Development, IOE, UCL's Faculty of Education and Society, London, UK