UCL

CENTRE FOR EDUCATION POLICY & EQUALISING OPPORTUNITIES

Working Paper No. 24-03

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

Laura A. Outhwaite
University College London

Pirjo Aunio
University of Helsinki

Jaimie K. Y. Leung
University College London

Jo Van Herwegen
University College London

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, 37 mathematical assessments and 22 screeners were identified. In addressing the first aim, most measurement tools were categorised as child-direct measures delivered individually with a trained assessor in a paper-based format. In addressing the second aim, the synthesis revealed four key findings. First, the majority of the identified measurement tools have not been evaluated for all aspects of reliability and validity, and only seven measurement tools met the common acceptability thresholds for more than two areas of psychometric evidence. Second, only three screeners demonstrated an acceptable ability to distinguish between typically developing children and those with or at-risk of mathematical learning difficulties. Third, although five mathematical assessments and six screeners included evaluations of predictive validity, none met the common acceptability threshold. Finally, only eight mathematical assessments and one screener were found to align with external measurement tools. Building on this current evidence and improving measurement quality is vital for raising methodological standards in mathematical learning and development research.

VERSION: March 2024

Suggested citation: Outhwaite, L.A., Aunio, P., Leung, J.K.Y., & Van Herwegen, J. (2024). *Mathematical Skills in Early Childhood: A Systematic Review of the Psychometric Properties of Early Maths Assessments and Screeners* (CEPEO Working Paper No. 24-03). Centre for Education Policy and Equalising Opportunities, UCL.

Disclaimer

Any opinions expressed here are those of the author(s) and not those of the UCL Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

CEPEO Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Highlights

- This systematic review provides an overview of 37 mathematical assessments and 22 screeners that have been evaluated for their psychometric properties for measuring mathematical skills in children aged 0-8 years.
- The reliability and validity evidence for these measurement tools is synthesised, including in relation to common acceptability thresholds.
- Based on the current evidence, recommendations are made with regards to the mathematical assessments and screeners that have the most promising psychometric evidence.
- This study is relevant to researchers, practitioners, and other stakeholders who are 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 mathematical learning difficulties.

Why does this matter?

When assessing children's mathematical skills, it is important that the chosen measurement tools are reliable and valid.

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]

[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, Finland

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

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, 37 mathematical assessments and 22 screeners were identified. In addressing the first aim, most measurement tools were categorised as child-direct measures delivered individually with a trained assessor in a paper-based format. In addressing the second aim, the synthesis revealed four key findings. First, the majority of the identified measurement tools have not been evaluated for all aspects of reliability and validity, and only seven measurement tools met the common acceptability thresholds for more than two areas of psychometric evidence. Second, only three screeners demonstrated an acceptable ability to distinguish between typically developing children and those with or at-risk of mathematical learning difficulties. Third, although five mathematical assessments and six screeners included evaluations of predictive validity, none met the common acceptability threshold. Finally, only eight mathematical assessments and one screener were found to align with external measurement tools. Building on this current evidence and improving measurement quality is vital for raising methodological standards in mathematical learning and development research.

Key words: mathematics; early childhood; assessment; screener; measurement

Address correspondence to: Dr Laura A. Outhwaite, l.outhwaite@ucl.ac.uk

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

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- 14% 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 (Alcock et al., 2016). 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 (Simms et al., 2019; Outhwaite et al., 2022), and ii) identify children with or at-risk of MLD (Lewis & Fisher, 2016).

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 only.

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). There are several models that 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). These models of mathematical development also include arithmetic skills, such as addition and subtraction presented in both single and multi-digit operations, as well as word problems (e.g., Aunio & Räsänen, 2016; Clements & Sarama, 2009; Purpura & Lonigan, 2015). Some of these models (e.g., Clements and Sarama, 2009) describe all 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.

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).

Previous reviews have summarised some assessments of children's mathematical skills, but only 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 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, validated, and produce reliable indications of children's skills in early childhood.

Defining Mathematical Learning Difficulties

Different terminology is frequently used to refer to children who struggle to learn mathematics. Some children may perform low on mathematical tasks, typically because of poor learning environments and may be referred to as 'low achievers'. Whereas, some children have persistent difficulties, despite good quality teaching and age-appropriate development in other learning domains (Muñez et al., 2023). 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 (Szücs & 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 in relation to 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, whilst 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, validated, and 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, APA & NCME, 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.

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

Psychometric Evidence	Example Analysis Methods	Common Acceptability Thresholds
Content validity	Expert panels 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)

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 context of use. It is essential to consider if the measurement is relevant and comprehensible 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), discriminative (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 a base of criterion validity evaluation. 'Gold Standard' measurement tools typically have undergone extensive development and are widely accepted as the best measurement tools currently available. 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 the learning difficulties over time. To be able to produce predictive evidence, longitudinal data is 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., 2020).

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).

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. 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 will be 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 (osf.io/6te7g) with ethical approval granted by IOE ethics committee. 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 Figures 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 backwards citation of included studies (n = 57) was also conducted. This search strategy was completed in March 2021 (from January 1990- present) and was updated in June 2023 (from January 2021- present).

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 (i.e., assessment or screener) based on the way in which it was presented in the included psychometric studies. 20% 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 on 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 Figure 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 were screened by an additional reviewer (see acknowledgements) to ensure high levels of agreement ($\kappa = .84$). An updated search strategy was completed in June 2023 (see Figure 2) and identified an additional 10 eligible studies. This resulted in an overall total of 71 included studies in the current review. Consistent with the initial search, one

reviewer (third author) was responsible for screening all records at both levels. A random 20% sample of records were also screened by an additional reviewer (first author) to ensure high levels of agreement ($\kappa = .93$).

Figure 1 *PRISMA Flow Diagram of studies through the systematic review (original search, March 2021)*

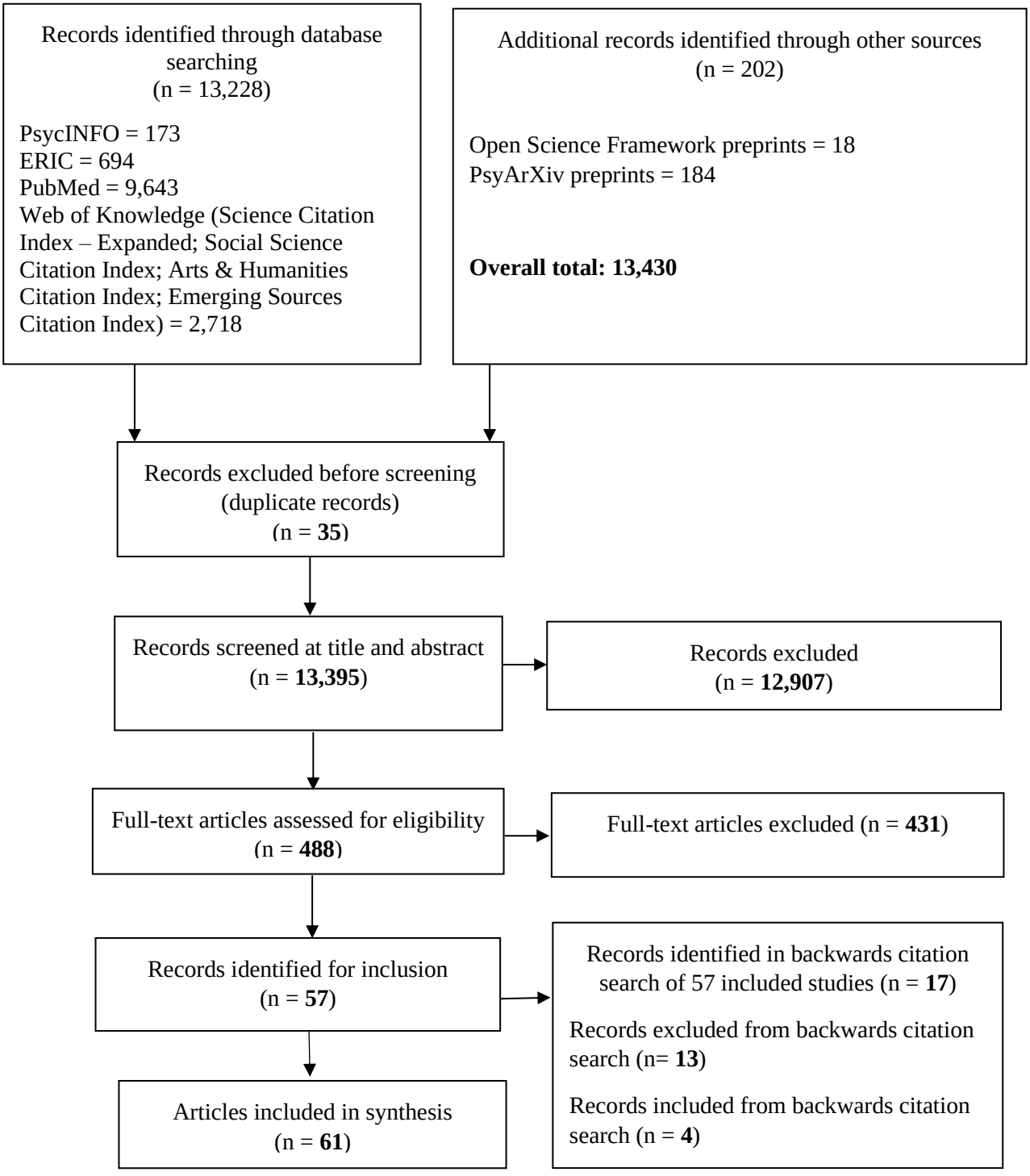
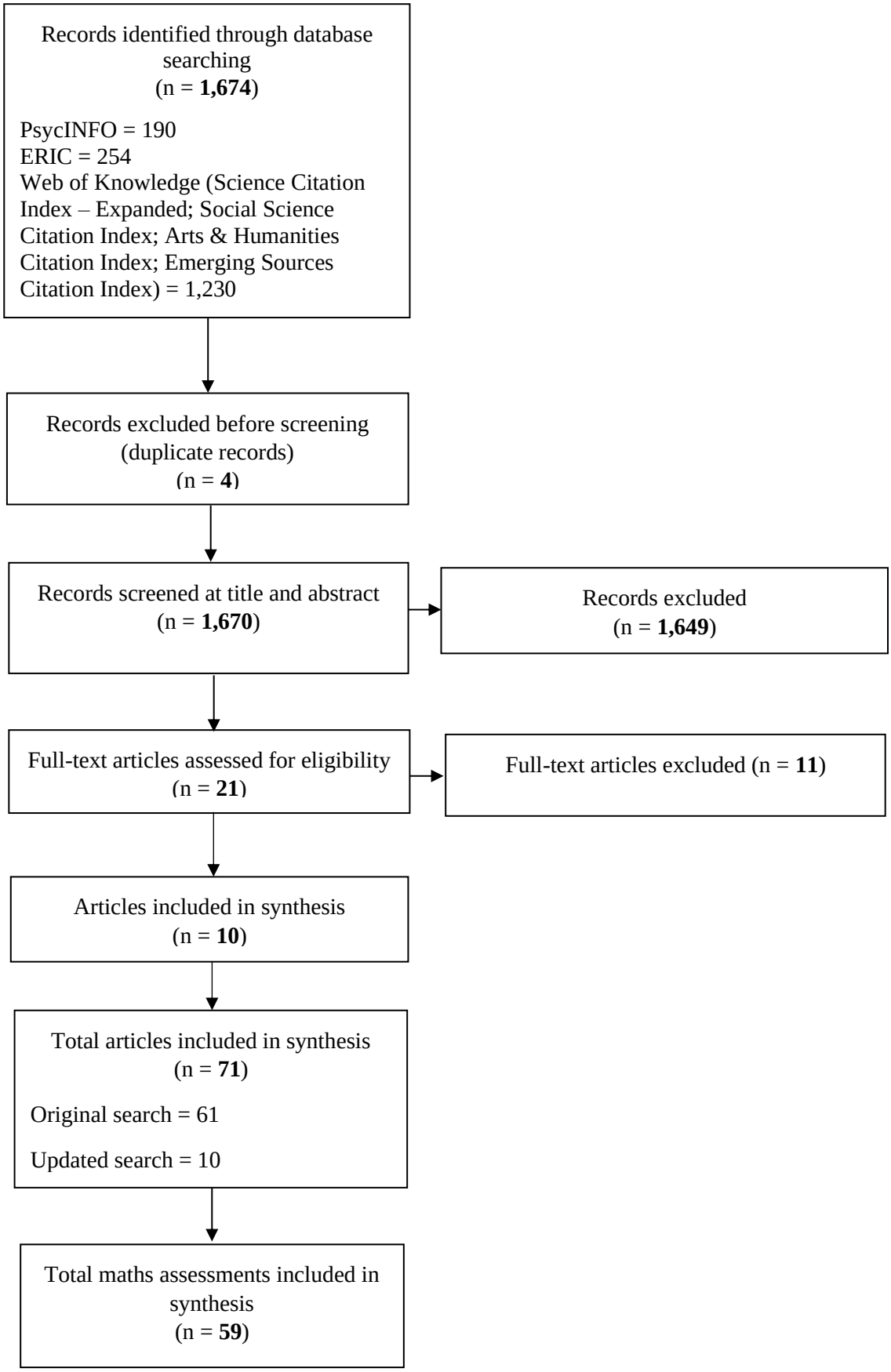


Figure 2 *PRISMA Flow Diagram of studies through the systematic review (updated search, June 2023)*



Coding Framework

To establish an overview of each of the measurement tools identified in the 71 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 (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 educational 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 (<.30 = Low; .3- .5 = Medium; >.5 = High/Acceptable) or were rated as 'Not applicable' (NA), if other forms of analysis were used.

Results

Overview of Measurement Tools

In total, 59 measurement tools were identified. This included 37 mathematical assessments designed for children aged 1-14 years and 22 screeners suitable for children aged 3-14 years. As summarised in Table 2, most measurement tools were child-direct measures (n = 52) administered individually (n = 52) with a trained assessor (n = 49) in a paper-based format (n = 41). Most measurement tools targeted number (n = 55) and/or arithmetic skills (n = 47), with less than half of the identified assessments and screeners measuring shape, space, and measure skills (n = 22).

Although the identified measurement tools were evaluated in 44 countries and 20 languages, over half of the assessments and screeners were developed in WEIRD societies and/or in English (n = 34). Only nine assessments and two 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 evaluation of the English and Spanish versions of the Birthday Party assessment (Lee, 2016), 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 38 assessments and 23 screeners.

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

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	-	-
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	-

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 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
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 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
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; Subitizing; Basic calculation.	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	-

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
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; Interpretating data; Problem solving.	Y	Y	Y
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; Subitizing; 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

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 & Dellatolas, 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	-
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

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
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	-
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 subitizing; Composing number; Geometry; Comparing shape; Identifying shape; Turns; Representing shape; Composing shape; Measuring; Patterning.	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	-	
Teaching Strategies GOLD (Lambert et al., 2014)	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

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
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 Utrecht Test of Early Numeracy (ENT) Test (Van Luit et al., 1994; Van de Rijt et al., 2003)	4-8	7 European countries (Various)[e]	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 (US- TEAM- short) (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 subitizing; Numerals; Composition of number; Arithmetic; Shape; Patterning shape; Compose shape.	Y	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

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
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	-
Basic Number Processing Test (BNPT) (Olkun et al., 2016)	6-9	Turkey (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	-
Early Numeracy (EN)- Test (Koponen et al., 2011)	5-8	Sweden; Finland (Swedish; Finnish)	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	-
Early Numeracy Screener (Lopez-Peterson et al., 2020)	6-7	Norway (Norwegian)	Child-direct; Paper-based	Group; Teacher	**52 items:** Numerical relational skills; Counting skills; Arithmetic skills.	Y	Y	-

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 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	-	-
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	-
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, 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	-	-

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
Number Sense Brief (NSB) Screener (Jordan et al., 2010)	5-6	USA (English)	Child-direct; Paper-based	Individual; Researcher/ trained assessor	**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 Sense Screener (NSS) (Akulun, 2019; Kiziltepe, 2018)	5-6	Turkey (Turkish)	Child-direct; Paper-based	Individual; Researcher/ trained assessor	**29 items:** Counting skills; Number recognition; Number comparisons; Nonverbal calculations; Story problems; 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	-	-
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	-	-

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 Diagnostic (PMA-D) (Brendefur et al., 2015)	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
Primary Math Assessment Screener (PMA-S) (Brendefur et al., 2015)	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	-	-

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

Content Validity

Content validity in the form of expert opinion on the suitability and adaptation of test items were only reported for six mathematical assessments (AAT, Ralston et al., 2018; 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 one screener (BNPT, Olkun et al., 2016). All were rated as acceptable.

Structural Validity

22 mathematical assessments included a measure of structural validity, of which Confirmatory Factor Analysis (CFA) was the most frequent approach (n = 10). However, only eight assessments met the common acceptability thresholds and were deemed to have good model fit (see Table 3). Eight screeners also included a measure of structural validity, of which CFA was also the most common method (n = 4) and four screeners met the acceptable threshold criteria (see Table 3).

Internal Consistency

Over half of the mathematical assessments reported internal consistency (n = 24) and most reached the acceptable threshold (n = 18) (see Table 3). However, of the 18 mathematical assessments with acceptable internal consistency, only one assessment reported disaggregated internal consistency results for the multiple dimensions identified in the structural validity evaluation (Birthday Party- Long Version- English, Lee, 2016).

Over half of the identified screeners also reported internal consistency (n = 12) with 10 meeting the acceptable thresholds. Within those that demonstrated acceptable internal consistency, only two screeners reported internal consistency for the different factors identified in the structural validity evaluation (EN- Test, Hellstrand et al., 2020; Early Numeracy Screener, Lopez-Peterson et al., 2020).

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
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

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
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)	NR	NR	NA	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
Early Grade Mathematics Assessment (EGMA) (Ketterlin-Geller et al., 2018; Perry, 2018)	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 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
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	*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

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
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
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

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
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
Teaching Strategies GOLD (Lambert et al., 2014; 2015)	CFA	6 factor model SRMR = .03, CFI = .93, RMSEA = .07	Low	Cronbach alpha	α = .94- .95	Acceptable
Test of Early Number and Arithmetic (TENA) (Bojorque et al., 2015)	NR	NR	NA	Cronbach alpha	α = .89- .91	Acceptable
The Utrecht Test of Early Numeracy (ENT) Test (Aunio et al., 2006)	NR	NR	NA	Cronbach alpha	α = .79- .90	Acceptable
Tools for Early Assessment in Math (US- TEAM-short) (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

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Tools for Early Assessment in Math Danish Version (DK-TEAM) (Sjoe et al., 2019)	NR	NR	NA	NR	NR	NA
Screeners						
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
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-Peterson et al., 2020)	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

Name of Measurement Tool (Related Papers)	Structural Validity			Internal Consistency		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Indicators of Basic Early Math Skills (IPAM) (de Leon 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
Mathematical School Readiness (MSR) (Mejias et al., 2019)	Unidimensionality	Evidence of unidimensionality for 3 tasks, r = .28 - .49	Mixed	Cronbach alpha	α = .63- .95	Mixed
Number Line Assessment 0-20, 0-100 (Clarke, 2020)	NR	NR	NA	Cronbach alpha	α = .83- .93	Acceptable
Number Line Assessment 0-100 (Sutherland et al., 2021)	NR	NR	NA	NR	NR	NA
Number Sense Brief (NSB) Screener (Jordan et al., 2010)	NR	NR	NA	Cronbach alpha	$\alpha > .80$	Acceptable
Number Sense Screener (NSS) (Aktulun; 2019; Kiziltepe, 2018)	Rasch model	Infit = .81- 1.35 Outfit = .60- 1.39	Acceptable Acceptable	Cronbach alpha	α = .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
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	α = .93 r = .90	Acceptable Acceptable

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) (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., 2015)	NR	NR	NA	Cronbach alpha	α = .82- .93	Acceptable
Primary Math Assessment Screener (PMA-S) (Brendefur et al., 2015)	NR	NR	NA	NR	NR	NA
Symbolic Magnitude Processing (SYMP) Test (Brankaer et al., 2017)	NR	NR	NA	NR	NR	NA

Notes: BIC = Bayesian Information Criterion; CAIC = Consistent Akaike's Information Criterion; CFA = Confirmatory Factor Analysis; CFI = Confirmatory Factor Index; IRT = 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

Reliability

12 mathematical assessments included indicators of test-retest reliability (controlled for age) with intervals ranging from 3-7 days to 2-6 months, and six were rated as acceptable. Seven assessments reported inter-rater reliability, of which six met the acceptable threshold (see Table 4). 10 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. But only three screeners were rated as having acceptable reliability using these methods (see Table 4).

Criterion Validity

Concurrent validity was evaluated with 20 mathematical assessments, with comparisons most frequently made with the Woodcock-Johnson Math subtests (n = 7). However, only eight mathematical assessments met acceptability thresholds (see Table 4). Divergent validity with standardised language and reading measurement tools was considered in four mathematical assessments, but only one was rated as acceptable. Predictive validity was also considered in five mathematical assessments, typically over 1–2-years. However, none were rated as acceptable on the common threshold criteria (see Table 4).

Concurrent validity was also evaluated with 10 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 were considered in two screeners, but only one was rated as acceptable. Predictive validity was considered in six screeners, over time periods ranging from 6 months to 3 years. However, all were rated as either mixed (n = 1) or low (n = 5) on the acceptability thresholds (see Table 4). Diagnostic accuracy was also considered in nine screeners. However, there were large variations in the reported sensitivity and specificity, with only three screeners meeting the acceptability thresholds (see Table 4).

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

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)	NR	NR	NA	NR	NR	NA
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
Early Grade Mathematics Assessment (EGMA) (Ketterlin-Geller et al., 2018; Perry, 2018)	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 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

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-Maths)	*r* = .50- .61	Mixed
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

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
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
Quantitative Reasoning Test (Nunes et al., 2015)	Test-retest (4.5 months)	r = .78	Acceptable	Predictive (Mathematical Reasoning Test; interval NR)	ΔR^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
School Achievement Test- 2nd Edition- Arithmetic Subtest (Viapiana et al., 2016)	NR	NR	NA	NR	NR	NA
Teaching Strategies GOLD (Lambert et al., 2014; 2015; Vitiello & Williford, 2021)	Inter-rater	k = .92	Acceptable	Concurrent (BBCS-R; WJ-III) Predictive (WJ-III; 1 term)	r = .27- .74 r = .39	Mixed Low
Test of Early Number and Arithmetic (TENA) (Bojorque et al., 2015)	Inter-rater	k = .92	Acceptable	Concurrent (ENT)	r = .80- .89	Acceptable
The Utrecht Test of Early Numeracy (ENT) Test (Aunio et al., 2006)	NR	NR	NA	NR	NR	NA

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Tools for Early Assessment in Math (US-TEAM- short) (Weiland et al., 2012)	NR	NR	NA	Concurrent (REMA; WJ-III-AP) Divergent (PPVT-III; WJ-III-LW)	*r* = 71- .74 *r* = .64	Acceptable Acceptable
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
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
Early Numeracy (EN)- Test (Hellstrand et al., 2020)	NR	NR	NA	NR	NR	NA

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Early Numeracy Screener (Lopez-Peterson et al., 2020)	NR	NR	NA	Predictive (Norwegian national test scores; 6 months)	*r* = .19- .25	Low
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
Indicators of Basic Early Math Skills (IPAM) (de Leon 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, 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 Brief (NSB) Screener (Jordan et al., 2010)	Test retest (17 months)	*r* = .61- 86	Mixed	Diagnostic accuracy	Sensitivity = .70- .86 Specificity = .35- .85	Mixed Mixed

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	Results	Rating	Method(s)	Results	Rating
Number Sense Screener (NSS) (Aktulun; 2019; Kiziltepe, 2018)	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	Acceptable Acceptable
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
Primary Math Assessment Diagnostic (PMA-D) (Brendefur et al., 2015)	NR	NR	NA	NR	NR	NA
Primary Math Assessment Screener (PMA-S) (Brendefur et al., 2015)	NR	NR	NA	NR	NR	NA

Name of Measurement Tool (Related Papers)	Reliability			Criterion Validity		
	Method(s)	**Results**	**Rating**	**Method(s)**	**Results**	**Rating**
Symbolic Magnitude Processing (SYMP) Test (Brankaer et al., 2017)	Test retest (interval NR)	r = .62- .77	Mixed	Diagnostic accuracy	TD children (≥35th percentile on standardized maths test) consistently outperformed MLD children (≤ the 10th percentile), except 10-11 years	Acceptable
				Concurrent (Standardized maths test)	r = .16- .40	Low

Notes: 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; 2005b; 2006); ***CPT*** = Calculation Performance Test (Olkun et al., 2013); ***DAS*** = Differential Ability Scales, Early Number Concepts Scale (Elliot, 2007); ***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); ***KRT-R*** = Kortrijkse Rekentest- Revisie (Baudonck et al., 2006); ***MAT*** = Math Achievement Test (Fidan, 2013); ***M-TEAM*** = 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., 2016); ***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); ***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).

Measurement Tools with Promising Evidence

Table 5 summarises the four mathematical assessments and three screeners with the most promising evidence identified within the current review.

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 Learning Outcomes Measure (ELOM)	-	✓	✓	✓	✓	4
Early Years Toolbox-Early Numeracy	-	-	-	✓	✓	2
Parent ratings of numeracy skills	-	✓	✓	-	-	2
Tools for Early Assessment in Math (US- TEAM- short)	-	✓	✓	-	✓	3
Screeners						
Assessing Student Proficiency of Early Number Sense (ASPENS)	-	-	-	✓	✓	2
Math Essential Skill Screener- Elementary Version (MESS-E)	-	-	✓	✓	✓	3
Preschool Early Numeracy Skills Screener- Brief Version (PENS-B) (English Version)	-	✓	✓	-	✓	3

Notes: Co.V = Content Validity; Cr.V = Criterion Validity; IC = Internal Consistency; R = Reliability; SV = Structural Validity

Discussion

This study reports the first pre-registered systematic review of the psychometric properties of mathematical assessments and screeners in early childhood. Specifically, this review first 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. Second, this review aimed to synthesise the reliability and validity of these measurement tools, including in relation to common acceptability thresholds. 71 individual studies relating to 59 measurement tools were identified. Of these measurement tools, 37 were mathematical assessments, and 22 were screeners. The psychometric properties of these measurement tools were then synthesised and appraised in line with five indicators of reliability and validity (content validity, structural validity, internal consistency, reliability, and criterion validity) from the Standards for Educational and Psychological measurements (AERA, APA & NCME, 2014) and COSMIN guidelines (Mokkink et al., 2016; Prinsen et al. 2018). This study is relevant to researchers, practitioners, and other stakeholders who are 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

In addressing the first aim, 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 44 countries and 20 languages, most assessments and screeners were developed in WEIRD societies and/or in English. Few measurement tools were evaluated across different countries, cultures, and/or language groups.

Psychometric Evaluations of the Identified Measurement Tools

In addressing the second aim, the synthesis revealed four key findings. First, the majority of the identified measurement tools have not been evaluated for all aspects of reliability and validity and few tools met the common acceptability thresholds for these indicators. For example, only four

assessments (ELOM, Snelling et al., 2019; Early Years Toolbox, Howard et al., 2022; Parent Ratings of Numeracy Skills, Lin et al., 2021; US-TEAM-Short, Weiland et al., 2012) and three screeners (ASPENS, Clarke et al., 2011; MESS-E, Erford et al., 1998; 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 seven 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.

Identifying Children with or at-risk of MLD

Second, in terms of diagnostic validity for identifying children with MLD, only the ASPENS (Clarke et al., 2011) 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 .70, the sensitivity results were below the common acceptability threshold of .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., 2020).

Third, predictive validity can also be used to evaluate the suitability of measurement tools for detecting children with or at-risk of MLD over time. However, this study found that only five mathematical assessments and six screeners included evaluations of predictive validity, and none met the common acceptability threshold. However, this may, in part, be due to issues relating consistencies with the external measurement tool. For example, the Early Numeracy Screener showed low predictive validity with the Norwegian national test scores measured 6 months later (Lopez-Peterson et al., 2020). In explaining these results, the authors highlighted inconsistencies in the types of items across the two measurement tools; whilst the Early Numeracy Screener includes untimed items and emphasises accuracy, the national test has timed items and focuses on fluency.

Lack of 'Gold Standard' Measurement Tools

Finally, this study found that only eight of the mathematical assessments and one of the screeners aligned with external measurement tools. The identified measurement tools that did show acceptable levels of concurrent validity were compared to the Differential Ability Scales, Early Number Concepts Scale (Elliot, 2007), Early Numeracy Test- Revised (Van Luit & Van de Rijt, 2009), Math Battery (Fuchs et al., 2003), Preschool Early Numeracy Skills Screener (Purpura & Lonigan, 2015), Research-Based Early Maths Assessment (REMA; Clements et al., 2008), Test of Early Mathematics Abilities- 3rd Version (TEMA-3; Ginsburg & Baroody, 2003), Wechsler Preschool and Primary Intelligence Scales (Wechsler, 1967) and the Woodcock-Johnson test batteries (Schrank et al., 2014; Woodcock & Johnson, 1989; Woodcock et al., 2001).

Many of these external measurement tools are widely recognised in the field of mathematical learning and development. However, there is no agreement on which tools constitute the 'Gold Standard(s)'. This may be, in part, due to the lack of consensus relating to the structure of early maths (Devlin et al., 2022). For example, the TEMA-3 focuses on number and arithmetic skills (Ginsburg & Baroody, 2003), whereas the REMA also includes shape, space, and measure items (Clements et al., 2008). Furthermore, although the Woodcock-Johnson Math subtests (Schrank et al., 2014; Woodcock & Johnson, 1989; Woodcock et al., 2001) and TEMA-3 (Ginsburg & Baroody, 2003), were most frequently used as the base for criterion validity evaluations and thus could be considered an indicator of a 'Gold Standard', these tools are not widely available in different languages and cultures, which limit their usability.

Limitation and Future Directions

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 due to the full-text availability eligibility criteria, which excluded many assessment tools with paywalled psychometric details (e.g., TEMA-3, Woodcock-Johnson Math subtests). A quality assessment of the included studies also fell beyond the scope of the current review.

Directions for future research should focus on developing the reliability and validity evidence of existing measurement tools to help establish 'Gold Standards' in the field of mathematical learning and development. Ideally, these 'Gold Standards' should be suitable for use in different languages, countries, and cultures. To support this process, future research should also work towards a commonly accepted definition of the structure of early mathematics (Devlin et al., 2022), and thus which skills should be included in 'Gold Standard' measurement tools. Future research should also work towards open-access measurement tools that can be used by practitioners (Hakkarainen et al., 2023) and other researchers in low-resource contexts (Pitchford & Outhwaite, 2016).

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 = 37) and screeners (n = 22) were identified in the current review, there remains significant gaps 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.

References

*71 studies identified through systematic review

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: 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.

Alcock, L., Ansari, D., Batchelor, S., Bisson, M. J., De Smedt, B., Gilmore, C., ... & Weber, K. (2016). Challenges in mathematical cognition: A collaboratively-derived research agenda. *Journal of Numerical Cognition, 2*(1), 20.

*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., & 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.

*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.

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.

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

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

*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.

Clements, D. H. & Sarama, J. (2009). *Learning and Teaching Early Math*. London: Routledge.

*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.

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

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.

*Csapó, B., Molnár, G. & Nagy, J. (2014). Computer-Based Assessment of School Readiness and Early Reasoning. *Journal of Educational Psychology*, *106*(3), 639–650.

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

*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 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). Nijmegen: 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., 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.

Dunn, L. M., & Dunn, L. M. (1965). Peabody picture vocabulary test.

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

*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.

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. P., & Pappas, S. (2016). Invitation to the birthday party: Rationale and description. *ZDM Mathematics Education*, *48*, 947-960.

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

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.

*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.

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.

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.

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

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

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

*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.

*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]. Arnhem, The Netherlands: Cito.

*Koponen, T., Salminen, J., Aunio, P., Polet, J., & Hellstrand, H. (2011). *LukiMat - Bedömning av lärandet: 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.

*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, 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.

*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.* Brooklyn: Wireless Generation Inc.

*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].New York, NY: 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.). Novato, CA: 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.

NCII. (2019). *Academic Progress Monitoring Tools Chart Rating Rubric.* National Centre on Intensive Intervention.

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

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., 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.

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

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*.

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.

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

*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.

*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]*, Göttingen, Hogrefe.

RTI International. (2014). *Early Grade Mathematics Assessment (EGMA) Toolkit.* Research Triangle, NC: RTI International

SASC. (2019). *SASC Guidance on assessment of Dyscalculia and Maths Difficulties within other Specific Learning Difficulties*. SASC: SpLD Assessment Standards Committee, UK.

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.

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*.

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.

*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]. *Doetinchem, The Netherlands: Graviant*.

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

*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*]. Frankfurt A.M., Germany: Harcourt Test Services.

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

Wechsler, D. (1997), *Wechsler Intelligence Scale for Children–Third edition* [Greek version]. Athens, Greece: 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.

Woodcock, R. W., & Johnson, M. B. (1989). Woodcock–Johnson: Tests of Achievement–Revised. Allen, TX: DLM.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III NU Complete. *Rolling Meadows, IL: Riverside Publishing*.

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