International Journal of Educational Research 119 (2023) 102172

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

A systematic review of functional numeracy measures for 9–12 -year-olds: Validity and reliability evidence

Airi Hakkarainen [a,b,*], Reinie Cordier [c,d], Lauren Parsons [d], Sangwon Yoon [a], Anu Laine [a], Pirjo Aunio [a], Renée Speyer [d,e,f]

[a] Department of Education, Faculty of Educational Sciences, University of Helsinki, Finland
[b] Department of Education, Faculty of Education and Psychology, University of Jyväskylä, Finland
[c] Department of Social Work, Education and Community Wellbeing, Faculty of Life Sciences, University of Northumbria, United Kingdom
[d] Curtin School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia
[e] Department Special Needs Education, University of Oslo, Norway
[f] Department of Otorhinolaryngology and Head and Neck Surgery, Leiden University Medical Centre, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Functional numeracy
Assessment
Validity
Reliability
Mathematical learning difficulties
COSMIN taxonomy

ABSTRACT

This systematic review aimed to summarize the characteristics and the measurement properties of functional numeracy measures developed for use by teachers among 9–12-year-old children with or without mathematical learning difficulties. A systematic search from five databases was conducted based on pre-defined criteria. PRISMA guidelines were followed for reporting the results. The terminology and classification of measurement properties adopted by the COSMIN taxonomy was used. Twenty-one studies of 18 measures met the inclusion criteria. Most of the identified measures did not report on several measurement properties due to incomplete or missing psychometric data. Knowledge of Mathematical Equivalence, BNPT, and MCS showed most promise based on the completeness of reporting measurement properties. Further validation is needed for all the included measures.

1. Introduction

Educators and researchers need valid and reliable measures to perform a trustworthy and meaningful assessment of students' numeracy skills. Measures with optimal discriminative power are required to identify students who may need additional support via interventions or special needs education. Existing measures available to educators can be divided into curriculum-based measures (CBM), screening measures for group level assessment and more specific measures for individual level. CBMs are frequently used in identifying students who need extra support in their learning of mathematics, but from a psychometric perspective, CBMs by themselves can seldom be used to make decisions (e.g., if a student has achieved particular standards in his/her mathematics learning; Lembke & Stecker, 2007). Further, sensitive screeners are needed at the beginning of the assessment process to identify students whose skills require further evaluation quickly and accurately, after which the more specific individual assessments can be conducted (Van Norman et al., 2018).

Numeracy skills can be divided into early numeracy (EN; Aunio & Räsänen, 2016) and functional numeracy (FN; Geary et al., 2013) skills. EN includes counting skills, basic arithmetic skills, understanding numerical relations, and symbolic and non-symbolic number

* Corresponding author at: University of Jyväskylä, Ruusupuisto, Building RUU, PO Box 35, FI-40014, Finland.
E-mail address: airi.m.hakkarainen@jyu.fi (A. Hakkarainen).

https://doi.org/10.1016/j.ijer.2023.102172
Received 15 September 2022; Received in revised form 9 March 2023; Accepted 17 March 2023
Available online 28 March 2023

sense (Aunio & Räsänen, 2016). These skills provide the basis for the development of FN skills in children aged 9 to 12 years. FN skills include whole number arithmetic, fractions, simple algebra, and measurement as part of problem-solving skills (Geary et al., 2013). Furthermore, the overall concept of FN refers to the fundamental mathematical skills that develop during formal schooling, which are necessary for success in work life in adulthood. Insufficient development in these skills may lead to mathematical learning difficulties, which once established, may be very persistent (e.g., Geary 2011; Vanbinst et al. 2014), leading to devastating problems in later adolescence (Hakkarainen et al., 2015) and adulthood (Geary, 2011). Thus, it is essential to assess children's FN skills regularly in the upper elementary grades when the demands for mastery of FN skills increase significantly (Gersten et al., 2012). For this purpose, teachers need valid and easy-to-use FN measures. Yet, there is a lack of synthesis of the characteristics and psychometric reporting of measures targeting the FN skills of 9–12-year-old children. Hence, general agreement among researchers and educators about the quality of the measurement properties of the measures that are being used to assess children's FN skills is missing. Psychometric reviews can support educators and researchers in selecting measures that are most suitable based on their measurement characteristics and reporting of their measurement properties to identify children in need of additional support when learning mathematics.

In this systematic review, we focus on psychometric properties of FN measures developed for use by teachers in their daily work, thus excluding national and international large-scale measures (e.g., PISA, TIMMS, NAEP, NAPLAN), measures meant for clinicians use only, and assessment batteries meant for general intelligence assessment (e.g., WISC-IV). As the existing knowledge of mathematical difficulties has developed extensively during last decades, the measures developed after 1995 were included.

1.1. Measurement properties

The terminology of measurement properties used in this study is based on the COnsensus based Standards for the selection of health Measurement INstrument (COSMIN) taxonomy (Mokkink et al., 2018). The COSMIN taxonomy (Table 1) comprises nine measurement properties subsumed into three main domains: (1) validity (includes five measurement properties), (2) reliability (includes three measurement properties), and (3) responsiveness (the ability to detect change over time in the construct to be measured).

According to the COSMIN taxonomy, *content validity* is the most important measurement property; each item must be relevant, comprehensive, and comprehensible concerning the construct of interest and the target population (Terwee et al., 2018). Relevance means that all items in a measure should be relevant both to the construct of the measure within the target population and the context of use. Further, every key aspect of the underlying construct should be included in the measure (comprehensiveness) and the target population should understand the items as intended (comprehensibility). Thus, content validity is about the extent to which the content of a measure adequately reflects the construct to be measured in a specified population. *Construct validity* is indicative of the degree to which the scores of a measure are consistent with hypotheses based on the assumption that a measure validly measures the construct to be measured and includes structural validity, hypothesis testing and cross-cultural validity. In addition, *criterion validity*, the degree to which the scores of the measure are an adequate reflection of a 'gold standard' (Mokkink et al., 2018).

Reliability shows the overall consistency of the measure (see Table 1). The reliability of the measure reveals the proportion of total score variance, which is due to true differences among respondents. Reliability includes properties like *test-retest, inter-rater*, and *intra-rater reliabilities*. Test-retest reliability is about the test consistency over time, inter-rater reliability about the degree of agreement among independent raters, and intrarater reliability about the consistence of the rating by the same rater (Mokkink et al., 2018). In addition, a reliable measure performs its measurements precisely and does not include much *measurement error*, which reflects the

Table 1
Definitions of measurement properties according to COSMIN (Mokkink et al., 2018).

Domain	Measurement property	Aspect of measurement property
Reliability	*Degree to which the measurement is free from measurement error*	
	Internal consistency	Degree of the interrelatedness among the items
	Reliability	Proportion of the total variance in the measurements which is because of "true" differences between patients
	Measurement error	Systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured.
Validity	*Degree to which an instrument measures the construct(s) it purports to measure*	
	Content validity	
		Degree to which the content of an instrument is an adequate reflection of the construct to be measured
	Face validity	
		Degree to which an instrument indeed looks as though they are an adequate reflection of the construct to be measured
	Construct validity	
		Degree to which the scores of an instrument are consistent with hypotheses based on the assumption that an instrument validly measures the construct to be measured.
	Structural validity	
		Degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured
	Hypotheses testing	
	Item construct validity	
	Cross-cultural validity	
		Degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument
	Criterion validity	
		Degree to which the scores of an instrument are an adequate reflection of a 'gold standard'
Responsiveness	*Ability of an instrument to detect change over time in the construct to be measured*	

systematic and random error of a respondent's score that is not due to true changes in the construct measured. *Internal consistency* reveals how well the items of the scales are interrelated (Mokkink et al., 2018). The third main domain, *responsiveness* (i.e., the ability of a measure to detect change over time) was out of the scope of this review.

To the best of our knowledge, no systematic reviews have been published to date investigating the measurement properties of existing FN measures for children aged between 9 and 12 years. Systematic reviews have been done mainly about interventions both in early numeracy (Park & Nelson, 2022) and in middle school children (Powell et al., 2021), but not about the measurement properties of the FN measures themselves. The main aim of this systematic review was to identify and describe FN measures used by teachers at elementary schools to identify children aged 9–12 years in need of additional support for learning mathematical skills, and to evaluate the measurement properties that have been reported of included measures.

2. Method

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement and checklist (Page et al., 2021). The COSMIN methodological guidelines and terminology (Prinsen et al., 2018; Terwee et al., 2018) were followed.

2.1. Eligibility criteria

The following inclusion and exclusion criteria were used to identify eligible measures: (1) measures assessed functional numeracy (FN) skills; (2) at least one subscale or a minimum of 50% of the total number of items of a measure relates to FN; (3) measures targeted children between nine and twelve years old; (4) measures were developed for use by elementary school teachers; and (5) measures were developed and studies of their measurement properties were published in English after year 1995. Psychometric studies were eligible if they reported on any measurement properties related validity or reliability of the included measures as defined in the COSMIN taxonomy (Mokkink et al., 2018); responsiveness was the only measurement property that was excluded. Measures on health literacy, intelligence (e.g., Wechsler Intelligence test for children [WISC]; Wechsler, 2014), large-scale psychological achievement tests (e.g., Kaufman Assessment Battery for Children [K-ABC]; Kaufman, 2005), and (inter)national large-scale mathematical tests (e.g., PISA, TIMMS, NAEP, NAPLAN) were outside the scope of this review.

2.2. Data sources and search strategies

Systematic literature searches were conducted across the following five databases to identify eligible studies: CINAHL, Embase, Eric, PsycINFO, and PubMed. Reference lists of eligible articles were checked for additional studies. Measure developers were contacted by e-mail if measures were not freely available, requesting access to the original measures.

Search strategies to identify FN measures and psychometric studies were conducted across all five electronic databases by combining both subject headings and free text terms related to numeracy and psychometrics. The full search strategies can be found in Appendix A. After the final measure selection, further literature searches were performed to identify additional psychometric studies, using names and acronyms of the included measures, and limiting results by measures' publication year.

2.3. Study selection and risk of bias of individual studies

Two independent reviewers first reviewed all titles and abstracts for eligibility. Next, both reviewers assessed the original articles for eligibility. Disagreements between reviewers were discussed and resolved via consensus by the research team. Weighted Kappa was calculated to assess the inter-rater agreement between reviewers: Kw = 0.86 (95% CI: 0.84–0.88), which was very good indicating no biased ratings between the raters. None of the reviewers or co-authors had formal or informal affiliations with any of the authors of the included studies and measures. Therefore, there was no evident bias in article selection or methodological study quality rating.

2.4. Data extraction

After completion of the selection process of the studies, the following data were extracted from the included studies and measures: (1) study characteristics (i.e., study population, age, grade); (2) characteristics of measures (i.e., measure names and acronyms, measured constructs, number of scales and subscales, number of items, response options and duration); and (3) study results on all available measurement properties (Mokkink et al., 2018). One reviewer extracted all data that were checked for accuracy by a second reviewer. Differences between the two reviewers were checked by a third reviewer.

2.5. The assessment of overall methodological quality

After completing the COSMIN procedure, the Qualsyst critical appraisal tool (Kmet et al., 2004) was used to assess the overall methodological quality and the potential risk of bias for each study. The Qualsyst is a commonly used quality assessment checklist for evaluating primary research papers from a variety of fields (Kmet et al., 2004). The checklist that consists of 14 criteria (2 = *meets the criterion*; 1 = *partially meets the criterion*; 0 = *does not meet the criterion*) was used to assess the overall methodological quality of individual studies. A total score was derived by adding up the scores from the 14 criteria, with the lowest possible score of 0 and the

(caption on next page)

Fig. 1. Flow diagram of the reviewing procedure based on PRISMA 2020 statement and checklist (Page et al., 2021).

highest possible total score varying from 22 to 28. The total score was converted to an overall quality percentage score by dividing the received total score by the possible total score, and the multiplying that value with 100. An overall quality percentage score of 80% or higher indicates strong methodological quality, a score between 70% and 79% indicates good quality, a score between 50% and 69% adequate quality, and a score below 50% poor quality. Two researchers scored the included studies against the 14 criteria independently. Disagreements about the scoring between the raters were resolved by consensus.

3. Results

3.1. Systematic literature searches

The literature search on May 6th 2021 from the five databases resulted in 6831 abstracts; after removing duplicates and those marked with "not article" by automation tools, 3522 articles were screened. A total of 1100 articles and 20 measures were assessed for eligibility resulting in 13 original studies reporting on validity and/or reliability properties of 11 FN measures. Further, additional literature searches on July 16th 2021 using names and acronyms of the 11 measures and reference checking of the included 13 articles, identified 46 research articles to be assessed for eligibility. Finally, seven new measures and eight psychometric studies were included, resulting in a total of 18 measures and 21 psychometric studies that were included in this review. In Appendix B, the 19 measures that were excluded and the reasons for exclusion are summarised. Fig. 1 presents the flow diagram of the included studies and measures.

3.2. Characteristics of included measures and studies on measurement properties

Of the 18 measures (Table 2), four were curriculum-based measures, 12 screening and two diagnostic assessments. Some of the measures were designed for several age groups from six up to 14 years of age (e.g., NSCT; Castro et al., 2017). Therefore, in the end, the age range of the participants was 5–16 years in the selected studies. Sixteen of the 18 measures were targeted both children with or without mathematical learning difficulties (MLD); one measure (DAA; Kunina-Habenicht et al., 2017) targeted children with below average mathematical skills; and one measure (ERT; Gebhardt et al., 2014) targeted children with special educational needs in learning. Descriptions and characteristics of the included measures are presented in Table 2.

All the measures were teacher-administered; five measures via computer: BNPT (Olkun et al., 2016), Division of fractions (Ketterlin-Geller et al., 2013), MAP (Klingbeil et al., 2019), NSCT (Castro et al., 2017), and Representing Rational Numbers (Ketterlin-Geller et al., 2019). Administration format was not reported in CBM-WPS-Fluency (Jitendra et al., 2014, 2005), ERT (Gebhardt et al., 2014), and South Africa Numeracy test (Kivilu, 2010). The rest of the measures (nine) were administered in paper-and-pencil format. The administration format of NAS (Looveer & Mulligan, 2009) was an interview. The content of the measures varied extensively and included problem-solving and strategies (mathematical creativity and word problems), symbolic and non-symbolic number sense, arithmetic skills, and fluency (addition, subtraction, multiplication, and division), place-value, enumeration, pattering, grouping, space concepts (geometry), measurement concepts (e.g., time, length, area), and early extensions of algebra (equal sign). Item numbers ranged from five (MCS; Akgul & Kahveci, 2016) to 250 item bank (Division of fractions; Ketterlin-Geller et al., 2013), and duration time from 2 min (IPAM; de León et al., 2021) to 90 min (DAA; Kunina-Habenicht et al., 2017). Two measures, ANWT (Moura et al., 2015) and Representing Rational Numbers (Ketterlin-Geller et al., 2019), did not have any time limits. For seven measures, the administration time of the test was not reported.

3.3. Validity evidence of the included measures

The validity properties of the measures that were found (content validity, and construct validity including structural validity and hypothesis testing) are summarised in the Table 3. Although content validity is the most important measurement property according to COSMIN taxonomy (Terwee et al., 2018), it was reported only for five measures: BNPT (Olkun et al., 2016), Division of fractions (Ketterlin-Geller et al., 2013), Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011), MCS (Akgul & Kahveci, 2016), and Representing Rational Numbers (Ketterlin-Geller et al., 2019) in six studies. For the three content validity domains (i.e., relevance, comprehensibility, and comprehensiveness) relevance of the items was evaluated for two measures (BNPT and MCS) in two studies; relevance and comprehensibility of the items for one measure (Knowledge of Mathematical Equivalence) in two studies; and relevance, comprehensiveness, and comprehensibility of the items for two measures (Division of fractions and Representing Rational Numbers) in two studies. Usually, feedback of the items was received from either panel experts or experts in mathematics education. Based on the experts' feedback, measures were adapted, items removed or added, or language rephrased. Feedback from students were only sought for one measure, Knowledge of Mathematical Equivalence, where children participated in a pilot study. The most thorough evaluation of content validity was done for the Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011) and Representing Rational Numbers (Ketterlin-Geller et al., 2019). For both measures, content validity was evaluated in terms of relevance, comprehensiveness, and comprehensibility.

According to the COSMIN taxonomy (Mokkink et al., 2018), the structure of a measure should be tested with a factor analysis when using classical test theory (CTT). In testing structural validity, confirmatory factor analysis (CFA) is preferred over exploratory factor analysis (EFA) and principal component analysis (PCA). Furthermore, Item Response Theory (IRT) could be applied using Rasch

Table 2
Summary of the measures included in the review.

Acronym Name of the Measure (Authors; Publication date)	*Grade level Age-group Target population*	*Assessment tasks (Items n): Task description*	*Administration Scoring*
ANWT Arabic Number-Writing Task (Moura et al., 2015)	Grades 1–4 6–12 years Children with or without MLD	Arabic number transcoding (n = 28): One- to four-digit numbers dictated and transcribed as Arabic numbers.	Administration: Pencil and paper; Individual and group; No time limits. Scoring: 0 = incorrect, 1 = correct
BNPT Basic Number Processing Test (Olkun et al., 2016)	Grade 1–4 6–11 years Children with or without MLD	Canonic dot counting (n = NR) Symbolic number comparison (n = NR) Mental number line (n = NR)	Administration: Tablet PCs; Individual; Time NR. Scoring: 0 = incorrect solution, 1 = correct solution; Response reaction time.
CBM – Math Computation Curriculum-based Measurement of Math Computation (Shapiro et al., 2006)	Grade 3–5 Age NR Children with or without MLD	Math computation (n = 25): Grade 1–2: addition and subtraction; Grade 3: simple multiplication and division; Grade 4: fractions and multi-digit multiplication; Grade 5: decimals, complex fractions, multi-digit division	Administration: Pencil and paper; Mode NR; Time NR. Scoring: Number of correct digits
CBM – Math Concept Application Curriculum-based Measurement of Math Concept Application (Shapiro et al., 2006)	Grade 3–5 Age NR Children with or without MLD	Math concepts and application (Grade 2 level n = 18; Grade 3–6 level n = 24): Concept areas included counting, number concepts, names of numbers, measurement, charts and graphs, money, fractions, applied computation, word problems. Problems required 1–3 responses (fill-in-the-blank, multiple choice).	Administration Pencil and paper; Mode NR; Time NR. Scoring: 0 = incorrect solution, 1 = correct solution
CBM-WPS-Fluency Curriculum-Based Measurement of Word Problem-Solving Fluency (Jitendra et al., 2005, 2014)	Grade 3 8–10 years Children with or without MLD	Mathematical word problems (n = 8): Change, group and compare problem types, requiring application of addition and/or subtraction computation skills of one- and two-digit numbers. 6 × 1-step problems and 2 × 2-step problems.	Administration: Format NR; Group; 10 min time limit. Scoring: 0 = incorrect number model, 1 = correct number model; 0 = incorrect solution, 1 = correct solution (possible total of 2 points per problem)
DAA Diagnostic Arithmetics Assessment (Kunina-Habenicht et al., 2017)	Grade 4 9–10 years Children with below-average mathematical ability	Context-free arithmetic problems (n = 40): 10 x addition, 10 x subtraction, 10 x multiplication, 10 x division items. Contextualized word problems (n = 18): Covers knowledge of numbers and operations (understanding number representations, proficiency in performing basic arithmetic operations, application of basic arithmetic operations to contextualized word problems), and measures modeling, and working with mathematical symbols and formal and technical elements of mathematics.	Administration: Pencil and paper; Group; 2 × 45 min school periods (10 min break between) Scoring: 0 = incorrect solution, 1 = correct solution
Division of fractions (Ketterlin-Geller et al., 2013)	Grade 5–7 Age NR Children with or without MLD	Division of fractions (n = 18; total item bank = 250): Computations and contextualised problems covering the component processes, strategies and knowledge inherent in the adopted model of divisions for fractions with understanding. Students randomly assigned 18 item test (5 = anchor items, 13 = unique items). Multiple choice response scale (3 distractors, 1 correct response).	Administration: Computer; Group administered; Time NR. Scoring: 0 = incorrect solution, 1 = correct solution
DMA Diagnostic Mathematics Assessment (Kunina-Habenicht et al., 2009)	Grades 3–4 8–11 years Children with or without MLD	Context-free arithmetic problems (n = 52): 14 x simple addition, 14 x simple subtraction, 12 x simple multiplication, 12 x simple division. Contextualized word problems (n = 35): 3 x addition only, 3 x subtraction only, 3 x multiplication only, 3 x division only, 22 x combination of two skills	Administration: Pencil and paper; Group; 2 × 45 min school periods (10 min break between) Scoring: 0 = incorrect solution, 1 = correct solution
ERT (Adapted) Eggenberger RechenTest (Gebhardt et al., 2014)	Grades 5–9 11–16 years Children with SEN-L	Basic arithmetical skills (n = 33): 13 x basic numeracy items taken from ERT 4+ (addition, subtraction, multiplication, division), 6 x place holder items ERT 4+ (e.g., __ + 8 = 21), 15 x items developed by authors. Word problems (n = 9): word problems taken from ERT 3+ Number series (n = 14): measures knowledge of place-value system (12 x ERT 4+ items, 2 x items developed by authors)	Administration: Format NR; Mode NR; Time NR. Scoring: 0 = incorrect solution, 1 = correct solution

(*continued on next page*)

Table 2 (*continued*)

Acronym Name of the Measure (Authors; Publication date)	*Grade level Age-group Target population*	*Assessment tasks (Items n): Task description*	*Administration Scoring*
		Writing numbers from dictation ($n = 14$): Items developed by authors	
IPAM Indicadores de Progreso de Aprendizaje en Matemáticas/ Indicators of Basic early Math Skills (de León et al., 2021)	Grade 1 & 3 6–8 years Children with or without MLD	Number comparison ($n = 64$): identify the largest number in a pair (e.g., [13–41]). Missing number ($n = 45$): identify a missing numbers from a set of three (e.g., [27, 29, _]). Single digit computation ($n = 45$): addition, subtraction, and multiplication problems with numbers 1–9. Multi-digit computation ($n = 45$): addition, subtraction, and multiplication problems with numbers 1–99. Place value ($n = 45$): identify a number presented in pictorial form based on 10-base blocks 3 parallel forms (A, B, C)	Administration: Pencil and paper; Group; 2 min time limit. Scoring: 0 = incorrect solution, 1 = correct solution
Knowledge of Mathematical Equivalence (Rittle-Johnson et al., 2011; Matthews et al., 2012)	Grades 2–6 8–12 years Children with or without MLD	Equation solving (Initial long form $n = 28$; revised short forms x 2 $n = 11$): ability to solve equations at four knowledge levels (rigid operational [e.g., $a + b = c$]; flexible operational [e.g., $c = a + b$]; basic relational [e.g., $a + b = c + d$]; comparative relational [e.g., If $56 + 85 = 141$, does $56 + 85 - 7 = 141 - 7$])0. Equation structure (Initial long form $n = 31$; revised short forms x 2 $n = 18$): knowledge of valid equation structures through: 1) evaluate4d equations as true/false; 2) explaining true/false evaluations; 3) reconstructing equations from memory Equal sign (Initial long form $n = 11$; revised short forms x 2 $n = 8$): define the =, rate given definitions of =, select best definition of = Initial long form developed, revised to two shorter forms ($n = 37$; 33 unique items/form, 4 overlapping items/form).	Administration: Pencil and paper; Group; Limited time per task: Equation solving = 10 min, Equation structure = 10 min, Equal sign = 5 min. Scoring: 0 = incorrect solution, 1 = correct solution (for computation items solutions considered correct if within 1 of correct answer)
MAP Measures of Academic Progress (Klingbeil et al., 2019)	Grades 6–8 Children with or without MLD	Math section of academic progress screen: ($n = 50$ items; Number administered a function of performance on previously administered items due to adaptive nature of test): Content covered: computation and problem solving, number sense, geometry, measurement, data, statistics, and probability, algebraic concepts. Multiple choice and short answer responses	Administration: Computer administered, adaptive test; Approx. 45 min Scoring: NR; Score range = 100–350
M-CBMs Math Curriculum-Based Measurements (Strait et al., 2015, 2018)	Grade 6 11–12 years Children with or without MLD	Math computation (n = NR; At least eight rows of problems per form; 4 forms): Problems included: 1) three-digit plus three-digit addition; 2) three-digit minus three-digit subtraction; 3) two-digit by two-digit multiplication; 4) three-digit divided by two-digit division.	Administration: Pencil and paper; Group; 5 min time limit Scoring: 1 point for each correct digit; Fluency score = sum of correct digits in 5 min; Accuracy score = number of correct digits/total digits attempted
MCS Mathematics Creativity Scale (Akgul & Kahveci, 2016)	Grades 5–8 10–15 years Children with or without MLD	Mathematical creativity task: ($n = 5$ items): Tasks ask students to find areas (geometry), generate mathematical problems according to two unknowns, use geometrical intuition, generate word problems for an arithmetic operation, and use logical thinking. Items scored for fluency (number of ideas produced per item), flexibility (number of categories ideas fell into) and originality (rare ideas given higher marks).	Administration: Pencil and paper; Group; 50 min time limit Scoring: Fluency score = one point for every idea produced; Flexibility score = one point for every category; Originality score = scored according to table provided (range 0–9)
NAS Numeracy Achievement Scale (Looveer & Mulligan, 2009)	Grades K–6 5–13 years Children with or without MLD	Understanding of mathematical concepts: (Kindergarten $n = 16$/form, Year 1 $n = 16$/ form, Year 2 $n = 20$/form, Year 3 $n = 20$/ form, Year 4 $n = 25$/form, Year 5 $n = 25$/ form, Year 6 $n = 25$/form): Concepts included were number (counting, place-value, numeration, pattering, grouping, four	Administration: Assessment interview; Individual; Time NR Scoring: 1= correct (if not guessed or response did not have an incorrect mathematical basis); 0 = incorrect, guessed, or incorrect mathematical basis

(*continued on next page*)

Table 2 (*continued*)

Acronym Name of the Measure (Authors; Publication date)	*Grade level Age-group Target population*	*Assessment tasks (Items n): Task description*	*Administration Scoring*
		arithmetic process, mental computation, fractions and decimals), space (two- and three-dimensional shapes, transformation, position, location and graphs), measurement (length, area, volume and capacity, mass and time). Two forms/Year Level; Some link items between forms for the same year level, and across different year levels.	
NSCT Nonsymbolic and Symbolic Comparison Tasks (Castro et al., 2017)	Grades 1–6 6–14 years Children with or without MLD	Nonsymbolic comparison task: (n = 60): select the square containing the most/least circles Symbolic comparison task: (n = 60): select the Arabic digit with the largest/smallest value Pairs included numerosities 1 to 9, and relationships between numbers within pairs varied among 4 conditions: small ratios (0.33 and 0.50), large ratios (0.66, 0.75 and 0.85), close numerical distances (1 and 2), and far numerical distances (4 and 5). Presented in two separate blocks of 30 stimuli each. Children instructed to answer quickly and accurately	Administration: Computer-based; Individual; 30–40 min Scoring: Accuracy score =% correct items; Adjusted reaction time (adjRT) = median reaction time of correct responses - median reaction time for the condition; Efficiency measure (EM) = adjRT/proportion of correct responses
Representing Rational Numbers (Ketterlin-Geller et al., 2019)	Grade 5–7 Age NR Children with or without MLD	Mathematical word problems (n = 39): Assessed representations of positive rational numbers. Core concepts evaluated: equivalent fractions, decimals, comparing fractions, conversion between representations. Multiple choice responses	Administration: Computer-based; Mode NR; Untimed Scoring: 1 = correct, 0 = incorrect
South Africa Numeracy Test (Kivilu, 2010)	Grade 3 Age NR Children with or without MLD	Counting and ordering tasks (n = 30 items) Operations-addition tasks (n = 30 items) Operations-subtraction tasks (n = 28 items) Operations-multiplication tasks (n = 26 items)	Administration: Format NR; Mode NR; Time NR. Scoring:% items correct

Note. MLD = Mathematical Learning Difficulties, SEN-*L* = Special Educational Needs in Learning, RT = Reaction time, NR = Not reported.

analyses to determine the dimensionality of a scale.

Structural validity was reported for 11 measures in 12 studies. Results of CFA was retrieved in three studies for three measures: DMA (Kunina-Habenicht et al., 2009), Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011), and Representing Rational Numbers (Ketterlin-Geller et al., 2019). The structural validity was evaluated using both CTT and IRT for the following two measures: Representing Rational Numbers (Ketterlin-Geller et al., 2019) and Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011). The following four measures evaluated structural validity using only IRT: Division of fractions (Ketterlin-Geller et al., 2013), ERT (Gebhardt et al., 2014), NAS (Looveer & Mulligan, 2009), and South African Numeracy test (Kivilu, 2010). Other analyses that were used in testing structural validity were EFA for MCS (Akgul & Kahveci, 2016), diagnostic classification model for DAA (Kunina-Habenicht et al., 2017), and bi-serial correlations for BNPT (Olkun et al., 2016).

In studies that reported structural validity, the results showed that structural validity of the measures was good indicating that the scores of these measures are an adequate reflection of the dimensionality of the construct of functional numeracy (FN). For DMA (Kunina-Habenicht et al., 2009) and MCS (Akgul & Kahveci, 2016), the factor loadings were available showing strong factor structure. The most thorough evaluation of structural validity was done for Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011). For this measure, structural validity was evaluated using CFA, PCA, and IRT in two studies. Evidence of structural validity was not evaluated for seven measures (see Table 3). This indicates a clear lack of evidence of the structural validity for those measures.

Table 3 shows that hypotheses testing for construct validity was reported for 11 measures in 15 studies and convergent validity, the correlations between different FN measures, was retrieved for nine measures in 12 articles. Further, discriminant validity between different groups (e.g., children with or without MLD, grade groups), was found for ANWT (Moura et al., 2015). Both convergent and discriminant validity were reported for two measures, MAP (Klingbeil et al., 2019) and NSCT (Castro et al., 2017). Evidence for convergent validity varied from medium to strong. Support for discriminant validity was indicated by significant differences between groups that were expected (e.g., children with or without MLD for the ANWT; Moura et al., 2015) and non-significant difference between groups that were not expected (e.g., NSCT between grades 1 and 2; Castro et al., 2017). Hence, the results produced evidence that was consistent with hypotheses that were formulated *a priori* based on the assumption that a measure would behave in a particular manner. Cross-cultural validity was reported only for South Africa Numeracy Test in terms of differential items functioning (Kivilu, 2010). Criterion validity was not evaluated because there is no gold standard and short and long versions of measures were not

Table 3
Summary of the validity properties reported in the included articles.

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity		*Hypothesis testing*	
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Aspect/Method*	*Results*
ANWT Arabic Number-Writing Task	Moura et al. (2015)	NR	NR	NR	NR	Discriminative validity	Significant difference in the individual items error rates between children with and without mathematical difficulties: $F = 7.63$, $p < .001$ (1st grade); $F = 153.36$, $p < .005$ (2nd grade); $F = 35.67$, $p < .001$ (3rd grade); $F = 15.14$, $p < .001$ (4th grade).
BNPT Basic Number Processing Test	Olkun et al. (2016)	Relevance	Panel experts gave feedback: feedback results NR, adaptations made to test based on feedback.	Bi-serial correlations	Canonical dot counting: Yr 1&2, all items ≥.5; Yr 3 = ≥.62 (2 items <0.15); Yr 4 = ≥.46 (1 item <0.15). Symbolic number comparison: Yr 1 = 0.2 - ≥.54; Yr 2 = ≥.5 (1 item <0.15); Yr 3 = ≥.50 (1 item <0.15); Yr 4 = ≥.50 (1 item <0.15) Mental number line: Yr 1 = 0.25–0.65; Yr 2 = ≥.5 (3 items <0.15); Yr 3 = ≥.3 3 items <0.15); Yr 4 = ≥.19 (1 item <0.15).	Convergent validity	Year 1: subtests explain 15% of variability in Match Achievement Test (MAT) score; small negative partial correlations with MAT: *r* ranged from −0.334 to −0.160. Year 2: subtests explain 60% of variability in MAT score; medium to large negative partial correlations with MAT: *r* ranged from −0.635 to −0.399. Year 3: subtests explain 45% of variability in MAT score; medium to large negative partial correlations with MAT: *r* ranged from −0.564 to −0.362. Year 4: subtests explain 39% of variability in MAT score; medium to large negative partial correlations with MAT: *r* ranged from −0.560 to −0.320.
						Convergent validity	Year 1: small negative partial correlations with CPT: *r* ranged from −0.374 to −0.253. Year 2: small to large negative partial correlations with CPT: *r* ranged from −0.536 to −0.111. Year 3: small to medium negative partial correlations with CPT: *r* ranged from −0.495 to −0.313. Year 4: small to medium negative partial correlations with CPT: *r* ranged from −0.567 to −0.196.
CBM – Math Computation	Shapiro et al. (2006)	NR	NR	NR	NR	Convergent validity	Trival to medium positive correlations with Pennsylvania

(*continued on next page*)

Table 3 (*continued*)

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity			
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Hypothesis testing Aspect/Method*	*Results*
Curriculum-based Measurement of Math Computation							System of School Assessment (PSSA) scores in Fall: *r* ranged from 0.072 to 0.408. Medium positive correlations with PSSA scores in Winter: *r* ranged from 0.505 to 0.525. Medium positive correlations with PSSA scores in Spring: *r* ranged from 0.519 to 0.521.
CBM – Math Concept Application Curriculum-based Measurement of Math Concept Application	Shapiro et al. (2006)	NR	NR	NR	NR	Convergent validity	Medium positive correlations with PSSA scores in Fall: *r* ranged from 0.457 to 0.479. Medium positive correlations with PSSA scores in Winter: *r* ranged from 0.613 to 0.641. Medium positive correlations with PSSA scores in Spring: *r* ranged from 0.561 to 0.644.
CBM-WPS Curriculum-Based Measurement of Word Problem Solving	Jitendra et al. (2005)	NR	NR	NR	NR	Convergent validity	Medium to large positive correlations with the two subtests (mathematics problem solving; mathematics procedure) of the Stanford Achievement Test: *r* ranged from 0.38 to 0.71. Medium to large positive correlations with the two subtests (mathematics concepts and applications; mathematics computation) of the TerraNova achievement test: *r* ranged from 0.48 to 0.69. Small to medium positive correlations with the Basic Math Computation Fluency Measure: *r* ranged from 0.26 to 0.45, $p <$ .05.
	Jitendra et al. (2014)	NR	NR	NR	NR	Convergent validity	Large positive correlations with the Number Combinations fluency: *r* ranged from 0.52 to 0.64. Medium positive correlation with the Measure of Academic Progress in mathematics: *r* ranged from 0.37 to 0.45.
DAA Diagnostic	Kunina-Habenicht et al. (2017)	NR	NR	DCM	4-factor structure (addition/ subtraction; multiplication/	Convergent validity	Large positive correlations with the National Educational

(*continued on next page*)

Table 3 (*continued*)

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity		*Hypothesis testing*	
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Aspect/Method*	*Results*
Arithmetics Assessment					division; modeling skills; skills for using measurement units): AIC = 72,096, BIC = 73,159; correlations between factors from 0.26 to 0.89.		Standards for mathematics in elementary school in Germany: $r = 0.61$.
				IRT	Unidimensional 1-factor structure (general arithmetic ability): AIC = 71,407, BIC = 72,163.		
Division of fractions	Ketterlin-Geller et al. (2013)	Relevance Comprehensiveness	Panel experts developed theory for construct via textbook consultation and discuss. Theory reviewed and refined by content experts. Problems fitting develop theory developed by research team.		IRT	Mean square of residual fit statistics: mean = 0.99 (SD=0.32). 13 items underfit, 4 items overfit. Item reliability = 0.83.	NR
NR		Relevance Comprehensibility	Items developed by research team and evaluated by panel experts. No items removed for relevance. Language in Items adjusted to improve comprehensibility				
DMA Diagnostic Mathematics Assessment	Kunina-Habenicht et al. (2009)	NR	NR	CFA	5-factor structure (addition, subtraction, multiplication, division, modeling): CFI = 0.951, RMSEA = 0.063 (children in grade 3); CFI = 0.981, RMSEA = 0.056 (children in grade 4); factor loadings of all items from 0.51 to 0.97 (in grade 3); factor loadings of all items from 0.45 to.99 (in grade 4).	NR	NR
ERT Eggenberger RechenTest	Gebhardt et al. (2014)	NR	NR	IRT	4 subtest structure (basic arithmetical skills; number series; word problems; writing numbers from dictation): No significant differences in item difficulty of all items for each factor across sample children with different numeracy abilities, $p > .01$; correlations between factors from 0.64 to 0.75.	NR	NR
IPAM Indicadores de Progreso de Aprendizaje en Matemáticas/	de León et al. (2020)	NR	NR	NR	NR	Convergent validity	Medium-large positive correlations of all forms with the Cálculo numérico/numerical computation measure (Sn), from La Batería de Aptitudes

(*continued on next page*)

Table 3 (*continued*)

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity		*Hypothesis testing*	
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Aspect/Method*	*Results*
Indicators of Basic early Math Skills							Diferenciales y Generales E2/ The Battery of Differential and General Abilities E2 (BADyG-E2): $r = 0.47–0.70$.
						Convergent validity	Medium-large positive correlations between forms: Fall/Winter = 0.57–0.86; Fall/ Spring = 0.54–0.86; Winter/ Spring = 0.59–0.90
	de León et al. (2021)	NR	NR	NR	NR	Convergent validity	Medium-large positive correlations of all forms with the Cálculo numérico/numerical computation measure (Sn), from La Batería de Aptitudes Diferenciales y Generales E2/ The Battery of Differential and General Abilities E2 (BADyG-E2): $r = 0.36–0.69$
						Convergent validity	Medium-large positive correlations between forms: Fall/Winter = 0.43–0.79; Winter/Spring = 0.50–0.80
Knowledge of Mathematical Equivalence	Matthews et al. (2012)	Relevance	Panel experts gave feedback: rated nearly all items; range from important to essential. Mean rating of 4.3; Five items from the original assessment were removed and eight items were added	PCA	Unidimensional 1-factor structure (equal-sign knowledge): First factor explained for 60% of total variance, and second factor explained only 2%; factor loadings of all items higher than 0.45.	Convergent validity	Large positive correlation with the Tennessee Comprehensive Assessment Program (TCAP) mathematic scores in grades 3–6: $r = 0.70$.
				IRT	Unidimensional 1-factor structure (equal-sign knowledge): Infit mean squares between 0.5 and 1.5.		
	Rittle-Johnson et al. (2011)	Relevance	4 experts in mathematics education were surveyed whether the items are relevant to measure the knowledge of mathematical equivalence; the irrelevant items were revised based on the experts' input.	PCA	Unidimensional 1-factor structure (mathematical equivalence knowledge): First factor explained for 57.2% of total variance, and second factor explained only 2.2%.	Convergent validity	Large positive correlation with the mathematic scores of Iowa Tests of Basic Skills in grades 3–6: r ranged from 0.79 to 0.80.
		Comprehensibility	24 students from 2nd to 4th grades were asked about whether the items are confusing to understand; the confusing items were eliminated or reworded based on the students' inputs.	CFA	Unidimensional 1-factor structure (mathematical equivalence knowledge): CFI = 0.980, SRMR = 0.121. A single factor captured a majority of the variance and performance on individual items, suggesting that the construct was unidimensional.		

(*continued on next page*)

Table 3 (*continued*)

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity			
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Hypothesis testing Aspect/Method*	*Results*
				IRT	Unidimensional 1-factor structure (mathematical equivalence knowledge): item-total correlations higher than 0.2; infit and outfit mean squares between 0.5 and 1.5.		
MAP Measures of Academic Progress	Klingbeil et al. (2019)	NR	NR	NR	NR	Convergent validity	Large positive correlations with the mathematic scores of 2016 and 2017 Forward Exam, and AIMSweb Match Computation and Math Concepts and Application probes grades 6–8: r ranged from 0.743 to 0.877.
						Discriminative validity	Significant difference in proficiency scores between participating schools in Grade 6: $\chi 2(1) = 5.91$, $p = .015$, and Grade 8: $\chi 2(1) = 6.89$, $p < .001$. No significant difference in proficiency scores between schools in Grade 7: $\chi 2\ (1) = 0.403$, $p = .525$.
M-CBMs Math Curriculum-Based Measurements	Strait et al. (2015)	NR	NR	NR	NR	Convergent validity	Medium-large positive correlations between forms: coefficients for fluency ranged from 0.41 to 0.81 ($M = 0.66$); coefficients for accuracy ranged from 0.47 to 0.78 ($M = 0.65$)
	Strait et al. (2018)					Convergent validity	M-CBM scores were strongly correlated (e.g., $r = 0.82$–.85) with the Palmetto Achievement Challenge Test (PACT) math section
MCS Mathematics Creativity Scale	Akgul and Kahveci (2016)	Relevance	7 experts were interviewed whether the items are relevant to measure mathematical creativity; items accepted by at least 5 of 7 experts were retained 40 mathematical education researchers and mathematics teachers were asked to answer the following question: "Which of the items in the test are able to measure the mathematical creativity of a middle school student. Did not lead to excluding items	EFA	Unidimensional 1-factor structure (mathematical creativity): First factor explained for 42% of total variance; factor loadings of all items from 0.60 to 0.71.	NR	NR

(*continued on next page*)

Table 3 (*continued*)

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity		*Hypothesis testing*	
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Aspect/Method*	*Results*
NAS Numeracy Achievement Scale	Looveer and Mulligan (2009)	NR	NR	IRT	Unidimensional 1-factor structure (numeracy): Item-person fit residuals between −2.5 and 2.5. Items were ordered according to chi-square statistics as produced by RUMM.	NR	NR
NSCT Nonsymbolic and Symbolic Comparison Tasks	Castro et al. (2017)	NR	NR	PCA	2-factor structure (nonsymbolic and symbolic numerical comparison; verbal and visuospatial working memory) solution: Both factors explained 75% of total variance.	Convergent validity Discriminative validity	Partial correlations between exact mental arithmetic and both, nonsymbolic ($r = –.29$, $p < .01$) and symbolic efficiency ($r = –.46$, $p < .001$). No statistically significant differences were found in the relationship between the tasks for children in grade 1 and grade 2, but both groups were significantly different from children of all the remaining grades. Second graders were not significantly different from third graders, but were significantly different compared to fourth, fifth and sixth graders. In contrast, third graders were only significantly different from sixth graders. No statistically significant differences among fourth, fifth and sixth graders were found.
Representing Rational Numbers	Ketterlin-Geller et al. (2019)	Relevance Comprehensiveness	In-depth analysis of the literature to develop and articulate conceptual framework. Framework refined by expert panel.	IRT	Item difficulty and fit assessed after piloting. Items retained if: 1) mean ability of students choosing the correct response > mean ability of students choosing an incorrect response; 2) item difficulty parameter = −4.0 to +4.0; 3) item discrimination parameter = 0.5 to 2.0; 4) item χ^2 fit statistic >0.01. 39 of the original 110 items retained.	NR	NR
		Comprehensibility	Pilot and cognitive interviewing with 20 students (Grades 5–8)		Item difficulty and fit assessed again: item difficulty estimates = −1.40 - 1.99 (1 item excepted); item discrimination parameters = 0.5– 2.0; acceptable model fit statistics		

(*continued on next page*)

Table 3 (*continued*)

Measure	*Reference(s)*	*Content validity*		*Construct validity*			
				Structural validity		*Hypothesis testing*	
		Aspect/Method	*Results*	*Aspect/ Method*	*Results*	*Aspect/Method*	*Results*
				CFA	Maximum likelihood estimation confirmed unidimensionality: $\chi^2(2)=7.17$, $p=.03$, RMSEA=0.08 (95% CI = 0.02–0.14)		
South Africa Numeracy Test	Kivilu (2010)	NR	NR	IRT	4 factor-structure (counting and ordering; operations-addition; operations-subtraction; operations-multiplication): No significant differences in item difficulty of all items for each factor between gender, $p > .05$.	NR	NR

Note. AIC = Akaike Information Criterion; BIC = Bayesian information criterion; RMSEA = root mean square error of approximation; SRMR = residual-based standardized root mean square residual, NR = not reported.

compared.

3.4. Reliability evidence of the included measures

The reliability measurement properties (internal consistency, reliability, and measurement error) of all measures are summarised in Table 4. According to the COSMIN taxonomy (Mokkink et al., 2018), when evaluating *internal consistency* of measures either Cronbach's alpha or Kuder-Richardson (KR-20) should be calculated for continuous or dichotomous scores, respectively, and for each unidimensional scale and subscales separately. Further, for IRT-based analyses standard error of the theta or the reliability coefficient of the estimated latent trait value should be calculated. In this review, internal consistency was reported for 11 measures in 14 studies. Cronbach's alpha or KR-20 with value 0.70 or above indicates acceptable internal consistency (Prinsen et al., 2018); in this review, the values of Cronbach's alpha ranged from low ($\alpha = 0.60$) to excellent ($\alpha = 0.95$), and the values of KR-20 were excellent (range from 0.91 to 0.96) In addition, if determined, item-total correlations or split-half analysis ranged from low to very strong ($r = 0.25$–.87). Finally, in the use of Attribute reliability, strong latent classification reliabilities were reported (ranging from 0.82 to 0.99), and in Rasch modeling (IRT), a strong person reliability ($r = 0.69$) was determined.

Interrater reliability was determined for four measures reporting either on Kappa coefficient ($\kappa = 0.95$; Jitendra et al., 2005) or correlation coefficients (for three measures; overall range from 0.81 to 1.00). *Test-retest reliability* was reported for three measures in four studies: overall range of the correlations was from 0.52 to 0.95; moreover, the time interval between administrations in these studies was about two weeks, which is in line with the COSMIN guidelines (Mokkink et al., 2018). Finally, for MCS (Akgul & Kahveci, 2016), excellent *intra-rater reliability* ($r = 0.88$–.96) was reported. *Measurement error* was not reported in any of the studies in this review.

3.5. Summary of the results

The summary of measurement properties evaluated for each measure are presented in Table 5. Overall, the strongest validity evidence in terms of content validity, structural validity and hypothesis testing was reported for BNPT (Olkun et al., 2016) and Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011), both targeting children with or without MLD. In addition, validity evidence in terms of content validity and structural validity was reported for Division of fractions (Ketterlin-Geller et al., 2013), MCS (Akgul & Kahveci, 2016), and Representing Rational Numbers (Ketterlin-Geller et al., 2019). Further, cross-cultural validity was reported only for South Africa Numeracy test (Kivilu, 2010); thus, more research is needed to enable researchers and teachers to be confident in these measures and trust that they work similarly between different groups (e.g., age groups and sex). Of note, evidence for content validity, the most important measurement property, was reported only for five out of 18 measures.

The strongest evidence on reliability (internal consistency, test-retest, intrarater, and interrater reliabilities) for the included measures was retrieved for CBM-WPS-Fluency (Jitendra et al., 2005, 2014), Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011), MCS (Akgul & Kahveci, 2016), and NAS (Looveer & Mulligan, 2009). Evidence on reliability was not available for five measures (see Table 5, "NR"). Furthermore, even though measurement error is an important measurement property, it was not reported to any of the included measures; thus, evidence on the capability of the included measures to precise measurement was not received.

The results from Qualsyst (Kmet et al., 2004) analysis showed that despite of the lacking information on detailed measurement properties, the overall methodological quality of the included studies was strong, and no bias of risk emerged (Fig. 2); in separate studies, it varied from adequate to strong. The mean methodological quality percentage score was 87%. The evidence of overall quality of the included studies was convincing as none of the studies fell into a category of poor methodological quality. Instead, most of the studies showed either strong or good methodological quality. Appendix C introduces the exact Qualsyst ratings of the included studies.

4. Discussion

The aim of this systematic review was to report on the characteristics of functional numeracy (FN) measures developed for teachers' use at the elementary school level and evaluate their validity and reliability evidence. Specifically, this study sought to investigate the measurement properties of existing FN measures (curriculum-based measures [CBMs], screening, and diagnostic measures concurrently) published since the year 1995 for children aged 9–12 years who may need additional support for mathematical learning difficulties. This review identified 21 individual studies relating to 18 FN measures that met the eligibility criteria and were included in this review. Of these measures four were curriculum-based measures, 12 screening measures, and two diagnostic assessments.

The measures studied in this review assessed a wide range of FN skills with multiple variations of items, time limitations, and purposes. The FN skills that are measured in children aged 9 to 12 years included symbolic number sense; counting skills, basic skills in arithmetic, and understanding of mathematical relationships with tasks using numbers with 1–3 digits, decimals and fractions. Arithmetic tasks included addition, subtraction, multiplication and division tasks. Furthermore, tasks in algebra, geometry and measurement were present in some measures. The measures used items presented with number symbols and/or word problems.

4.1. Content validity of the included fn measures

For a measure to be rated as having good content validity, the measure should be relevant, comprehensive, and comprehensible

Table 4
Summary of the reliability properties reported in the included articles.

Measure	*Refs.*	*Reliability Internal consistency Method*	*Results*	*Reliability Method*	*Results*
ANWT Arabic Number-Writing Task	Moura et al. (2015)	KR-20 Split-half analysis	KR-20 = 0.91 $r = 0.94$	NR	NR
BNPT Basic Number Processing Test	Olkun et al. (2016)	KR-20 Cronbach's alpha	KR-20 = 0.72–0.96 (mental number line tasks) α = 0.69–0.79 (canonic dot counting, symbolic number comparison tasks)	NR	NR
CBM – Math Computation Curriculum-based Measurement of Math Computation	Shapiro et al. (2006)	NR	NR	NR	NR
CBM – Math Concept Application Curriculum-based Measurement of Math Concept Application	Shapiro et al. (2006)	NR	NR	NR	NR
CBM-WPS-Fluency Curriculum-Based Measurement of Word Problem Solving	Jitendra et al. (2005)	Cronbach's alpha	Separately for each 8 probes: range α = 0.60 to α = 0.75; for scores aggregated across each pair of odd and even probes: range α = 0.76 to α = 0.83.	Interrater: Kappa coefficient	κ = 0.95
	Jitendra et al. (2014)	Cronbach's alpha	α = 0.68 (T1), α = 0.67 (T2), α = 0.71 (T3)	Test-retest: Correlation	$r = 0.65$ (between T1 and T2); $r = 0.52$ (between T1 and T3), $r = 0.70$ (between T2 and T3); all correlations statistically significant ($p < .01$)
DAA Diagnostic Arithmetics Assessment	Kunina-Habenicht et al. (2017)	Attribute Reliability	The latent classification reliabilities: (1) addition/ subtraction: 0.94, and (2) multiplication/ division: 0.99, (3) modeling skills: 0.84, and (4) skills using measurement units: 0.82	NR	NR
Division of fractions	Ketterlin-Geller et al. (2013)	IRT	Person reliability = 0.69	NR	NR
DMA Diagnostic Mathematics Assessment	Kunina-Habenicht et al. (2017)	NR	NR	NR	NR
ERT Eggenberger RechenTest	Gebhardt et al. (2014)	Cronbach's alpha	Four subscales: α = 0.92 (Basic arithmetic skills), α = 0.86 (Number series), α = 0.72 (Word problems), α = 0.85 (Writing numbers)	NR	NR
IPAM Indicadores de Progreso de Aprendizaje en Matemáticas/ Indicators of Basic early Math Skills	de León et al. (2020)	Item-total correlations	Between items and total correlations ranges per form: Fall = 0.25–0.81; Winter = 0.44–0.87; Spring = 0.42–0.87.	NR	NR
	de León et al. (2021)	Item-total correlations	Between items and total correlations ranges per form: Fall = 0.38–0.85; Winter = 0.39–0.84; Spring = 0.40–0.82.	NR	NR
Knowledge of Mathematical Equivalence	Matthews et al. (2012)	Cronbach's alpha	Two forms: α = 0.93 (Form 1), α = 0.94 (Form 2)	Interrater: Independent rater coded 20% of the sample	Form 1 agreement 0.99 (range 0.96 to 1.00), Form 2 agreement 0.97 (range 0.87 to 1.00).
	Rittle-Johnson et al. (2011)	Cronbach's alpha	Two forms: α = 0.94 (Form 1), α = 0.95 (Form 2)	Interrater: Independent rater coded 20% of the sample	Form 1 agreement 0.99 (range 0.96 to 1.00), Form

(*continued on next page*)

Table 4 (*continued*)

Measure	*Refs.*	*Reliability Internal consistency Method*	*Results*	*Reliability Method*	*Results*
					2 agreement 0.97 (range 0.87 to 1.00)
				Test-retest: Correlation between a subset of 28 items included in both the initial instrument (T1) and the shortened versions of the assessment (From 1, From 2) administered at T2.	Form 1: r(26) = 0.94 Form 2: r(26) = 0.95
MAP Measures of Academic Progress	Klingbeil et al. (2019)	NR	NR	NR	NR
M-CBMs Math Curriculum-Based Measurements	Strait et al. (2015)	NR	NR	Test-retest: Pearson product moment correlation	M-CBM fluency scores ranged from 0.49 to 0.75 (M = 0.66) Accuracy scores ranged from 0.61 to 0.75 (M = 0.68)
	Strait et al. (2018)	NR	NR	Test-retest	Mean fluency or accuracy scores from four M-CBMs produced more reliable test–retest estimates (e.g., r = 0.86–.88) compared to individual scores (e.g., r = 0.55–.73).
MCS Mathematics Creativity Scale	Akgul and Kahveci (2016)	Cronbach's alpha Item-total correlation	α =0.80 between items and total scores varied from 0.49 to 0.72, and between items from 0.33 to 0.82	Interrater: Two independent raters Test-retest: Pearson product-moment correlation Intra-rater reliability	CC per item from 0.81 to 0.91 Group of 40 people, two months apart CC per item from 0.58 to 0.74 CC per item from 0.88 to 0.96
NAS Numeracy Achievement Scale	Looveer and Mulligan (2009)	Cronbach's alpha	calculated to each year level; values ranged from 0.73 to 0.90	Interrater: Project manager and university researcher (second author)	r = 0.92
NSCT Nonsymbolic and Symbolic Comparison Tasks	Castro et al. (2017)	NR	NR	Test-retest: Correlations	Significant correlations (r = 0.66–0.86; $p < .001$) for adjRT and EMs for nonsymbolic and symbolic tasks under each experimental condition (global, small ratio, large ratio, close distance, far distance). Low or no statistically significant correlations between the accuracy measures.
Representing Rational Numbers	Ketterlin-Geller et al. (2019)	NR	NR	NR	NR
South Africa Numeracy Test	Kivilu (2010)	KR-20	KR-20 = 0.96	NR	NR

Note. Measurement error has not been included in this table as no study reported on measurement error; KR-20 = Kuder–Richardson Formula 20, CC = correlation coefficient, NR = not reported.

(Terwee et al., 2018). Although content validity is the most important measurement property according to COSMIN taxonomy (Terwee et al., 2018), only 5 of the 18 included measures were evaluated for content validity; relevance of the final version of measures was mostly evaluated by asking experts. In MCS (Akgul & Kahveci, 2016) also mathematics teachers evaluated the relevance. The two most thoroughly developed measures with respect to content validity were Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011) and Representing Rational Numbers (Ketterlin-Geller et al., 2019). Importantly, these two measures were developed also through feedback from students, whilst the other sixteen measures incorporated no student feedback even though the age of students would have allowed their participation. Yet, it is crucial that the items and wordings in measures are clear and comprehensible, thus students are important informants to ensure the content validity of a measure. In addition, when using CBMs, the match between educational context and test content should be considered (Foegen et al., 2007) and the development of CBMs should

Table 5
Summary of measurement properties evaluated for each measure.

Measurement Instrument	Content Validity	Structural Validity (factor analysis - CTT; dimensionality -IRT [Rasch analysis])	Hypothesis Testing (Hypothesis about other instruments and relation)	Cross-cultural Validity (Differential item functioning – IRT; Multi-group Confirmatory Factor Analysis - CTT)	Criterion Validity (gold standard or short form)	Internal Consistency	Reliability (test-retest, intra-rater, inter-rater; ICC or Kappa)	Measurement Error (SDC, LoA, MIC, MDC)
ANWT	NR	NR	Yes	NA	NA	Yes	NR	NR
BNPT	Yes	Yes	Yes	NA	NA	Yes	NR	NR
CBM – Math Computation	NR	NR	Yes	NA	NA	NR	NR	NR
CBM – Math Concept Application	NR	NR	Yes	NA	NA	NR	NR	NR
CBM-WPS-Fluency	NR	NR	Yes	NA	NA	Yes	Yes	NR
DAA	NR	Yes	Yes	NA	NA	Yes	NR	NR
Division of fractions	Yes	Yes	NR	NA	NA	Yes	NR	NR
DMA	NR	Yes	NR	NA	NA	NR	NR	NR
ERT (Adapted)	NR	Yes	NR	NA	NA	Yes	NR	NR
IPAM	NR	NR	Yes	NA	NA	Yes	NR	NR
Knowledge of Mathematical Equivalence	Yes	Yes	Yes	NA	NA	Yes	Yes	NR
MAP	NR	NR	Yes	NA	NA	NR	NR	NR
M-CBMs	NR	NR	Yes	NA	NA	NR	Yes	NR
MCS	Yes	Yes	NR	NA	NA	Yes	Yes	NR
NAS	NR	Yes	NR	NA	NA	Yes	Yes	NR
NSCT	NR	Yes	Yes	NA	NA	NR	Yes	NR
Representing Rational Numbers	Yes	Yes	NR	NA	NA	NR	NR	NR
South Africa Numeracy Test	NR	Yes	NR	Yes	NA	Yes	NR	NR

Note. NR = not reported; NA = not applicable; CTT = classic test theory; IRT = item response theory; ICC = intraclass correlation coefficient; SDC = smallest detectable change; LoA = limits of agreement; MIC = minimal important change; MDC = minimal detectable change).

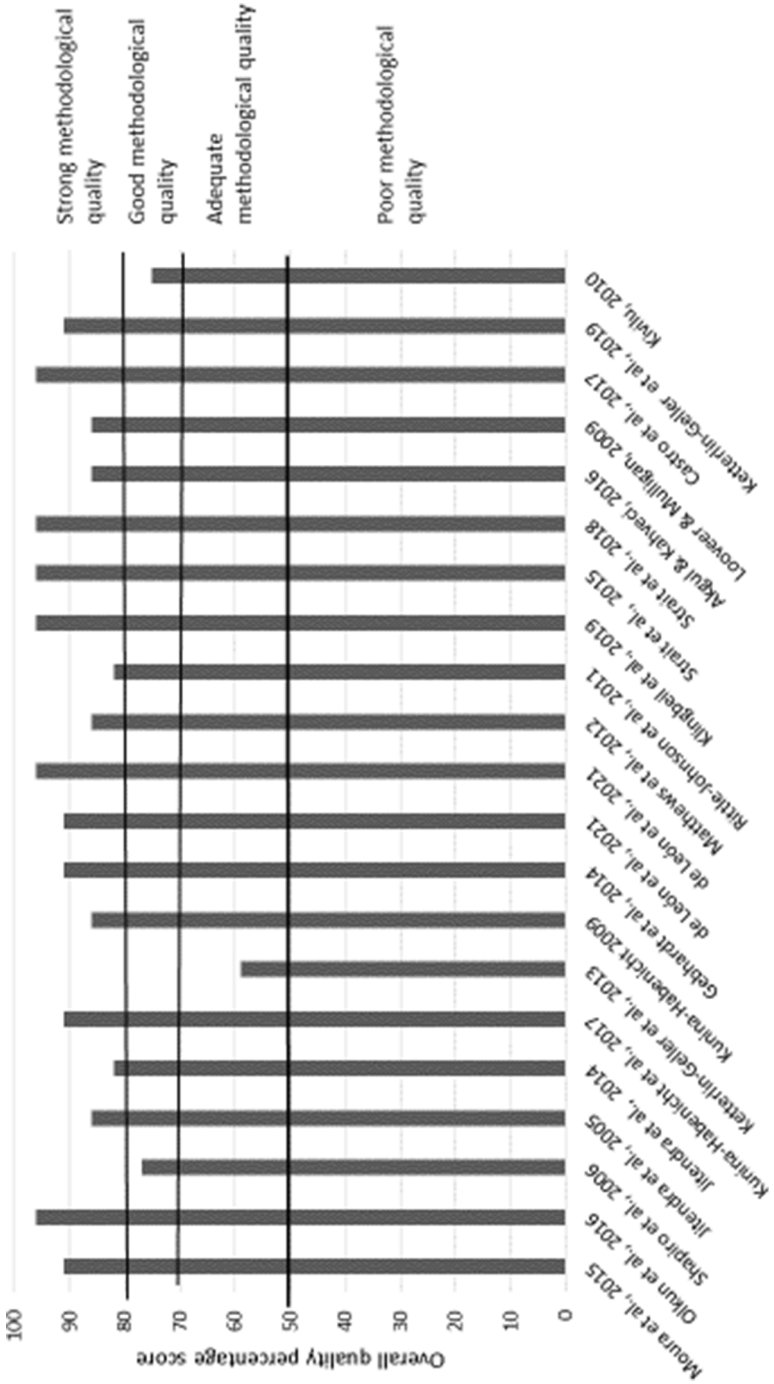

Fig. 2. Overall methodological quality of the included studies according to Qualsyst (Kmet et al., 2004).

be ongoing to improve instrumentation (Christ et al., 2008).

Only the data concerning the criterion "asking professionals about the relevance of the measure's items" was clearly reported for the measures included in this review. Content validity criteria related to comprehensiveness and comprehensibility were not reported, reducing the amount of the content validity data. For most measures (13 of 18) content validity was not reported at all, making it difficult to draw any substantive conclusions about the content validity of measures based on these studies alone. Therefore, findings from this review indicate that evidence of the quality of content validity of FN measures for 9–12-year-old children is still uncertain and further investigation is required to enable a conclusive assessment of available FN measures 9–12-year-old children.

4.2. Construct validity of the included fn measures

Construct validity seeks to determine the degree to which the scores of a measure are consistent with hypotheses based on the assumption that a measure validly measures the construct under investigation (Mokkink et al., 2018). This domain of validity encompasses structural validity, cross-cultural validity, and hypothesis testing for construct validity.

In terms of structural validity, the COSMIN criteria (Mokkink et al., 2018) favours confirmatory factor analysis (CFA) over exploratory factor analysis (EFA) and principal component analysis (PCA). The CFA and PCA of Knowledge of Mathematical Equivalence (Rittle-Johnson et al., 2011) were performed using the same data, making it possible to yield over optimistic model fit indices and parameter estimates (Fokkema & Greiff, 2017). In some studies, the sample sizes were small, and the reporting was lacking details such as missing data and rotation method.

The second aspect of construct validity, hypothesis testing, includes many aspects (e.g., convergent and discriminant validity) which may increase the likelihood for hypotheses testing to be evaluated. That was the case in this review also, studies included reported hypothesis testing for 10 out of 18 measures, mostly about convergent validity. Finally, the third aspect of construct validity, cross-cultural validity, was reported only for South Africa Numeracy Test (Kivilu, 2010) showing that the items in the measure were not biased since there were no sex differences in item difficulty.

4.3. Reliability of the included fn measures

Reliability, as a domain of measurement according to COSMIN, contains three measurement properties: internal consistency, reliability, and measurement error (Mokkink et al., 2018). Internal consistency with the use of Cronbach's alpha was the most reported property, also split-half method, item-total correlations, and KR-20 were used. Worryingly, report on internal consistency was lacking from nine measures; further, for some measures (CBM Math Computation, CBM Math Concept Application, and MAP) internal consistency coefficients were reported from earlier studies, thus marked in this study with not reported (NR). This was done because internal consistency statistics should be calculated separately for every scale and subscale (Mokkink et al., 2018), and in the target population intended, i.e., in each data, because sample size has an effect on the Cronbach's alpha (Bujang et al., 2018). Further, test-retest reliability, interrater and intrarater reliabilities were reported as other reliability properties. The recommendation for time interval for test-retest is two weeks (Mokkink et al., 2018), but it is not always possible to follow that recommendation. For example, the test-retest time interval for MCS (Akgul & Kahveci, 2016) was two months instead of the recommended two-week period.

Although most studies reported on either internal consistency or reliability, none of the included studies reported on all three measurement properties for the reliability domain (Mokkink et al., 2018). Surprisingly, measurement error was not evaluated in any of the studies included in this review. This is very significant deficiency in the reliability of the included measures: for measurement error, Mokkink et al. (2018) note that a reliable measure performs its measurements precisely and does not include much measurement error, which reflects the systematic and random error of a respondent's score that is not due to true changes in the construct measured. Furthermore, cross-cultural validity was evaluated only for South Africa Numeracy test (Kivilu, 2010).

To sum up, Knowledge of Mathematical Equivalence (Matthews et al., 2012; Rittle-Johnson et al., 2011) was the only measure of which data about five; BNPT (Olkun et al., 2016) and MCS (Akgul & Kahveci, 2016) data about four of the eight measurement properties was available; and data for the rest of the measures were ranging from one measurement property to three. Further, despite its practical relevance, none of the measures reported measurement error. This is a significant deficiency, as measures with low measurement error are better able to detect changes sensitively and help professionals decide when and how to make interventions. Although the results from the Qualsyst (Kmet et al., 2004) analysis revealed that the overall methodological quality of the included studies was strong and no bias of risk emerged, further development of the measures in this review and new ones under development is needed to ensure the quality of the measures intended for teachers' use.

4.4. Limitations

This review has some limitations. First, the data in this review consisted only of peer-review original articles, thus, information about measurement properties of the measures was not reached e.g., from manuals. Thus, additional reviews with somewhat different predefined criteria are needed to capture a deeper and wider understanding of the psychometrics of the measures that are used to assess students' mathematical skills. Second, this review did not consider all nine measurement properties from the COSMIN taxonomy (Mokkink et al., 2018) as responsiveness excluded from the scope for this review. The decision was taken because evaluating responsiveness would have required a review of all the studies that have used the included measures as an outcome measure and would have required an additional and different search strategy. Third, interpretability (i.e., the degree to which one can assign qualitative meaning to a measure's quantitative scores or change in scores) and feasibility (i.e., the ease of application of the measure in its

intended context of use), though essential when considering and recommending the most suitable measure, were outside the scope of this review as they are not measurement properties (Prinsen et al., 2018).

4.5. Directions for future research

For researchers who want to comprehensively understand the measurement properties of all current FN measures used by teachers for 9–12-year-old children, this systematic review highlights the need for further validation studies of the measures included in this review. Especially, as content validity is considered the most important measurement property, attention should be focused reporting it bearing in mind that each item must be relevant and comprehensible concerning the construct of interest and the target population (Terwee et al., 2018). In addition, to ensure the measure's ability to validly measure the construct it is supposed to measure, its structural validity needs to be studied. Hence, at the very least, developers of new measures should comprehensively report on the content validity and internal structure (structural validity, internal consistency, and measurement invariance [when measures are adapted for different populations]). Moreover, given the gaps identified related to the measurement properties of current measures, this systematic review illustrates the need for the future development of high-quality FN measures. Such measures should be carefully developed to ensure both content and construct validity. Indeed, to ensure good measurement properties, appropriate statistics for each measurement property should be calculated and reported, in accordance with the criteria for good measurement properties (Prinsen et al., 2018).

5. Conclusion

This systematic review evaluated the measurement properties of 18 FN measures using the COSMIN guidelines. Evidence regarding measurement properties was limited, with no measures being evaluated for measurement error; criterion validity was not evaluated because there was no gold standard and short and long versions of measures were not compared. Further validation is recommendable for all measures to determine if they are of sufficiently high quality to assess FN skills in children aged 9–12 years. This is important as teachers are expected to be able to differentiate their teaching according to their students' needs to ensure the development of every student's FN skills. Furthermore, we want to address that the peer-reviewing process in developing and validating new measures is irreplaceable. In many cases, psychometric information of a measure can be found in manuals, but they lack the peer reviewing process, which we think is very important to ensure the quality of the measures used in assessing children. In addition, it would be desirable that publishers and authors would make the development and psychometric data on measures freely available to researchers who evaluate measures.

Declaration of Competing Interest

None.

Acknowledgment

We would like to acknowledge M.Ed. Annukka Relander (University of Helsinki), whose master's thesis laid a foundation for this research article.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijer.2023.102172.

References

Akgul, S., & Kahveci, N. G. (2016). A Study on the development of a Mathematics Creativity Scale. *Eurasian Journal of Educational Research, 62*, 57–76. https://doi.org/10.14689/ejer.2016.62.5

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. https://doi.org/10.1080/1350293x.2014.996424

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's Alpha test: A simple guide for researchers. *Malaysian Journal of Medical Sciences, 25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Castro, D., Estevez, N., Gomez, D., & Dartnell, P. R. (2017). Reliability and validity of nonsymbolic and symbolic comparison tasks in school-aged children. *The Spanish Journal of Psychology, 20*, E75. https://doi.org/10.1017/sjp.2017.68

Christ, T., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research. curriculum-based measurement of math computation. *Assessment of Effective Intervention, 33*(4), 198–205. https://doi.org/10.1177/1534508407313480

de León, S. C., Jiménez, J. E., García, E., & Gutiérrez, N. (2021). Identification of Spanish third graders at risk of math problems: Usefulness of number sense based screening measures. *Psychology in the Schools, 58*(7), 1416–1431. https://doi.org/10.1002/pits.22525

de León, S. C., Jiménez, J. E., García, E., Gutiérrez, N., & Gil, V. (2020). Universal screening in mathematics for Spanish students in first grade. *Learning Disability Quarterly, 44*(2), 123–135. https://doi.org/10.1177/0731948720903273

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of literature. *The Journal of Special Education, 41*(2), 121–139. https://doi.org/10.1177/00224669070410020101

Fokkema, G., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it. *European Journal of Psychological Assessment, 33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460. *: Official Organ of the European Association of Psychological Assessment.*

Geary, D. C. (2011). Consequences, characteristics, and causes of mathematical learning disabilities and persistent low achievement in mathematics. *Journal of Developmental & Behavioral Pediatrics, 32*(3), 250–263. https://doi.org/10.1097/DBP.0b013e318209edef

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PloS one, 8*(1), e54651. https://doi.org/10.1371/journal.pone.0054651

Gebhardt, M., Zehner, F., & Hessels, M. (2014). Basic arithmetical skills of students with learning disabilities in the secondary special schools: An exploratory study covering fifth to ninth grade. *Frontline Learning Research*. https://doi.org/10.14786/flr.v2i1.73

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K. S., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78*(4), 423–445. https://doi.org/10.1177/001440291207800403

Hakkarainen, A. M., Holopainen, L. K., & Savolainen, H. K. (2015). A five-year follow-up on the role of educational support in preventing dropout from upper secondary education in Finland. *Journal of Learning Disabilities, 48*(4), 408–421. https://doi.org/10.1177/0022219413507603

Jitendra, A. K., Dupuis, D. N., & Zaslofsky, A. F. (2014). Curriculum-based measurement and standards-based mathematics: Monitoring the arithmetic word problem-solving performance of third-grade students at risk for mathematics difficulties. *Learning Disability Quarterly, 37*(4), 241–251. https://doi.org/10.1177/0731948713516766

Jitendra, A. K., Sczesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review, 34*(3), 358–371. https://doi.org/10.1080/02796015.2005.12086291

Kaufman, A. S. (2005). *Essentials of KABC-II assessment*. John Wiley & Sons.

Ketterlin-Geller, L. R., Shivraj, P., Basaraba, D., & Yovanoff, P. (2019). Considerations for using mathematical learning progressions to design diagnostic assessments. *Measurement: Interdisciplinary Research and Perspectives, 17*(1), 1–22. https://doi.org/10.1080/15366367.2018.1479087

Ketterlin-Geller, L. R., Yovanoff, P., Jung, E., Liu, K., & Geller, J. (2013). Construct definition using cognitively based evidence: A framework for practice. *Educational Assessment, 18*(2), 122–146. https://doi.org/10.1080/10627197.2013.790207

Kivilu, J. M. (2010). Determination of Differential Bundle Functioning (DBF) of numeracy and literacy tests administered to grade 3 learners in South Africa. *South African Journal of Psychology, 40*(3), 308–317. https://doi.org/10.1177/008124631004000309

Klingbeil, D. A., Maurice, S. A., Van Normann, E. R., Nelson, P. M., Birr, C., Hanrahan, A. R., et al. (2019). Improving mathematics screening in middle school. *School Psychology Review, 48*(4), 383–398. https://doi.org/10.17105/SPR-2018-0084.V48-4

Kmet, L. M., Lee, R. C., & Cook, L. S ((2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Alberta Heritage Foundation for Medical Research (AHFMR). AHFMR - HTA Initiative #13 http://www.ihe.ca/advanced-search/standard-quality-assessment-criteria-for-evaluating-primary-research-papers-from-a-variety-of-fields.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*(2–3), 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2017). Incremental validity of multidimensional proficiency scores from diagnostic classification models: An illustration for elementary school mathematics. *International Journal of Testing, 17*(4), 277–301. https://doi.org/10.1080/15305058.2017.1291517

Lembke, E., & Stecker, P.M. (2007). *Curriculum-based measurement in mathematics: An evidence-based formative assessment procedure*. RMC Research Corporation, Center on Instruction. https://files.eric.ed.gov/fulltext/ED521574.pdf.

Looveer, J., & Mulligan, J. (2009). The efficacy of link items in the construction of a numeracy achievement scale-from kindergarten to year 6. *Journal of Applied Measurement, 10*(3), 247–265. PMID: 19671988.

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education, 43*(3), 316–350. https://doi.org/10.5951/jresematheduc.43.3.0316

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., & Bouter, L. M. (2018a). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*, 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L.B., Prinsen, C.A.C., Patrick, D.L., Alonso, J., Bouter, L.M., de Vet, H.C.W. et al (2018). *COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)*. https://cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf.

Moura, R., Lopes-Silva, J. B., Vieira, L. R., Paiva, G. M., de Almeida Prado, A. C., Wood, G., et al. (2015). From "five" to 5 for 5 min: Arabic number transcoding as a short, specific, and sensitive screening tool for mathematics learning difficulties. *Archives of Clinical Neuropsychology, 30*(1), 88–98. https://doi.org/10.1093/arclin/acu071, 2015.

Olkun, S., Altun, A., Gocer Sahin, S., & 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.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., & Mulrow, C. D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71 *(Clinical research ed.)*.

Park, S., & Nelson, G. (2022). The quality of outcome measure reporting in early numeracy intervention studies. *Psychology in the Schools, 59*, 1721–1736. https://doi.org/10.1002/pits.22726

Powell, S. R., Mason, E. N., Bos, S. E., Hirt, S., Ketterlin-Geller, L. R., & Lembke, E. S. (2021). A systematic review of mathematics interventions for middle-school students experiencing mathematics difficulty. *Learning Disabilities Research & Practice, 36*(4), 295–329. https://doi.org/10.1111/ldrp.12263

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., et al. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1147–1157. https://doi.org/10.1007/s11136-018-1798-3

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology, 103*(1), 85–104. https://psycnet.apa.org/doi/10.1037/a0021334.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*(1), 19–35. https://psycnet.apa.org/doi/10.1177/0734282905285237.

Strait, G. G., Smith, B. H., & McQuillin, S. D. (2018). Aggregated randomly generated math curriculum-based measurements for middle school students: Reliability, predictive validity, and cut score precision. *Assessment for Effective Intervention, 44*(1), 58–64. https://doi.org/10.1177/1534508418761231

Strait, G. G., Smith, B. H., Pender, C., Malone, P. S., Roberts, J., & Hall, J. D. (2015). The reliability of randomly generated math curriculum-based measurements. *Assessment for Effective Intervention, 40*(4), 247–253. https://doi.org/10.1177/1534508415588075

Terwee, C.B., Prinsen, C.A.C., Chiarotto, A., de Vet, H.C.W., Bouter, L.M., Alonso, J. et al (2018). *COSMIN methodology for assessing the content validity of PROMs*. https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf.

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., et al. (2018b). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*, 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Van Norman, E. R., Nelson, P. M., Klingbeil, D. A., Cormier, D. C., & Lekwa, A. J. (2018). Gated screening frameworks for academic concerns: The influence of redundant information on diagnostic accuracy outcomes. *Contemporary School Psychology, 23*(2), 152–162. https://doi.org/10.1007/s40688-018-0183-0

Vanbinst, K., Ghesquiere, 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. https://doi.org/10.1016/j.ridd.2014.06.023

Wechsler, D. (2014). *WISC-V: Technical and interpretive manual*. Pearson.