Measuring the Four Cognitive Strands of Mathematical Proficiency in Primary School: A Systematic Review of Measurement Properties

Sovia Khotami*, Gunawan Suryoputro, Ai Fatimah Nur Fuad

Universitas Muhammadiyah Prof. DR. HAMKA, Jakarta, Indonesia
*Corresponding author, email: soviakhotami@uhamka.ac.id

https://doi.org/10.17977/um065.v7.i2.2027.11

Article history
Submitted: 19 June 2026
Revised 13 August 2026
Accepted: 25 August 2026
Published: 27 August 2026

Keywords
Adaptive reasoning
Instrument validation
Mathematical proficiency
Primary education
Rasch measurement

Abstract
Mathematical proficiency in the primary years comprises four interwoven cognitive strands conceptual understanding, procedural fluency, strategic competence, and adaptive reasoning yet how well existing instruments measure them is unclear. This systematic review examined how these strands are operationalised in primary-school students and appraised the psychometric quality of the instruments used. Following the PRISMA 2020 statement and COSMIN methodology, Scopus and Web of Science were searched for English-language articles published between 2010 and 2026, yielding 40 studies. Conceptual understanding and procedural fluency were assessed far more often than strategic competence and adaptive reasoning; only one study operationalised all four strands, and its factor analysis recovered just two. Structural validity dominated by the Rasch/item-response-theory family was the most frequently reported property, whereas inter-rater reliability, content validity, and measurement invariance were seldom examined, and responsiveness was not established in any study. The field is thus strong on internal structure but weak on score consistency, comparability, and sensitivity to change. We propose an evidence-based blueprint sufficient theory-aligned items per strand, rubric scoring with multiple raters, bifactor or many-facet modelling, and a full validity argument to guide the development of curriculum-aligned, psychometrically defensible instruments.

1. Introduction

Numeracy is a foundational capability that shapes learning, economic participation, and informed citizenship across the life course (Gal et al., 2020; Reder et al., 2020). In the primary years it is especially consequential: the mathematical competencies children build in the early grades form the platform on which all later mathematics rests (Björklund et al., 2020). Longitudinal evidence consistently shows that early numeracy and number sense predict mathematics achievement years later (Aunio et al., 2020; Jordan et al., 2022; Salminen et al., 2021), and that gaps present at school entry tend to widen rather than close (Chang, 2023; Ersan & Rodriguez, 2020). Conversely, weak early numeracy constrains later access to science- and technology-intensive pathways, making its timely, accurate identification a matter of both educational quality and equity (Aunio et al., 2020). Helping every child become mathematically proficient is therefore a central goal of primary education and an explicit policy priority in many systems, including Indonesia, where the Kurikulum Merdeka and the national Minimum Competency Assessment foreground numeracy and higher-order thinking (Hidayah et al., 2021; Purnomo et al., 2022; Rohmah et al., 2022; Tanudjaya & Doorman, 2020).

Crucially, mathematical proficiency is more than computation. The widely adopted framework of Kilpatrick et al. (2001) conceptualises it as five interwoven strands conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition of which the first four are cognitive and assessable through performance. Contemporary curricula echo this breadth, reframing mathematics learning around reasoning, problem solving, and the wider 21st-century competencies (Drake & Reid, 2020; Yayuk et al., 2020), and stressing that strategic and reasoning-oriented capabilities, not procedures alone, distinguish proficient learners (Copur-Gencturk & Doleck, 2021). These higher-order strands, however, are also the hardest to assess: they must be elicited through open, cognitively demanding tasks and are frequently proxied or displaced by easier-to-score computation, so that what is measured can drift from what is valued. Because the strands are mutually reinforcing, a balanced account of a child's proficiency requires valid evidence on all four cognitive strands rather than on computation alone.

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

This ambition has sharpened the demand for high-quality assessment. The disruption of the COVID-19 pandemic eroded foundational numeracy and widened achievement gaps (Conto et al., 2021; Gore et al., 2021), intensifying the need for instruments that can diagnose where children stand and monitor recovery. At the same time, assessment is rapidly digitalising, with computer-based, game-based, and adaptive formats and, most recently, artificial-intelligence-mediated tools entering primary mathematics (Alabdulaziz, 2021; Benavides-Varela et al., 2020; Kasneci et al., 2023; Wang et al., 2022; Zhang et al., 2025). These developments raise, rather than lower, the bar for measurement: a score is only as useful as the validity and reliability of the instrument that produced it (Messick, 1995). For the four cognitive strands specifically, this means instruments whose internal structure recovers the intended constructs, whose scores are consistent across items and where rubric scoring is used across raters, and whose content is demonstrably aligned with the proficiency framework.

Yet whether existing instruments meet these requirements is unclear, because the measurement of mathematical proficiency in primary school has not been systematically and critically appraised. Prior syntheses have been valuable but partial: they have catalogued early-numeracy screeners used within intervention studies (Akther & Powell, 2025) or summarised the predictive validity of curriculum-based measures (Park et al., 2024), rather than examining how the full four-strand construct is operationalised at the item level and how strong the underlying psychometric evidence is. Methodological advances now make such an appraisal feasible: dimensionality can be evaluated with bifactor and higher-order models that respect the interdependence of mathematical abilities (Pokropek et al., 2022), and consensus standards such as COSMIN provide a transparent basis for grading measurement-property evidence and reporting reviews of measurement instruments (Elsman et al., 2024; Gagnier et al., 2021; Mokkink et al., 2024). Applying this lens to primary-mathematics assessment would reveal which strands are well measured, which are neglected, and where the evidence base is fragile information that is essential both for selecting instruments and for designing new ones. Without it, instrument choice rests on face validity, and inferences about children's proficiency together with the instructional and policy decisions that follow may stand on unstable foundations.

The present study addresses this gap through a systematic review of the measurement properties of instruments assessing mathematical proficiency operationalised as Kilpatrick's four cognitive strands in primary-school students. Synthesising the evidence is intended to inform the evidence-based development of robust, curriculum-aligned instruments, an aim of direct relevance to assessment reform in contexts such as Indonesia. Three questions guide the review. The first asks what instruments are used to measure mathematical proficiency, or its cognitive strands, in primary-school students, and what formats and scoring schemes these instruments employ. The second asks how each of the four cognitive strands is operationalised at the item level, and which strands are most and least represented across the corpus. The third asks what measurement properties are reported for these instruments and how methodologically sound that evidence is. Taken together, the three questions move from describing the available instruments, to examining the fidelity with which they represent the four-strand construct, to appraising the strength of the psychometric evidence that supports their use.

2. Method

2.1. Review Design and Reporting

This study is a systematic review of measurement properties that synthesises how mathematical proficiency and its four cognitive strands defined by Kilpatrick et al. (2001): conceptual understanding, procedural fluency, strategic competence, and adaptive reasoning is operationalised and measured in primary-school students. The review followed the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) methodology for systematic reviews of outcome measurement instruments (Prinsen et al., 2018) and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al., 2021). A protocol specifying the objectives, eligibility criteria, search strategy, and analysis plan was defined a priori following PRISMA-P (Moher et al., 2015) and is available from the authors. Because the included instruments are performance-based achievement tests rather than patient-reported measures, COSMIN was used as the appraisal framework while construct validity was interpreted within the unified validity perspective of Messick (1995) and the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

2.2. Eligibility Criteria

Eligibility followed a Construct–Population–Instrument frame. Studies were included if they (i) measured mathematical proficiency or at least one of its four cognitive strands, or a directly mappable construct (e.g., number sense, numeracy, mathematical reasoning, or fraction understanding); (ii) sampled primary-school students (grades 1–6; ages 6–12), with multi-grade samples eligible when at least 50% of participants fell within this range; (iii) employed a mathematics achievement test or task-based assessment; and (iv) developed or evaluated at least one measurement property content validity, structural validity, internal consistency, reliability, measurement invariance, criterion validity, hypotheses testing, or responsiveness or reported

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

classical item analysis, Rasch, or item-response-theory modelling. Intervention and media-development studies that merely used a test without reporting measurement evidence, and constructs outside the four cognitive strands (e.g., affect, creativity, computational thinking), were excluded. Only peer-reviewed journal articles and reviews published in English between 2010 and 2026 were considered. The full criteria are summarised in Table 1.

Table 1. Inclusion and Exclusion Criteria

Parameter	Inclusion	Exclusion
Construct	Mathematical proficiency or ≥1 of the four Kilpatrick cognitive strands; mappable constructs (number sense, numeracy, reasoning, fractions)	Affective constructs; creativity/originality; computational thinking; non-curricular neurocognitive measures
Population	Primary grades 1–6 (ages 6–12); multi-grade samples with ≥50% in range	Preschool, secondary, undergraduate, teachers; exclusively clinical samples
Instrument	Mathematics achievement test or task-based assessment	Self-report attitude scales; process observation without achievement scores
Measurement focus	Develops/evaluates ≥1 measurement property, or reports item analysis / Rasch / IRT	Intervention or media studies using a test without measurement evidence
Publication	Peer-reviewed article or review; English; 2010–2026	Conference abstracts, books/chapters, editorials; non-English

2.3. Information Sources and Search Strategy

Two databases were searched on April 2026: Scopus and the Web of Science Core Collection. These two databases were selected because they are the most comprehensive, curated, multidisciplinary citation indexes for peer-reviewed research and are the sources most commonly used in systematic reviews of measurement instruments. The query combined three conceptual blocks with the Boolean operator AND: (A) Construct (mathematical-proficiency and strand terms), (B) Population (primary-school terms), and (C) Measurement (instrument terms intersected with a measurement-property filter adapted from Terwee et al., 2009). The 2010–2026 window was chosen to capture a full decade of psychometric work while remaining current. After applying language (English) and document-type (article, review) limits, the search returned 453 records in Scopus and 317 in Web of Science (Table 2). The complete Boolean strings executed in each database are reported below.

Scopus (TITLE-ABS-KEY): (TITLE-ABS-KEY ("mathematical proficiency" OR "mathematics proficiency" OR "mathematical competence" OR "conceptual understanding" OR "procedural fluency" OR "strategic competence" OR "adaptive reasoning" OR "mathematical reasoning" OR "mathematical problem solving" OR "mathematical literacy" OR "number sense" OR numeracy OR "fraction*") AND TITLE-ABS-KEY ("primary school" OR "elementary school" OR "primary education" OR "elementary education" OR "primary student*" OR "elementary student*" OR "primary pupil*" OR "elementary pupil*" OR "grade 1" OR "grade 2" OR "grade 3" OR "grade 4" OR "grade 5" OR "grade 6" OR "fifth grade*" OR "fourth grade*" OR "key stage 1" OR "key stage 2") AND TITLE-ABS-KEY ((instrument OR test OR assessment OR scale OR measure OR rubric OR questionnaire OR inventory) AND (valid* OR reliab* OR "measurement propert*" OR psychometric* OR "Rasch" OR "item response theory" OR "IRT" OR "factor analysis" OR "internal consistency" OR "Cronbach" OR "inter-rater" OR "interrater" OR "content validity" OR "construct validity" OR "criterion validity" OR "structural validity" OR responsiveness OR "item analysis" OR "difficulty index" OR "discrimination index" OR calibration OR "Aiken"))) AND PUBYEAR > 2009 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "re")) AND (LIMIT-TO (LANGUAGE , "English")).

Web of Science Core Collection (TS = Topic; years 2010–2026; document types Article or Review; language English): TS = (("mathematical proficiency" OR "mathematics proficiency" OR "mathematical competence" OR "conceptual understanding" OR "procedural fluency" OR "strategic competence" OR "adaptive reasoning" OR "mathematical reasoning" OR "mathematical problem solving" OR "mathematical literacy" OR "number sense" OR numeracy OR fraction*) AND ("primary school" OR "elementary school" OR "primary education" OR "elementary education" OR "primary student*" OR "elementary student*" OR "primary pupil*" OR "elementary pupil*" OR "grade 1" OR "grade 2" OR "grade 3" OR "grade 4" OR "grade 5" OR "grade 6" OR "fifth grade*" OR "fourth grade*" OR "key stage 1" OR "key stage 2") AND ((instrument OR test OR assessment OR scale OR measure OR rubric OR questionnaire OR inventory) AND (valid* OR reliab* OR "measurement propert*" OR psychometric* OR Rasch OR "item response theory" OR IRT OR "factor analysis" OR "internal consistency" OR Cronbach OR "inter-rater" OR interrater OR "content validity" OR "construct validity" OR "criterion validity" OR "structural validity" OR responsiveness OR "item analysis" OR "difficulty index" OR "discrimination index" OR calibration OR Aiken))).

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Table 2. Search Strategy Summary

Element	Detail
Databases	Scopus; Web of Science Core Collection
Date searched	April 2026
Block A Construct	"mathematical proficiency", "conceptual understanding", "procedural fluency", "strategic competence", "adaptive reasoning", "mathematical reasoning", "number sense", numeracy, fraction*
Block B Population	"primary school", "elementary school", "grade 1"–"grade 6", "key stage 1/2"
Block C Measurement	(instrument OR test OR assessment OR scale OR rubric) AND (validity OR reliability OR psychometric* OR Rasch OR "item response theory" OR "factor analysis" OR "item analysis" OR Aiken ...)
Limits	English; document type article or review; 2010–2026
Records retrieved	Scopus 453; Web of Science 317 (total 770)

2.4. Study Selection

The 770 records were exported and deduplicated by normalised digital object identifier and normalised title, removing 194 duplicates and leaving 576 unique records. Two reviewers independently screened titles and abstracts against the eligibility criteria; disagreements were resolved through discussion with a third senior reviewer, and inter-rater agreement was quantified with Cohen's κ (McHugh, 2012). Fifty-six reports were sought for full-text retrieval, of which seven could not be obtained. The remaining 49 full-text reports were assessed; nine were excluded (intervention/effectiveness or media development, n = 5; construct outside the four strands, n = 3; correlational design with the instrument as a secondary measure, n = 1), yielding 40 studies for synthesis. The selection process is depicted in Figure 1.

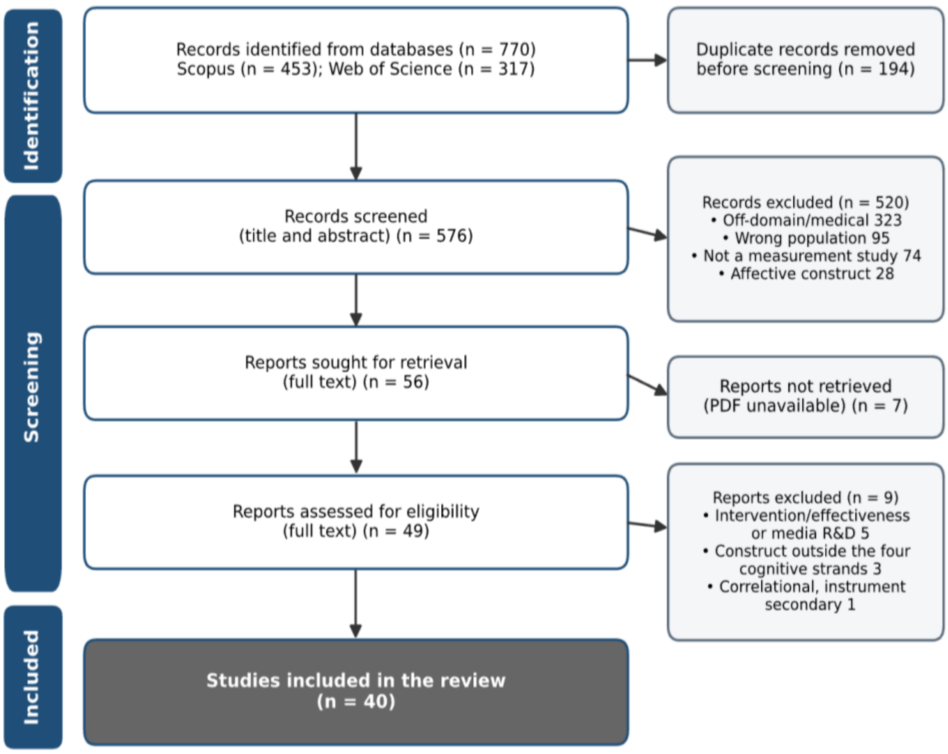

Figure 1. PRISMA 2020 Flow Diagram

2.5. Data Extraction and Quality Appraisal

Data were charted with a piloted extraction form capturing study and design characteristics, sample (size, grade, country), instrument (name, format, item count, scoring), the item-level operationalisation of each Kilpatrick strand, and all reported measurement properties with their statistics. Extraction was performed by one reviewer and verified by a second on a 20% random sample. Methodological quality was appraised per measurement property with the COSMIN Risk of Bias checklist (Mokkink et al., 2018), and results were rated against the COSMIN criteria for good measurement properties (Prinsen et al., 2018; Terwee et al., 2018), using the COSMIN taxonomy of measurement properties (Mokkink et al., 2010). Given the heterogeneity of constructs, instruments, and designs, the evidence was synthesised narratively rather than meta-analytically; findings were organised around (i) the instruments and formats used, (ii) the coverage and item-level operationalisation of

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

the four cognitive strands, and (iii) the breadth and methodological quality of the reported measurement properties.

3. Results and Discussion

3.1. Study Characteristics

The 40 included studies originated from 22 countries, four of which involved international collaboration. The United States contributed the largest share (n = 7), followed by China (n = 4) and by Indonesia and Thailand (n = 3 each); the remaining studies were spread thinly across Europe, Asia, Africa, and Latin America, most contributing only a single study (Figure 2b). This concentration of evidence in a few settings, coupled with the predominance of single-study countries, signals limited cross-context replication of any one instrument. Publication output spanned 2010–2026 and rose over the period 17 studies (43%) appeared in 2022 or later, with a marked peak of nine studies in 2019 and a further rise to seven in 2025 (Figure 2a) signalling a sustained and intensifying interest in the measurement of primary-school mathematical proficiency. Sample sizes varied by two orders of magnitude, from 91 pupils (Howard et al., 2017) to 18,405 (Hellstrand et al., 2024), and several studies drew on large, population-level samples (e.g., Bellini et al., 2019, N = 2,935; Nezhnov et al., 2015, N = 2,216; Gökçe et al., 2023, N = 2,034; Miao et al., 2022, N = 2,642).

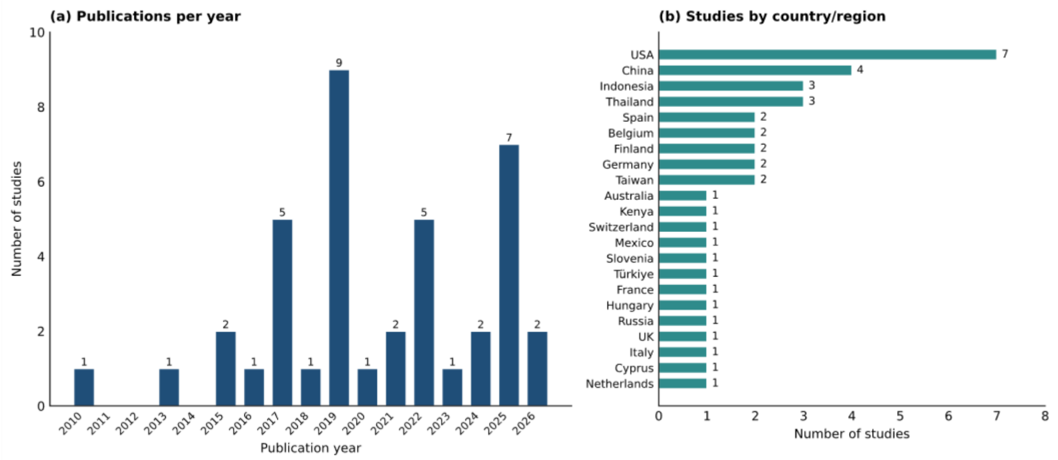

Figure 2. Temporal and Geographic Distribution

Across the corpus, every primary grade band was represented, although coverage was densest at the upper-primary level (grades 4–6) and comparatively thin in the earliest grades. Several instruments spanned the boundaries of primary schooling reaching down into kindergarten (Hirsch et al., 2018; Van Norman et al., 2026; Park et al., 2024) or up into lower secondary (Hellstrand et al., 2024, grades 3–9) and were retained under the pre-specified rule that at least half of the sampled participants fell within grades 1–6.

In terms of design, the corpus was dominated by instrument-development and cross-sectional validation studies; a smaller number were diagnostic-accuracy or longitudinal-predictive studies (e.g., Hirsch et al., 2018; Van Norman et al., 2026), and two were syntheses of existing measures (Akther & Powell, 2025; Park et al., 2024). Most instruments were researcher-developed rather than commercially standardised, which is consistent with the review's focus on the development and evaluation of measurement properties. The analytic approaches were dominated by the Rasch and item-response-theory (IRT) family, which appeared in the majority of studies that reported structural validity, whereas factor-analytic methods were largely confined to scales and questionnaires. Three studies were conducted in Indonesia (Ahyan et al., 2025; Diputra et al., 2025; Yanti et al., 2026) the national context most relevant to the present programme of work each developing a test aligned to the national numeracy or curriculum framework. Full characteristics of the 40 studies are presented in Table 3.

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Table 3. Characteristics of the 40 Included Studies

Study	Country	N	Grade	Instrument	Strand(s)	Key measurement evidence
Ahyan et al. (2025)	Indonesia	participants + 12 raters	primary	Numeracy test	—	Rasch/IRT, reliability, content
Akther & Powell (2025)	USA	61 studies	preschool–Grade 2	Early numeracy screeners/pretests	CU, PF	(synthesis)
Balt et al. (2020)	Germany	101	Grade 1	MARKO-D + Learning Progress Assessment	CU	Rasch/IRT, invariance
Bellini et al. (2019)	Switzerland	2,935	Grade 4	Mathematical Competence Scale (6 domains)	CU, PF	Rasch/IRT, content
Brankaer et al. (2017)	Belgium	1,588	Grades 1–6	SYMP Test (paper–pencil)	CU	reliability
Burte et al. (2019)	USA	132→88 (adults) + primary	primary 10–12 yr	Make-A-Dice Test	AR	reliability
Chaiarwut et al. (2025)	Thailand	330 (EFA) + 1,916 (CFA)	Grade 4–6	Executive Maths Problem-Solving (self-assess.)	SC, AR	α, EFA/CFA
Chen et al. (2017)	China	1,207	Grade 3–5	Number sense + rule-space model	CU	Rasch/IRT
de León et al. (2021)	Spain	236	Grade 3	M-CBM IPAM (number sense)	CU, PF, SC	EFA/CFA, AUC/ROC
Diputra et al. (2025)	Indonesia	220	Grade 5	Mathematical proficiency test + questionnaire	CU, PF, SC, AR	EFA/CFA
Díaz-Cárdenas et al. (2019)	Mexico	primary classes	primary	Fraction multiplication/division assessment	CU, PF	α, EFA/CFA, Rasch/IRT
Gökçe et al. (2023)	Türkiye	2,034	Grade 2–4	Number sense tests with anchor items	CU	invariance
Hakkarainen et al. (2025)	Finland	358	Grade 3, 5, 7	FUNA-DB dyscalculia screener	CU, PF	α, EFA/CFA, reliability, invariance
Hellstrand et al. (2024)	Finland	18,405	Grade 3–9	FUNA-DB dyscalculia screener	CU, PF	EFA/CFA, AUC/ROC, invariance
Hirsch et al. (2018)	Netherlands/ Germany	1,728	Kindergarten→Grade 6	Basic numerical competences	CU	α, EFA/CFA
Howard et al. (2017)	Australia	91	Grade 2	NAPLAN numeracy subtest (Rasch)	PF, SC	Rasch/IRT
Jiang et al. (2021)	China	310	Grade 6	Fraction subconstructs test	CU	Rasch/IRT
Jones et al. (2019)	United Kingdom	188	Year 5	Comparative judgement (conceptual underst.)	CU	α, reliability
Karika & Csíkos (2022)	Hungary	124	Grade 5	Simple-fractions understanding test	CU	α
Kattou et al. (2013)	Cyprus	359	Grade 4–6	Mathematical Ability + Creativity test	CU	α, EFA/CFA, content
Khontee & Poonputta (2025)	Thailand	530	Grade 6	Mathematical Reasoning Ability Test	AR	KR-20, α, EFA/CFA, content
Kosko (2019)	USA	196	Grade 4–5	Multiplicative Reasoning Assessment	SC, AR	Rasch/IRT

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Study	Country	N	Grade	Instrument	Strand(s)	Key measurement evidence
Lee & Lembke (2016)	USA	~300/grade	K–Grade 3	CBM early numeracy/maths	CU, PF, SC	α, reliability
Li & Yang (2010)	Taiwan	~1,200+; retest 228	Grade 5	Computer-administered Number Sense Scale	CU	EFA/CFA, reliability
Marwiang et al. (2017)	Thailand	520	Grade 5–6	Maths Problem-Solving Skill framework	SC, AR	α, Rasch/IRT
Mejias et al. (2019)	France/ Luxembourg	346	Grade 1	Mathematical School Readiness test	CU, PF	—
Miao et al. (2022)	China	2,642	Grade 5–6	Rational numbers & proportional reasoning	SC, AR	Rasch/IRT
Morra et al. (2019)	Italy	779	preschool–Grade 5	Quantitative Central Conceptual Structures	CU	—
Nezhnov et al. (2015)	Russia	2,216	Grade 4	SAM-Math	CU, PF	Rasch/IRT
Park et al. (2024)	USA	22 studies / 4,130 participants	preschool–Grade 1	Early Numeracy CBM	CU, PF	— (meta-analysis)
Rodrigues et al. (2019)	USA	411/362/304	Grade 4–6	Fraction measures (screener)	CU, PF	AUC/ROC
Rodriguez-Segura & Schueler (2022)	Kenya	2,644	Grade 3, 5, 6	Phone-based numeracy assessment	PF, SC	α, Rasch/IRT
Schadl & Lindmeier (2025)	Germany	147	Grade 5	Digital SMDS & PRS (fraction prior knowledge)	PF, SC, AR	Rasch/IRT
Sánchez-Barbero et al. (2022)	Spain	318	Grades 1–6	Reversibility-of-thought scale (word problems)	SC, AR	α, EFA/CFA, content
Van Hoof et al. (2015)	Belgium	grade 4 + multi-grade	Grade 4	Natural-number-bias / rational-number test	CU, AR	EFA/CFA, Rasch/IRT
Van Norman et al. (2026)	USA	211 (K) + 79 (G1)	Kindergarten + Grade 1	Early-math CBM vs CAT (i-Ready)	CU, PF	reliability
Vidmar et al. (2017)	Slovenia/ Germany	468 + 328	school entry ~6 yr + G1–2	PIPS (literacy + numeracy)	PF, SC	EFA/CFA, reliability, invariance
Yang (2019)	Taiwan	819	Grade 5	Three-Tier Number Sense Test	CU, AR	α, content
Yanti et al. (2026)	Indonesia	500	Grade 4	Word-problem-solving instrument	SC, AR	Rasch/IRT, invariance, content
Yuan et al. (2022)	China	817	Grade 5	Fraction-operations test	PF	Rasch/IRT

Note. CU = conceptual understanding; PF = procedural fluency; SC = strategic competence; AR = adaptive reasoning; CBM = curriculum-based measurement; CAT = computer-adaptive test; EFA/CFA = exploratory/confirmatory factor analysis; IRT = item-response theory; AUC/ROC = area under the receiver-operating-characteristic curve; α = Cronbach's alpha; KR-20 = Kuder–Richardson 20. "—" denotes no quantitative measurement-property statistic extracted (e.g., evidence syntheses).

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

3.2. Instruments and Assessment Formats

The corpus encompassed a broad range of instruments. The most common were curriculum-based measures (CBMs) of early numeracy and arithmetic (Lee & Lembke, 2016; de León et al., 2021; Park et al., 2024; Rodrigues et al., 2019; Van Norman et al., 2026) and number-sense tests or scales (Brankaer et al., 2017; Li & Yang, 2010; Yang, 2019; Gökçe et al., 2023; Chen et al., 2017; Mejias et al., 2019). A second cluster comprised fraction- and rational-number tests (Jiang et al., 2021; Karika & Csíkos, 2022; Yuan et al., 2022; Schadl & Lindmeier, 2025; Van Hoof et al., 2015; Díaz-Cárdenas et al., 2019), while a third addressed mathematical reasoning and problem solving (Kosko, 2019; Miao et al., 2022; Khontee & Poonputta, 2025; Marwiang et al., 2017; Yanti et al., 2026; Sánchez-Barbero et al., 2022). A smaller set targeted integrated proficiency or competence (Bellini et al., 2019; Nezhnov et al., 2015; Diputra et al., 2025) and diagnostic screeners for mathematical learning difficulties (Hakkarainen et al., 2025; Hellstrand et al., 2024).

Response formats were equally varied. Beyond conventional multiple-choice and open-ended items, the corpus included timed CBM probes (Lee & Lembke, 2016), three-tier items combining an answer, a reason, and a confidence rating (Yang, 2019), comparative-judgement assessment (Jones et al., 2019), a self-assessment questionnaire (Chaiarwut et al., 2025), and several digital or online formats, including a phone-based assessment administered during school closures (Rodriguez-Segura & Schueler, 2022) and computer-administered or item-response-driven tests (Li & Yang, 2010; Schadl & Lindmeier, 2025). The content focus was concentrated on a few curricular domains number sense, whole-number operations, and fractions or rational numbers reflecting their foundational role in primary mathematics; fractions alone were the explicit target of seven instruments, while whole-number operations and place value featured prominently across the curriculum-based measures. Notably, formats designed to capture higher-order reasoning or conceptual understanding rather than answer accuracy alone were comparatively rare: the three-tier test (Yang, 2019), comparative judgement (Jones et al., 2019), the mixed multiple-choice-and-essay design analysed with many-facet Rasch measurement (Ahyan et al., 2025), and the open-ended tasks scored with a partial-credit model (Marwiang et al., 2017) were the principal exemplars. Overall, most instruments measured number sense, numeracy, or fractions, and few were designed to assess mathematical proficiency as an integrated, multi-strand construct.

3.3. Operationalisation of the Four Cognitive Strands

Coverage of the four Kilpatrick strands was markedly uneven (Figure 3a). Conceptual understanding was operationalised most frequently, in 26 of the 40 studies (65%), typically through number sense, fraction understanding, or conceptual numerical knowledge (e.g., Brankaer et al., 2017; Jiang et al., 2021; Yang, 2019; Balt et al., 2020). Procedural fluency appeared in 18 studies (45%), usually as arithmetic fluency, number processing, or computation (e.g., Hellstrand et al., 2024; Lee & Lembke, 2016; Yuan et al., 2022). Strategic competence was operationalised in 13 studies (33%), predominantly via mathematical word problems and problem-solving tasks (e.g., Marwiang et al., 2017; Yanti et al., 2026; de León et al., 2021). Adaptive reasoning was the least represented strand, explicit in only 12 studies (30%); where it was assessed, it took the form of multiplicative reasoning (Kosko, 2019), proportional reasoning (Miao et al., 2022), a dedicated mathematical-reasoning-ability test (Khontee & Poonputta, 2025), the reasoning tier of a three-tier instrument (Yang, 2019), or rational-number reasoning (Van Hoof et al., 2015). The strands also tended to co-occur in predictable pairings: conceptual understanding and procedural fluency were frequently assessed together within number-sense and fraction instruments, whereas strategic competence and adaptive reasoning co-occurred within problem-solving and reasoning instruments, mirroring the conceptual proximity of these strand pairs. The scoring of strands likewise differed systematically: conceptual understanding and procedural fluency were most often scored dichotomously and analysed with Rasch/IRT models, whereas strategic competence and adaptive reasoning being elicited through open-ended or justification tasks more often required polytomous, rubric-based scoring (e.g., the partial-credit scoring of Marwiang et al., 2017), which in turn raises the question of inter-rater consistency addressed. Table 4 synthesises, for each strand, its typical operationalisation, the predominant item format and scoring, and the complete set of studies that operationalised it.

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Table 4. Operationalisation of the Four Cognitive Strands, With All Studies Measuring Each Strand

Strand	Typical operationalisation	Predominant format & scoring	Studies
Conceptual understanding (CU)	Number sense; fraction and part–whole understanding; conceptual numerical knowledge; representation	Multiple-choice or short-answer; mostly dichotomous; some multi-tier	Akther & Powell (2025); Balt et al. (2020); Bellini et al. (2019); Brankaer et al. (2017); Chen et al. (2017); de León et al. (2021); Diputra et al. (2025); Díaz-Cárdenas et al. (2019); Gökçe et al. (2023); Hakkarainen et al. (2025); Hellstrand et al. (2024); Hirsch et al. (2018); Jiang et al. (2021); Jones et al. (2019); Karika & Csíkos (2022); Kattou et al. (2013); Lee & Lembke (2016); Li & Yang (2010); Mejias et al. (2019); Morra et al. (2019); Nezhnov et al. (2015); Park et al. (2024); Rodrigues et al. (2019); Van Hoof et al. (2015); Van Norman et al. (2026); Yang (2019)
Procedural fluency (PF)	Arithmetic fluency; number processing; computation; fraction operations	Timed; dichotomous	Akther & Powell (2025); Bellini et al. (2019); de León et al. (2021); Diputra et al. (2025); Díaz-Cárdenas et al. (2019); Hakkarainen et al. (2025); Hellstrand et al. (2024); Howard et al. (2017); Lee & Lembke (2016); Mejias et al. (2019); Nezhnov et al. (2015); Park et al. (2024); Rodrigues et al. (2019); Rodriguez-Segura & Schueler (2022); Schadl & Lindmeier (2025); Van Norman et al. (2026); Vidmar et al. (2017); Yuan et al. (2022)
Strategic competence (SC)	Contextual word-problem solving; mathematical modelling	Open-ended; polytomous / rubric	Chaiarwut et al. (2025); de León et al. (2021); Diputra et al. (2025); Howard et al. (2017); Kosko (2019); Lee & Lembke (2016); Marwiang et al. (2017); Miao et al. (2022); Rodriguez-Segura & Schueler (2022); Schadl & Lindmeier (2025); Sánchez-Barbero et al. (2022); Vidmar et al. (2017); Yanti et al. (2026)
Adaptive reasoning (AR)	Justification; multiplicative and proportional reasoning; evaluation of solutions	Open-ended or reason-tier; polytomous / rubric	Burte et al. (2019); Chaiarwut et al. (2025); Diputra et al. (2025); Khontee & Poonputta (2025); Kosko (2019); Marwiang et al. (2017); Miao et al. (2022); Schadl & Lindmeier (2025); Sánchez-Barbero et al. (2022); Van Hoof et al. (2015); Yang (2019); Yanti et al. (2026)

Critically, only one study operationalised all four cognitive strands within a single instrument: Diputra et al. (2025) administered a 12-item test (scored 0–3) to 220 fifth-grade pupils to capture conceptual understanding, procedural fluency, strategic competence, and adaptive reasoning. However, its exploratory factor analysis (Kaiser–Meyer–Olkin = 0.762) yielded only two factors, with conceptual understanding and procedural fluency collapsing together, indicating that the four strands were not empirically separable at that item density. By contrast, studies that successfully recovered an intended internal structure did so by aligning items to a theory-based hierarchy and validating it with Rasch modelling: Nezhnov et al. (2015) designed the SAM-Math test in blocks of three items per mastery level (procedural, conceptual, and functional) and confirmed the hierarchy through item difficulty ordering; Marwiang et al. (2017) anchored open-ended tasks to a four-level progress map; and Balt et al. (2020) drew five items per developmental level from a calibrated item pool. These patterns indicate that recovering distinct strands or levels depends less on covering all strands than on a sufficient number of theory-aligned items per target and an appropriate measurement model.

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

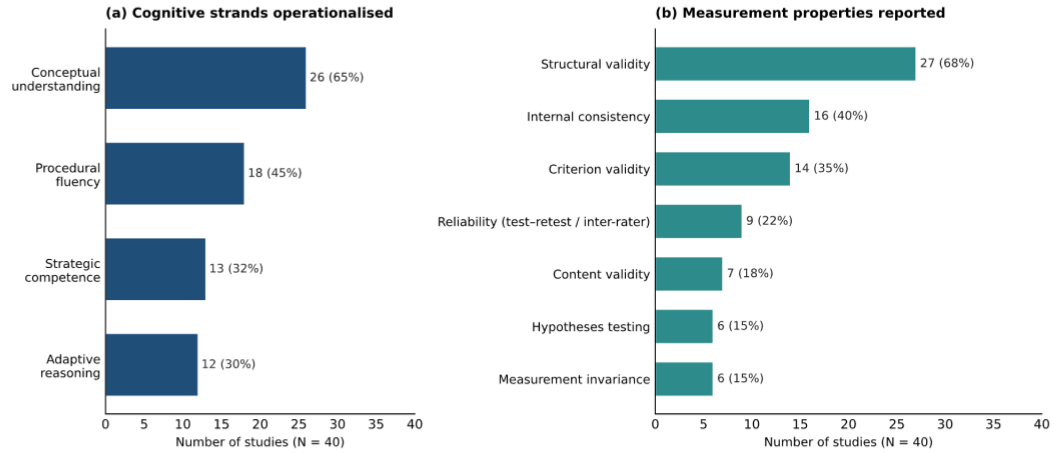

Figure 3. Coverage of (a) the Four Cognitive Strands and (b) Reported Measurement Properties

3.4. Reported Measurement Properties and Methodological Quality

The breadth of reported measurement properties was likewise uneven (Figure 3b); Table 5 complements that overview by pairing each property with the methods most often used, representative values, and the complete set of studies reporting it. Structural validity was the most frequently reported property, in 27 studies (68%), and constituted the methodological strength of the corpus. It was dominated by the Rasch/IRT family rather than by factor-analytic modelling of the strand structure: applications included many-facet Rasch measurement, which simultaneously models item difficulty and rater severity (Ahyan et al., 2025); the partial-credit model for polytomous, open-ended responses (Marwiang et al., 2017); cognitive-diagnostic models such as the rule-space model (Chen et al., 2017) and the hierarchical diagnostic classification model (Yuan et al., 2022); and the mixture Rasch model for latent classes (Miao et al., 2022). Exploratory and confirmatory factor analysis were used mainly for scales and questionnaires—for instance, a two-factor structure for the dyscalculia screener (Hakkarainen et al., 2025; Hellstrand et al., 2024), a five-factor structure for early numeracy (Hirsch et al., 2018), a second-order model of mathematical ability (Kattou et al., 2013), and an EFA-to-CFA validation of an executive problem-solving questionnaire (Chaiarwut et al., 2025) in which six factors explained 76% of the variance and the confirmatory model fitted well.

Table 5. Reported Measurement Properties with Representative Evidence and All Studies Reporting Each Property

Measurement property	Typical method(s)	Representative values / criteria	Studies
Structural validity	Rasch/IRT, many-facet Rasch, partial-credit, cognitive-diagnostic models; EFA/CFA for scales	Acceptable item fit; CFI = 1.00, RMSEA = .000 (Khontee & Poonputta, 2025); two-factor solution (Hakkarainen et al., 2025)	Ahyan et al. (2025); Balt et al. (2020); Bellini et al. (2019); Chaiarwut et al. (2025); Chen et al. (2017); de León et al. (2021); Diputra et al. (2025); Díaz-Cárdenas et al. (2019); Hakkarainen et al. (2025); Hellstrand et al. (2024); Hirsch et al. (2018); Howard et al. (2017); Jiang et al. (2021); Kattou et al. (2013); Khontee & Poonputta (2025); Kosko (2019); Li & Yang (2010); Marwiang et al. (2017); Miao et al. (2022); Nezhnov et al. (2015); Rodriguez-Segura & Schueler (2022); Schadl & Lindmeier (2025); Sánchez-Barbero et al. (2022); Van Hoof et al. (2015); Vidmar et al. (2017); Yanti et al. (2026); Yuan et al. (2022)
Internal consistency	Cronbach's α; KR-20	α = 0.73–0.95; KR-20 = 0.92 / 0.95	Burte et al. (2019); Chaiarwut et al. (2025); Díaz-Cárdenas et al. (2019); Hakkarainen et al. (2025); Hirsch et al. (2018); Jones et al. (2019); Karika & Csíkos (2022); Kattou et al. (2013); Khontee & Poonputta (2025); Lee & Lembke (2016); Li & Yang (2010); Marwiang et al. (2017); Rodriguez-Segura & Schueler (2022); Sánchez-Barbero et al. (2022); Van Norman et al. (2026); Yang (2019)
Criterion validity	ROC/AUC; concurrent correlation	AUC ≥ 0.80–0.85	Ahyan et al. (2025); Balt et al. (2020); Brankaer et al. (2017); de León et al. (2021); Hellstrand et al. (2024); Hirsch et al. (2018); Jones et al. (2019); Lee & Lembke (2016); Nezhnov et al. (2015); Park et al. (2024); Rodrigues et al. (2019); Rodriguez-Segura & Schueler (2022); Van Norman et al. (2026); Vidmar et al. (2017)
Reliability (test–retest / inter-	Test–retest r; alternate-	r ≥ .70; alternate-form 0.62–0.99	Ahyan et al. (2025); Brankaer et al. (2017); Burte et al. (2019); Hakkarainen et al. (2025); Jones et al. (2019);

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Measurement property	Typical method(s)	Representative values / criteria	Studies
rater / alternate-form)	form; rater-severity facet (MFRM)		Lee & Lembke (2016); Li & Yang (2010); Van Norman et al. (2026); Vidmar et al. (2017)
Content validity	Expert review; Aiken's V; item-objective congruence	IOC = 0.60–1.00	Ahyan et al. (2025); Bellini et al. (2019); Kattou et al. (2013); Khontee & Poonputta (2025); Sánchez-Barbero et al. (2022); Yang (2019); Yanti et al. (2026)
Measurement invariance	Differential item functioning; multi-group CFA	Invariant by grade/ gender/ language; weak invariance not supported across countries	Balt et al. (2020); Gökçe et al. (2023); Hakkarainen et al. (2025); Hellstrand et al. (2024); Vidmar et al. (2017); Yanti et al. (2026)
Hypotheses (construct) testing	Convergent / discriminant correlations	Convergent with achievement; discriminant with literacy	Brankaer et al. (2017); Hakkarainen et al. (2025); Howard et al. (2017); Morra et al. (2019); Schadl & Lindmeier (2025); Yang (2019)
Responsiveness	—	Not established in any included study	—

Internal consistency was reported in 16 studies (40%), with coefficients ranging from α = 0.73 (Hirsch et al., 2018) to α = 0.948 (Karika & Csíkos, 2022); other strong values included α = 0.83 for a partial-credit problem-solving test (Marwiang et al., 2017), α = 0.90–0.97 across subscales of a digital screener (Hakkarainen et al., 2025), and Kuder–Richardson coefficients of 0.92 and 0.95 for reasoning and knowledge tests (Khontee & Poonputta, 2025). Classical item analysis complemented these indices in several test-development studies for example, item difficulties between 0.15 and 0.80 and discrimination indices at or above 0.30 guided item retention for a number-sense scale (Li & Yang, 2010). Criterion validity appeared in 14 studies (35%), most often as classification accuracy for screening, with areas under the curve at or above 0.80 (de León et al., 2021; Rodrigues et al., 2019), and as concurrent comparisons against in-person standardised tests (Rodriguez-Segura & Schueler, 2022).

By contrast, three properties were conspicuously under-reported. Reliability in the sense of test–retest, inter-rater, or alternate-form agreement was reported in only nine studies (23%); test–retest evidence appeared in a few instruments (Brankaer et al., 2017; Li & Yang, 2010) and alternate-form reliability in early-math measures (Van Norman et al., 2026), but inter-rater reliability was almost absent, surfacing only through the rater-severity facet of many-facet Rasch measurement (Ahyan et al., 2025) and the scale-separation and inter-rater estimates of comparative judgement (Jones et al., 2019). This scarcity is consequential, the strands most dependent on rubric- or open-response scoring (strategic competence and adaptive reasoning) are precisely those whose scores are most vulnerable to rater variability. Content validity was quantified in only seven studies (18%), through expert review, Aiken's V (Ahyan et al., 2025; Yanti et al., 2026), or the index of item-objective congruence (Khontee & Poonputta, 2025). Measurement invariance was examined in six studies (15%): it was supported across grade, gender, and language for a large-scale screener (Hakkarainen et al., 2025; Hellstrand et al., 2024) and tested through differential item functioning (Gökçe et al., 2023; Yanti et al., 2026), but weak invariance was not supported for a school-entry instrument compared across Slovenia and Germany (Vidmar et al., 2017), underscoring the fragility of cross-context comparability. Responsiveness the ability to detect change over time was not formally established as a measurement property in any included study, even where instruments were used for progress monitoring.

Considered against the COSMIN framework, the methodological quality of structural-validity evidence was generally adequate where Rasch/IRT models were applied with acceptable fit and dimensionality checks. However, few studies triangulated multiple properties: only a minority (e.g., Hakkarainen et al., 2025; Khontee & Poonputta, 2025; Sánchez-Barbero et al., 2022) reported a comprehensive battery spanning content, structural, internal-consistency, and at least one further property. Moreover, sound structural validity did not guarantee precise person measurement: Yanti et al. (2026) reported low person reliability (0.58, then 0.67 across two Rasch trials) for a 30-item word-problem test, demonstrating that item quantity alone does not ensure measurement precision. The strength of structural-validity claims also varied with sample size: large calibration samples (e.g., Bellini et al., 2019; Hellstrand et al., 2024; Nezhnov et al., 2015) supported stable Rasch/IRT estimation, whereas several validations rested on modest samples (e.g., Howard et al., 2017, N = 91; Karika & Csíkos, 2022, N = 124; Schadl & Lindmeier, 2025, N = 147), which constrains the precision of their structural inferences. Read together, Figure 3b and Table 5 reveal a steep gradient: properties tied to internal structure and score consistency are comparatively well covered, whereas those securing cross-context comparability (invariance) and sensitivity to change (responsiveness) are largely or wholly absent a profile characteristic of a field still concentrated on the earliest stages of instrument validation. In sum, the corpus establishes structural validity anchored in the Rasch/IRT family as the field's principal strength, while inter-

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

rater reliability, quantified content validity, measurement invariance, and responsiveness remain its most pronounced gaps.

3.5. Assessment delivery and digital formats

Although most instruments were administered on paper, a substantial minority were delivered in digital, online, computer-adaptive, or telephone-based formats, reflecting the field's gradual digitalisation. These included the online FUNA-DB dyscalculia screener (Hakkarainen et al., 2025; Hellstrand et al., 2024), a computer-administered number-sense scale (Li & Yang, 2010), digital fraction tests analysed with IRT (Schadl & Lindmeier, 2025), a web-based three-tier number-sense test (Yang, 2019), a computer-adaptive test benchmarked against curriculum-based measures (Van Norman et al., 2026), and a phone-based numeracy assessment deployed during pandemic school closures (Rodriguez-Segura & Schueler, 2022). Where delivery modes were compared, the digital format generally preserved the instrument's psychometric properties but introduced administration-specific considerations: phone-based scores were valid for aggregate but not individual-level inference (Rodriguez-Segura & Schueler, 2022), and a digital fraction sub-test retained good reliability whereas its companion proportional-reasoning sub-test required further adaptation before it performed adequately (Schadl & Lindmeier, 2025). Digital delivery therefore expands reach and efficiency but does not, in itself, confer measurement quality, which remained contingent on item design and the underlying measurement model an observation of direct relevance to the development of technology-based mathematics assessments.

3.6. Discussion

This review set out to map how mathematical proficiency and its four cognitive strands are operationalised and measured in primary-school students, and to appraise the quality of the resulting measurement evidence. Three findings stand out: coverage of the strands is uneven and tilts away from adaptive reasoning; the strands are difficult to separate empirically at the item densities typically used; and the field's measurement practice is anchored in the Rasch/item-response-theory (IRT) family yet leaves inter-rater reliability, content validity, measurement invariance, and responsiveness largely unexamined. We discuss each in turn and derive an evidence-based blueprint for instrument development (Figure 4). Unlike earlier syntheses that catalogued early-numeracy screeners (Park et al., 2024; Akther & Powell, 2025), this review evaluates the measurement of the full four-strand proficiency construct together with the psychometric quality of the evidence base.

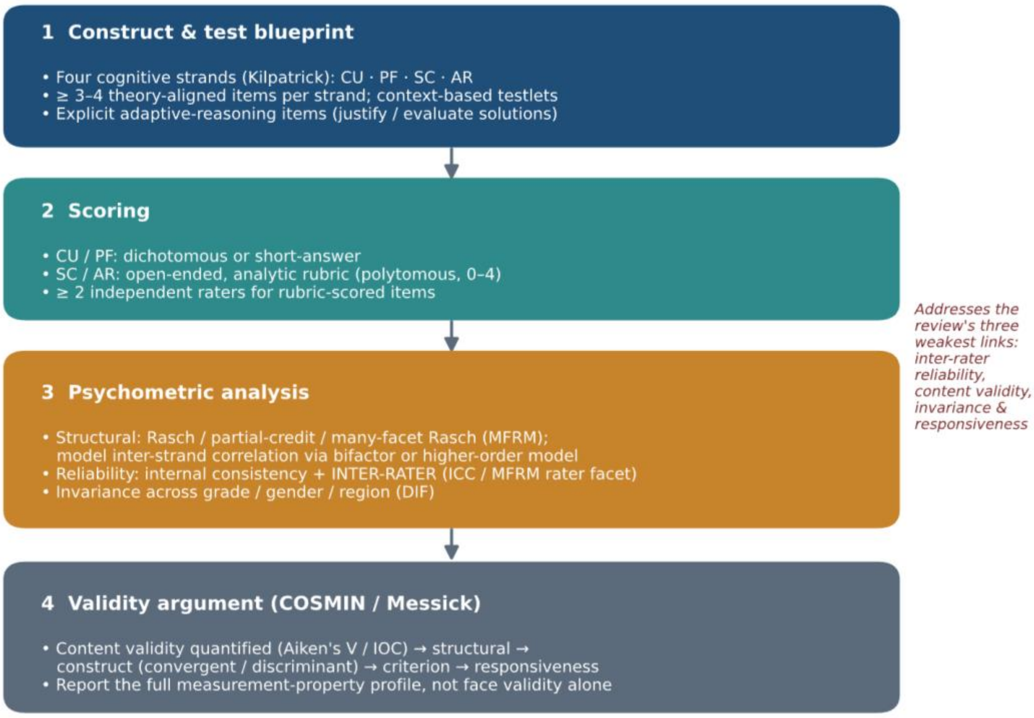

Figure 4. Blueprint for Instrument Development

3.7. An Uneven Map of the Four Strands

The dominance of conceptual understanding and procedural fluency, and the relative neglect of strategic competence and especially adaptive reasoning, echoes a long-standing concern that most mathematics tests address only part of the proficiency construct. Kilpatrick et al. (2001) insisted that the strands are interwoven and mutually reinforcing, yet our corpus shows them being assessed largely in isolation, with adaptive reasoning

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

explicit in fewer than a third of studies. The affective fifth strand, productive disposition, lay outside our cognitive-measurement scope, so the imbalance documented here concerns the four cognitive strands alone and is striking precisely because these are the strands a performance test can, in principle, capture. This imbalance is not incidental: reasoning and strategic problem solving are intrinsically harder to elicit and score than computation. Copur-Gencturk and Doleck (2021) describe strategic competence for multistep fraction problems as "overlooked," and studies that foreground reasoning typically rely on richer, more demanding designs realistic mathematics tasks (Palinussa et al., 2021), justification items (Kosko, 2019), or reason tiers (Yang, 2019). Where conceptual understanding is targeted, it is frequently operationalised through number sense and fraction understanding (Copur-Gencturk, 2021; Jiang et al., 2021), domains that lend themselves to dichotomous scoring and large-scale calibration. The practical consequence is a literature that measures the more tractable strands well and the harder strands seldom precisely the inverse of what a balanced proficiency framework requires.

3.8. The Dimensionality Problem

The single study that operationalised all four cognitive strands in one instrument (Diputra et al., 2025) recovered only two factors, with conceptual understanding and procedural fluency coalescing. This is the predictable result of two compounding pressures: the strands are theoretically interdependent, and twelve items cannot sustain a stable four-factor solution. The measurement literature offers a constructive resolution. Because mathematical abilities comprise a strong general dimension alongside narrower specific abilities (Pokropek et al., 2022), bifactor and higher-order models rather than a set of orthogonal factors are better suited to representing proficiency (Akaeze et al., 2022), provided that model fit and ancillary indices are evaluated rigorously (Goretzko et al., 2023); indeed, the factor structure of early-mathematics constructs is itself an active question (Purpura et al., 2020). Equally important is item density: studies that recovered an intended structure did so by anchoring a sufficient number of theory-aligned items to each target blocks of items per mastery level (Nezhnov et al., 2015) or per developmental level (Balt et al., 2020). The implication for instrument design is twofold: provide enough items per strand we suggest at least three to four and model the inter-strand correlation explicitly rather than assuming separability.

3.9. Strong Structural Validity, Weak Score-Consistency Evidence

The corpus's reliance on the Rasch/IRT family is a genuine strength: these models place persons and items on a common scale, support unidimensionality checks, and through the partial-credit and many-facet extensions accommodate the polytomous, rubric-scored responses that strategic competence and adaptive reasoning demand. Yet only the many-facet Rasch application (Ahyan et al., 2025) and comparative judgement (Jones et al., 2019) reported inter-rater evidence, leaving the rubric-scored strands most exposed to the very source of error they cannot ignore. Rater effects can account for a substantial share of score variance, and many-facet Rasch measurement is well established as the means to model rater severity and bias within performance assessment (Uto & Ueno, 2020; Uto, 2021; Li, 2022), including in Indonesian rubric-based settings (Permana & Widodo, 2022). The near-absence of inter-rater reliability is therefore not a minor omission but a threat to the reproducibility of exactly the scores that matter most for the higher-order strands.

3.10. Under-Examined Validity Evidence

Content validity was quantified in fewer than a fifth of studies, measurement invariance in fewer than a sixth, and responsiveness in none. These gaps matter for both inference and equity. Quantified content validity through expert-agreement indices such as Aiken's V or the index of item-objective congruence is the first link in a defensible validity argument (Beck, 2020; Kaldaras et al., 2024), and its scarcity means many claims of "measuring proficiency" rest on face validity alone. Measurement invariance is the precondition for any cross-group comparison; its fragility was demonstrated directly when weak invariance failed across two national versions of a school-entry instrument (Vidmar et al., 2017). Contemporary methodological work cautions that invariance is neither automatic nor a box to be ticked (Robitzsch & Lüdtke, 2023) and that multigroup and alignment approaches are required to test it properly (Immekus, 2021; Ding et al., 2022). Read against the COSMIN framework (Gagnier et al., 2021; Mokkink et al., 2020, 2024), the corpus exhibits a steep evidentiary gradient strong on internal structure, thin on the properties that secure comparability and sensitivity to change.

3.11. Digital Delivery and Implications for Practice

Digital, computer-adaptive, and telephone-based formats are entering primary mathematics assessment, and our findings indicate that they preserve psychometric quality only when the underlying item design and measurement model are sound consistent with advances in computerized adaptive testing built on cognitive diagnosis (Junpeng et al., 2020; Zhang et al., 2025) and with cautions about validity in technology- and AI-mediated assessment (Kasneci et al., 2023; Getenet, 2024).

These findings translate into a concrete blueprint for developing proficiency instruments (Figure 4): specify the four strands with at least three to four theory-aligned items each and context-based testlets; score strategic competence and adaptive reasoning with analytic rubrics applied by at least two independent raters;

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

analyse with Rasch, partial-credit, or many-facet models within a bifactor or higher-order structure; and assemble a full validity argument that begins with quantified content validity and extends through inter-rater reliability, invariance, and responsiveness. For the Indonesian context in particular where the Minimum Competency Assessment has made numeracy a national priority (Purnomo et al., 2022; Hidayah et al., 2021) and primary-level numeracy and higher-order thinking remain developmental concerns (Tanudjaya & Doorman, 2020; Rohmah et al., 2022) such a blueprint offers a locally relevant route to instruments that are simultaneously curriculum-aligned and psychometrically defensible. More broadly, the gradient documented here suggests a field maturing from instrument creation toward instrument validation; closing the inter-rater, content-validity, and invariance gaps is the decisive next step toward measures whose scores can be trusted, compared, and acted upon.

3.12. Limitations

Several limitations temper these conclusions. The search was restricted to English-language articles in two databases, and seven full texts could not be retrieved, so some relevant instruments particularly those reported in national-language journals may be under-represented. The COSMIN apparatus was developed for patient-reported outcomes and was adapted here for performance-based achievement tests; although we anchored its use in the educational validity tradition, its criteria for "sufficient" structural validity translate imperfectly to Rasch-scaled achievement data. Finally, the heterogeneity of constructs, instruments, and designs precluded meta-analysis, so the synthesis is narrative and the property-coverage counts should be read as indicative rather than definitive. These constraints notwithstanding, the convergent pattern across 40 studies strong structural validity alongside weak score-consistency and comparability evidence is unlikely to be an artefact of selection.

4. Conclusion

Across 40 studies, the measurement of primary-school mathematical proficiency shows uneven strand coverage conceptual understanding and procedural fluency dominate, strategic competence is less frequent, and adaptive reasoning is rarest, with no single instrument convincingly separating all four strands and a steep evidentiary gradient: structural validity, anchored in the Rasch and item-response-theory family, is a genuine strength, whereas inter-rater reliability, quantified content validity, measurement invariance, and responsiveness remain largely unexamined. Future instruments should therefore specify each strand with several theory-aligned items and context-based testlets, score reasoning and problem-solving tasks with analytic rubrics applied by at least two independent raters, model the interdependent strands with bifactor or many-facet approaches, and build a full validity argument beginning with quantified content validity a blueprint of direct relevance to assessment reform in contexts such as Indonesia, where numeracy is a national priority.

Author Contributions

Every author was involved in conceiving the study, developing the methodology, conducting the investigation, curating the data, performing formal analysis, validating the results, and writing the manuscript. All authors have reviewed and given final approval of the manuscript and accept responsibility for all aspects of this work.

Funding

No funding support was received.

Declaration of Conflicting Interests

The authors confirm that there are no known financial conflicts or personal relationships that might have been perceived to bias the work presented in this paper.

Data Availability

All data supporting this study's findings are presented in the article itself. Additional materials including the review protocol, search strategies with corresponding hit counts, records of screening and eligibility decisions, the data-extraction matrix, the bank of instruments used, and the quality-appraisal sheets can be obtained from the corresponding author upon reasonable request.

Declaration on AI Use

In developing this manuscript, the authors employed Claude (Anthropic) to help draft certain sections and to refine the English-language writing. Following the use of this tool, the authors carefully reviewed and revised the resulting content, and they take complete responsibility for the final published material.

References

Ahyan, S., Supiyati, S., Fahrurrozi, & Hassan, M. N. (2025). An application of many-facet Rasch measurement to validate the numeracy test for elementary students. *Infinity Journal, 14*(4), 861–876. https://doi.org/10.22460/infinity.v14i4.p861-876

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Akaeze, H., Lawrence, F., & Wu, J. H. (2022). Resolving dimensionality in a child assessment tool: An application of the multilevel bifactor model. *Educational and Psychological Measurement, 83*(1), 93–115. https://doi.org/10.1177/00131644221082688

Akther, S. S., & Powell, S. R. (2025). A synthesis on early numeracy assessments within intervention for students in preschool through grade 2. *Exceptionality, 33*(5), 375–407. https://doi.org/10.1080/09362835.2025.2541576

Alabdulaziz, M. S. (2021). COVID-19 and the use of digital technology in mathematics education. *Education and Information Technologies, 26*(6), 7609–7633. https://doi.org/10.1007/s10639-021-10602-3

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aunio, P., Korhonen, J., Ragpot, L., Törmänen, M., & Henning, E. (2020). An early numeracy intervention for first-graders at risk for mathematical learning difficulties. *Early Childhood Research Quarterly, 55*, 252–262. https://doi.org/10.1016/j.ecresq.2020.12.002

Balt, M., Fritz, A., & Ehlert, A. (2020). Insights into first grade students' development of conceptual numerical understanding as drawn from progression-based assessments. *Frontiers in Education, 5*, Article 80. https://doi.org/10.3389/feduc.2020.00080

Beck, K. (2020). Ensuring content validity of psychological and educational tests—The role of experts. *Frontline Learning Research, 8*(6), 1–37. https://doi.org/10.14786/flr.v8i6.517

Bellini, D., Crescentini, A., Zanolla, G., Cubico, S., Favretto, G., Faccincani, L., Ardolino, P., & Gianesini, G. (2019). Mathematical Competence Scale (MCS) for primary school: The psychometric properties and the validation of an instrument to enhance the sustainability of talents development through the numeracy skills assessment. *Sustainability, 11*(9), Article 2569. https://doi.org/10.3390/su11092569

Benavides-Varela, S., Callegher, C. Z., Fagiolini, B., Leo, I., Altoè, G., & Lucangeli, D. (2020). Effectiveness of digital-based interventions for children with mathematical learning difficulties: A meta-analysis. *Computers & Education, 157*, Article 103953. https://doi.org/10.1016/j.compedu.2020.103953

Björklund, C., van den Heuvel-Panhuizen, M., & Kullberg, A. (2020). Research on early childhood mathematics teaching and learning. *ZDM Mathematics Education, 52*(4), 607–619. https://doi.org/10.1007/s11858-020-01177-3

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*(4), 1361–1373. https://doi.org/10.3758/s13428-016-0792-3

Burte, H., Gardony, A. L., Hutton, A., & Taylor, H. A. (2019). Make-A-Dice Test: Assessing the intersection of mathematical and spatial thinking. *Behavior Research Methods, 51*(2), 602–638. https://doi.org/10.3758/s13428-018-01192-z

Chaiarwut, S., Kwangmuang, P., & Srikoon, S. (2025). Development of executive mathematics problem-solving self-assessment questionnaire for upper primary school students. *Social Sciences and Humanities Open, 12*, Article 101749. https://doi.org/10.1016/j.ssaho.2025.101749

Chang, I. (2023). Early numeracy and literacy skills and their influences on fourth-grade mathematics achievement: A moderated mediation model. *Large-scale Assessments in Education, 11*(1). https://doi.org/10.1186/s40536-023-00168-6

Chen, F., Yan, Y., & Xin, T. (2017). Developing a learning progression for number sense based on the rule space model in China. *Educational Psychology, 37*(2), 128–144. https://doi.org/10.1080/01443410.2016.1239817

Conto, C. A., Akseer, S., Dreesen, T. D., Kamei, A., Mizunoya, S., & Rigole, A. (2021). Potential effects of COVID-19 school closures on foundational skills and country responses for mitigating learning loss. *International Journal of Educational Development, 87*, Article 102434. https://doi.org/10.1016/j.ijedudev.2021.102434

Copur-Gencturk, Y. (2021). Teachers' conceptual understanding of fraction operations: Results from a national sample of elementary school teachers. *Educational Studies in Mathematics, 107*(3), 525–545. https://doi.org/10.1007/s10649-021-10033-4

Copur-Gencturk, Y., & Doleck, T. (2021). Strategic competence for multistep fraction word problems: An overlooked aspect of mathematical knowledge for teaching. *Educational Studies in Mathematics, 107*(1), 49–70. https://doi.org/10.1007/s10649-021-10028-1

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

Díaz-Cárdenas, A. F., Díaz-Furlong, A., Díaz-Furlong, H. A., Sankey-García, M. R., & Zago-Portillo, G. (2019). Multiplication and division of fractions: Numerical cognition development and assessment procedures. *Revista Latinoamericana de Investigación en Matemática Educativa, 22*(3), 333–362. https://doi.org/10.12802/relime.19.2234

Ding, Y., Hansen, K. Y., & Klapp, A. (2022). Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychology of Education, 38*(2), 709–732. https://doi.org/10.1007/s10212-022-00623-y

Diputra, K. S., Agustika, G. N. S., Utami, I. A. M. I., Julianto, P., & Arifuddin, A. (2025). Investigating mathematical proficiency of elementary school students: A foundation for effective learning models. *Edelweiss Applied Science and Technology, 9*(5), 314–323. https://doi.org/10.55214/25768484.v9i5.6831

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Drake, S. M., & Reid, J. L. (2020). 21st century competencies in light of the history of integrated curriculum. *Frontiers in Education, 5*, Article 122. https://doi.org/10.3389/feduc.2020.00122

Elsman, E. B. M., Mokkink, L. B., Terwee, C. B., Beaton, D., Gagnier, J. J., Tricco, A. C., Baba, A., Butcher, N. J., Smith, M., Hofstetter, C., Aiyegbusi, O. L., Berardi, A., Farmer, J., Haywood, K. L., Krause, K. R., Markham, S., Mayo-Wilson, E., Mehdipour, A., Ricketts, J., . . . Offringa, M. (2024). Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. *Journal of Clinical Epidemiology, 173*, Article 111422. https://doi.org/10.1016/j.jclinepi.2024.111422

Ersan, O., & Rodriguez, M. C. (2020). Socioeconomic status and beyond: A multilevel analysis of TIMSS mathematics achievement given student and school context in Turkey. *Large-scale Assessments in Education, 8*(1). https://doi.org/10.1186/s40536-020-00093-y

Gagnier, J. J., Lai, J., Mokkink, L. B., & Terwee, C. B. (2021). COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research, 30*(8), 2197–2218. https://doi.org/10.1007/s11136-021-02822-4

Gal, I., Grotlüschen, A., Tout, D., & Kaiser, G. (2020). Numeracy, adult education, and vulnerable adults: A critical view of a neglected field. *ZDM Mathematics Education, 52*(3), 377–394. https://doi.org/10.1007/s11858-020-01155-9

Getenet, S. (2024). Pre-service teachers and ChatGPT in multistrategy problem-solving: Implications for mathematics teaching in primary schools. *International Electronic Journal of Mathematics Education, 19*(1), Article em0766. https://doi.org/10.29333/iejme/14141

Gökçe, S., Güner, P., & Baştuğ, M. (2023). Monitoring proficiency: Growth of number sense in primary school. *Psychology in the Schools, 60*(3), 707–728. https://doi.org/10.1002/pits.22786

Gore, J., Fray, L., Miller, A., Harris, J., & Taggart, W. (2021). The impact of COVID-19 on student learning in New South Wales primary schools: An empirical study. *The Australian Educational Researcher, 48*(4), 605–637. https://doi.org/10.1007/s13384-021-00436-w

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement, 84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Hakkarainen, A., Väisänen, E., Aunio, P., Hellstrand, H., Laakso, M.-J., Laine, A., Räsänen, P., & Korhonen, J. (2025). Psychometric evidence for the Functional Numeracy Assessment Dyscalculia Battery (FUNA-DB Screener): An online assessment of mathematical learning difficulties. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000895

Hellstrand, H., Holopainen, S., Korhonen, J., Räsänen, P., Hakkarainen, A., Laakso, M.-J., Laine, A., & Aunio, P. (2024). Arithmetic fluency and number processing skills in identifying students with mathematical learning disabilities. *Research in Developmental Disabilities, 151*, Article 104795. https://doi.org/10.1016/j.ridd.2024.104795

Hidayah, I. R., Kusmayadi, T. A., & Fitriana, L. (2021). Minimum competency assessment (AKM): An effort to photograph numeracy. *Journal of Mathematics and Mathematics Education, 11*(1), 14. https://doi.org/10.20961/jmme.v11i1.52742

Hirsch, S., Lambert, K., Coppens, K., & Moeller, K. (2018). Basic numerical competences in large-scale assessment data: Structure and long-term relevance. *Journal of Experimental Child Psychology, 167*, 32–48. https://doi.org/10.1016/j.jecp.2017.09.015

Howard, S. J., Woodcock, S., Ehrich, J., & Bokosmaty, S. (2017). What are standardized literacy and numeracy tests testing? Evidence of the domain-general contributions to students' standardized educational test performance. *British Journal of Educational Psychology, 87*(1), 108–122. https://doi.org/10.1111/bjep.12138

Immekus, J. C. (2021). Multigroup CFA and alignment approaches for testing measurement invariance and factor score estimation: Illustration with the schoolwork-related anxiety survey across countries and gender. *Methodology, 17*(1), 22–38. https://doi.org/10.5964/meth.2281

Jiang, Z., Mok, I. A. C., & Li, J. (2021). Chinese students' hierarchical understanding of part-whole and measure subconstructs. *International Journal of Science and Mathematics Education, 19*(7), 1441–1461. https://doi.org/10.1007/s10763-020-10118-1

Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal, 45*(3), 662–680. https://doi.org/10.1002/berj.3519

Jordan, N. C., Devlin, B. L., & Botello, M. (2022). Core foundations of early mathematics: Refining the number sense framework. *Current Opinion in Behavioral Sciences, 46*, Article 101181. https://doi.org/10.1016/j.cobeha.2022.101181

Junpeng, P., Marwiang, M., Chinjunthuk, S., Suwannatrai, P., Chanayota, K., Pongboriboon, K., Tang, K. N., & Wilson, M. (2020). Validation of a digital tool for diagnosing mathematical proficiency. *International Journal of Evaluation and Research in Education, 9*(3), 665. https://doi.org/10.11591/ijere.v9i3.20503

Kaldaras, L., Akaeze, H., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education, 9*, Article 1399377. https://doi.org/10.3389/feduc.2024.1399377

Karika, T., & Csíkos, C. (2022). A test for understanding simple fractions among 5th grade students at the beginning of lower secondary education. *Eurasia Journal of Mathematics, Science and Technology Education, 18*(2), Article em2081. https://doi.org/10.29333/ejmste/11654

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G.

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

(2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kattou, M., Kontoyianni, K., Pitta-Pantazi, D., & Christou, C. (2013). Connecting mathematical creativity to mathematical ability. *ZDM Mathematics Education, 45*(2), 167–181. https://doi.org/10.1007/s11858-012-0467-1

Khontee, K., & Poonputta, A. (2025). Causal factors influencing the mathematical reasoning ability of Grade 6 students. *Educational Process: International Journal, 17*, Article e2025297. https://doi.org/10.22521/edupij.2025.17.297

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. National Academy Press. https://doi.org/10.17226/9822

Kosko, K. W. (2019). A multiplicative reasoning assessment for fourth and fifth grade students. *Studies in Educational Evaluation, 60*, 32–42. https://doi.org/10.1016/j.stueduc.2018.11.003

Lee, Y.-S., & Lembke, E. (2016). Developing and evaluating a kindergarten to third grade CBM mathematics assessment. *ZDM Mathematics Education, 48*(7), 1019–1030. https://doi.org/10.1007/s11858-016-0788-6

Li, A. W. (2022). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing, 35*(10), 2409–2431. https://doi.org/10.1007/s11145-022-10279-1

Li, M.-N. F., & Yang, D.-C. (2010). Development and validation of a computer-administered number sense scale for fifth-grade children in Taiwan. *School Science and Mathematics, 110*(4), 220–230. https://doi.org/10.1111/j.1949-8594.2010.00024.x

Marwiang, M., Klaha, J., Saree, L., & Junpeng, P. (2017). Assessing students' mathematical problem solving skill through the innovative lesson study and open approach. *Turkish Online Journal of Educational Technology*.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mejias, S., Muller, C., & Schiltz, C. (2019). Assessing mathematical school readiness. *Frontiers in Psychology, 10*, Article 1173. https://doi.org/10.3389/fpsyg.2019.01173

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Miao, Z., Bokhove, C., Reynolds, D., & Charalambous, C. Y. (2022). Rational numbers and proportional reasoning in Chinese primary schools: Patterns, latent classes, and reasoning processes. *Asian Journal for Mathematics Education, 1*(4), 408–436. https://doi.org/10.1177/27527263221143802

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), Article 1. https://doi.org/10.1186/2046-4053-4-1

Mokkink, L. B., Boers, M., van der Vleuten, C. P. M., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: A Delphi study. *BMC Medical Research Methodology, 20*(1), Article 293. https://doi.org/10.1186/s12874-020-01179-5

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Elsman, E. B. M., & Terwee, C. B. (2024). COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Quality of Life Research, 33*(11), 2929–2939. https://doi.org/10.1007/s11136-024-03761-6

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Morra, S., Bisagno, E., Caviola, S., Delfante, C., & Mammarella, I. C. (2019). Working memory capacity and the development of quantitative central conceptual structures. *Cognition and Instruction, 37*(4), 483–511. https://doi.org/10.1080/07370008.2019.1636797

Nezhnov, P., Kardanova, E., Vasilyeva, M., & Ludlow, L. (2015). Operationalizing levels of academic mastery based on Vygotsky's theory: The study of mathematical knowledge. *Educational and Psychological Measurement, 75*(2), 235–259. https://doi.org/10.1177/0013164414534068

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Palinussa, A. L., Molle, J. S., & Gaspersz, M. (2021). Realistic mathematics education: Mathematical reasoning and communication skills in rural contexts. *International Journal of Evaluation and Research in Education, 10*(2), 522. https://doi.org/10.11591/ijere.v10i2.20640

Park, S., Nelson, G., Shin, J., Clarke, B., Cook, M. A., & Hermida, J. (2024). Criterion validity of early numeracy curriculum-based measurement: A meta-analysis. *Learning Disabilities: A Contemporary Journal*.

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Permana, R. A. H. A., & Widodo, A. (2022). Validity and inter-rater reliability of the scoring rubrics for the science teacher TPACK test instrument. *Jurnal Penelitian dan Pembelajaran IPA, 8*(1), 27. https://doi.org/10.30870/jppi.v8i1.11164

Pokropek, A., Marks, G., Borgonovi, F., Koc, P., & Greiff, S. (2022). General or specific abilities? Evidence from 33 countries participating in the PISA assessments. *Intelligence, 92*, Article 101653. https://doi.org/10.1016/j.intell.2022.101653

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (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

Purnomo, H., Sa'dijah, C., Hidayanto, E., Sisworo, S., Permadi, H., & Anwar, L. (2022). Development of instrument numeracy skills test of Minimum Competency Assessment (MCA) in Indonesia. *International Journal of Instruction, 15*(3), 635–648. https://doi.org/10.29333/iji.2022.15335a

Purpura, D. J., King, Y. A., Rolan, E., Hornburg, C. B., Schmitt, S. A., Hart, S. A., & Ganley, C. M. (2020). Examining the factor structure of the home mathematics environment to delineate its role in predicting preschool numeracy, mathematical language, and spatial skills. *Frontiers in Psychology, 11*, Article 1925. https://doi.org/10.3389/fpsyg.2020.01925

Reder, S., Gauly, B., & Lechner, C. M. (2020). Practice makes perfect: Practice engagement theory and the development of adult literacy and numeracy proficiency. *International Review of Education, 66*(2–3), 267–288. https://doi.org/10.1007/s11159-020-09830-5

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal, 30*(6), 859–870. https://doi.org/10.1080/10705511.2023.2191292

Rodrigues, J., Jordan, N. C., & Hansen, N. (2019). Identifying fraction measures as screeners of mathematics risk status. *Journal of Learning Disabilities, 52*(6), 480–497. https://doi.org/10.1177/0022219419879684

Rodriguez-Segura, D., & Schueler, B. E. (2022). Can learning be measured by phone? Evidence from Kenya. *Economics of Education Review, 90*, Article 102309. https://doi.org/10.1016/j.econedurev.2022.102309

Rohmah, A. N., Sutama, S., Hidayati, Y. M., Fauziati, E., & Rahmawati, L. E. (2022). Planning for cultivation numerical literacy in mathematics learning for Minimum Competency Assessment (AKM) in elementary schools. *Mimbar Sekolah Dasar, 9*(3), 503–516. https://doi.org/10.53400/mimbar-sd.v9i3.51774

Salminen, J., Khanolainen, D., Koponen, T., Torppa, M., & Lerkkanen, M. (2021). Development of numeracy and literacy skills in early childhood—A longitudinal study on the roles of home environment and familial risk for reading and math difficulties. *Frontiers in Education, 6*, Article 725337. https://doi.org/10.3389/feduc.2021.725337

Sánchez-Barbero, B., Cáceres, M. J., & Chamoso, J. M. (2022). Validation of a scale to assess the reversibility of thought in verbal arithmetic problems. *Acta Scientiae, 24*(6), 370–401. https://doi.org/10.17648/acta.scientiae.7328

Schadl, C., & Lindmeier, A. (2025). Preparing for digital learning monitoring in the fraction context: Assessment of students' prior knowledge according to evidence-based cognitive models. *International Journal of Science and Mathematics Education, 23*(7), 2199–2224. https://doi.org/10.1007/s10763-024-10531-w

Tanudjaya, C. P., & Doorman, M. (2020). Examining higher order thinking in Indonesian lower secondary mathematics classrooms. *Journal on Mathematics Education, 11*(2), 277–300. https://doi.org/10.22342/jme.11.2.11000.277-300

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115–1123. https://doi.org/10.1007/s11136-009-9528-5

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research, 27*(5), 1159–1170. https://doi.org/10.1007/s11136-018-1829-0

Uto, M. (2021). A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. *Behaviormetrika, 48*(2), 425–457. https://doi.org/10.1007/s41237-021-00144-w

Uto, M., & Ueno, M. (2020). A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika, 47*(2), 469–496. https://doi.org/10.1007/s41237-020-00115-7

Van Hoof, J., Janssen, R., Verschaffel, L., & Van Dooren, W. (2015). Inhibiting natural knowledge in fourth graders: Towards a comprehensive test instrument. *ZDM Mathematics Education, 47*(5), 849–857. https://doi.org/10.1007/s11858-014-0650-7

Van Norman, E. R., Truman, K. J., & Hojnoski, R. L. (2026). A comparison of the validity of growth estimates from early math curriculum based measures and a computer adaptive test. *Psychology in the Schools, 63*, 457–470. https://doi.org/10.1002/pits.70105

Vidmar, M., Niklas, F., Schneider, W., & Hasselhorn, M. (2017). On-entry assessment of school competencies and academic achievement: A comparison between Slovenia and Germany. *European Journal of Psychology of Education, 32*(2), 311–331. https://doi.org/10.1007/s10212-016-0294-9

Wang, L., Chen, B., Hwang, G.-J., Guan, J., & Wang, Y. (2022). Effects of digital game-based STEM education on students' learning achievement: A meta-analysis. *International Journal of STEM Education, 9*(1). https://doi.org/10.1186/s40594-022-00344-0

Yang, D.-C. (2019). Development of a three-tier number sense test for fifth-grade students. *Educational Studies in Mathematics, 101*(3), 405–424. https://doi.org/10.1007/s10649-018-9874-8

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174

Yanti, S., Tjalla, A., Sutisna, A., & Ramdhan, T. W. (2026). Constructing an instrument to measure and evaluate problem-solving skills in elementary school students: A focus on mathematical word problem-solving. *Kasetsart Journal of Social Sciences, 47*(1), Article 470113. https://doi.org/10.34044/j.kjss.2026.47.1.13

Yayuk, E., Purwanto, P., As'ari, A. R., & Subanji, S. (2020). Primary school students' creative thinking skills in mathematics problem solving. *European Journal of Educational Research, 9*(3), 1281–1295. https://doi.org/10.12973/eu-jer.9.3.1281

Yuan, L., Liu, Y., Chen, P., & Xin, T. (2022). Development of a new learning progression verification method based on the hierarchical diagnostic classification model: Taking Grade 5 students' fractional operations as an example. *Educational Measurement: Issues and Practice, 41*(3), 69–82. https://doi.org/10.1111/emip.12488

Zhang, Y., Zhang, L., Zhang, H., & Wu, X. (2025). Development of a computerized adaptive assessment and learning system for mathematical ability based on cognitive diagnosis. *Journal of Intelligence, 13*(9), Article 114. https://doi.org/10.3390/jintelligence13090114

© 2026 The Author(s)
This is an Open Access article under the Creative Commons Attribution-ShareAlike 4.0 International License
ISSN: 2797-3174