

Rasch Model Applications in Validity and Reliability Testing of Educational Measurement Instruments: A PRISMA-Guided Systematic Literature Review (2020–2026)

Cynthia Mayang Sari¹, Ulya Amelia¹, Agustian Ramadana Putra¹, Syafa'at Ariful Huda¹, S. Nudia Mastur¹

¹Universitas Negeri Jakarta, Indonesia

ABSTRACT

Purpose – This study analyzes the use of the Rasch Model to test the validity and reliability of educational measurement instruments published between 2020 and 2026, based on 13 empirical articles retained from 5,097 records retrieved from ScienceDirect, ProQuest, and Google Scholar. The review examines the characteristics of the publications and instruments analyzed, the psychometric indicators most frequently reported, the comprehensiveness of psychometric evidence, and the methodological gaps identified in previous studies.

Methods – This study employed a systematic literature review approach guided by the PRISMA 2020 statement. After deduplication and title, abstract and full text screening against explicit inclusion and exclusion criteria, 13 empirical articles were retained for narrative synthesis.

Findings – The results of the study show that the Rasch model has been used in instruments that measure various fields in education. The most consistently reported indicators are item fit, infit-outfit mean squares, ZSTD, point-measure correlation, person and item reliability, and the separation index. However, reporting on unidimensionality, local independence, response category functions, and differential item functioning remains inconsistent.

Research Implications – The findings imply that psychometric reporting practices in Rasch-based validation studies need to be standardized. This study is limited by its concentration on literature from specific educational contexts, particularly Asian countries, and by variation in how Rasch indicators are reported across studies, indicating the need for broader cross-country research, more standardized psychometric reporting, and further validation through DIF, response category analysis, external validity evidence, and integration with CFA or EFA.

 OPEN ACCESS

ARTICLE HISTORY

Received: 09-07-2026

Revised: 30-09-2026

Accepted: 01-10-2026

KEYWORDS

rasch model, validity and reliability, educational measurement instruments, psychometric evidence, differential item functioning

Corresponding Author:

Cynthia Mayang Sari

Master's Program, Educational Management, Universitas Negeri Jakarta, Indonesia

Email: cynthiams@unj.ac.id

Introduction

Measurement underpins educational decision-making, whether assessing learning outcomes, evaluating instruction, mapping teacher competencies, or tracking constructs such as motivation, creativity, digital literacy, and higher-order thinking skills. Instruments lacking sufficient evidence of validity and reliability risk erroneous score interpretations, inaccurate student classifications, and irrelevant learning recommendations. Validity is therefore not a label permanently attached to an instrument but an accumulation of evidence supporting the interpretation and use of scores in a specific context (Messick, 1995), a view reinforced by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).

In educational research practice, instrument quality is still often judged with classical test theory indicators such as item-total correlations and alpha coefficients. These remain useful but say little about the relationship between respondent ability and item difficulty, the function of response categories, or item bias toward particular groups. The Rasch model offers a modern alternative that transforms ordinal data into interval measures on a logit scale, provided its assumptions are met, allowing researchers to evaluate item and person fit, item difficulty, person ability, reliability and separation, unidimensionality, response category functions, and differential item functioning (DIF) (Boone, 2016; Peeters et al., 2025).

The use of the Rasch model in educational research has grown significantly in recent years to develop and validate instruments for diverse constructs, including digital literacy (Avinç & Doğan, 2024), digital citizenship (Prasetyo et al., 2022), inductive reasoning (Soeharto & Csapó, 2022), and physics analytical skills (Nurussaniah et al., 2025). It has likewise been applied to 21st-century skills and learning environments, such as computational thinking dispositions (Sovey et al., 2022) and artificial intelligence literacy (Dong et al., 2025). Sari et al. (2024) combined confirmatory factor analysis and the Rasch Model to develop a numeracy learning environment instrument, underscoring the value of complementary psychometric evidence. Instrument formats are equally varied, covering Likert scales, multiple-choice and two-tier tests, performance-based instruments, attitude questionnaires, and competency scales, as in Bui et al. (2020), Alhadabi and Aldhafri (2021), and Aghekyan (2019).

Despite this growth, the literature remains fragmented, with most studies validating a single instrument in a specific context. Constructs, respondent groups, sample sizes, software, and evaluation indicators differ across studies, so evidence on best practice in Rasch-based validation is scattered and difficult to compare, particularly regarding item fit, person fit, reliability and separation, unidimensionality, response category function, DIF, and decisions to retain, revise, or delete items.

The first gap is that evidence on the characteristics of educational instruments analyzed with the Rasch Model has not been integrated. Primary studies cover digital literacy, creativity, inductive reasoning, teacher–student relationships, numeracy learning environments, physics analytical skills, and AI literacy (Alhadabi & Aldhafri, 2021; Avinç & Doğan, 2024; Dong et al., 2025; Indihadi et al., 2022; Sari et al., 2024; Soeharto & Csapó, 2022), whereas existing systematic reviews of Rasch measurement address other domains: language assessment (Aryadoust et al., 2021) and nursing research (Stolt et al., 2022). Neither addresses educational measurement instruments across diverse constructs, levels, and formats, so no systematic review comprehensively maps the constructs, educational levels, academic fields, respondent characteristics, instrument types, and development designs in Rasch-based educational measurement research.

The second gap concerns variation in the reporting of Rasch-based validity and reliability evidence. Some studies emphasize reliability and separation, while others focus on item fit, unidimensionality, DIF, or rating scale function, so the terms “valid” and “reliable” are not always backed by equivalent psychometric evidence. Instrument quality cannot rest on a single indicator; it requires integrated evidence on internal structure, model fit, response category appropriateness, person–item separation, and equivalence of item functioning across groups.

The third gap is the limited assessment of the methodological soundness of Rasch applications. Many publications report Rasch results without transparently explaining model selection, decision criteria, handling of misfitting items, testing of unidimensionality, DIF evaluation, or implications for instrument revision. A systematic review is therefore needed that not only counts studies and fields of application but also evaluates reporting quality and the completeness of the psychometric evidence in each study.

Thus, this study aims to systematically analyze the development of the use of the Rasch Model in testing the validity and reliability of educational measurement instruments. Specifically, this study seeks to answer the following questions: (1) what are the characteristics of the publications and educational instruments analyzed using the Rasch Model; (2) which psychometric indicators are most frequently used in validity and reliability testing; and (3) to what extent have previous studies reported evidence of model fit, reliability, unidimensionality, response category function, and DIF.

Methods

This study employed a systematic literature review (SLR) design to identify, select, evaluate, and synthesize empirical evidence on the use of the Rasch Model in analyzing the validity and reliability of educational measurement instruments. The review was designed and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines so that identification, screening, eligibility

assessment, and inclusion were transparent and replicable (Page et al., 2021). To address RQ3, each included article was evaluated against an explicit completeness rubric covering five categories of psychometric evidence: (a) item and person fit statistics, (b) reliability and separation indices, (c) unidimensionality and local independence, (d) response category functioning, and (e) differential item functioning (DIF). Studies reporting only fit and reliability evidence were classified as basic, those additionally reporting one further category as intermediate, and those addressing all five categories as comprehensive; this three-level classification underpins the RQ3 synthesis reported below.

The literature search covered three databases selected for their broad coverage of reputable international journals and support for field-specific search syntax: ScienceDirect, ProQuest, and Google Scholar. It covered publications from 2020 to 2026. Boolean operators “AND” and “OR” combined search terms restricted to titles, abstracts, and keywords, using English keywords to retrieve English-language literature and Indonesian keywords to retrieve Indonesian-language literature, as summarized in Table 1. The records identified, duplicates removed, records screened, exclusions with reasons, and articles included are summarized in the PRISMA 2020 flow diagram (Figure 1).

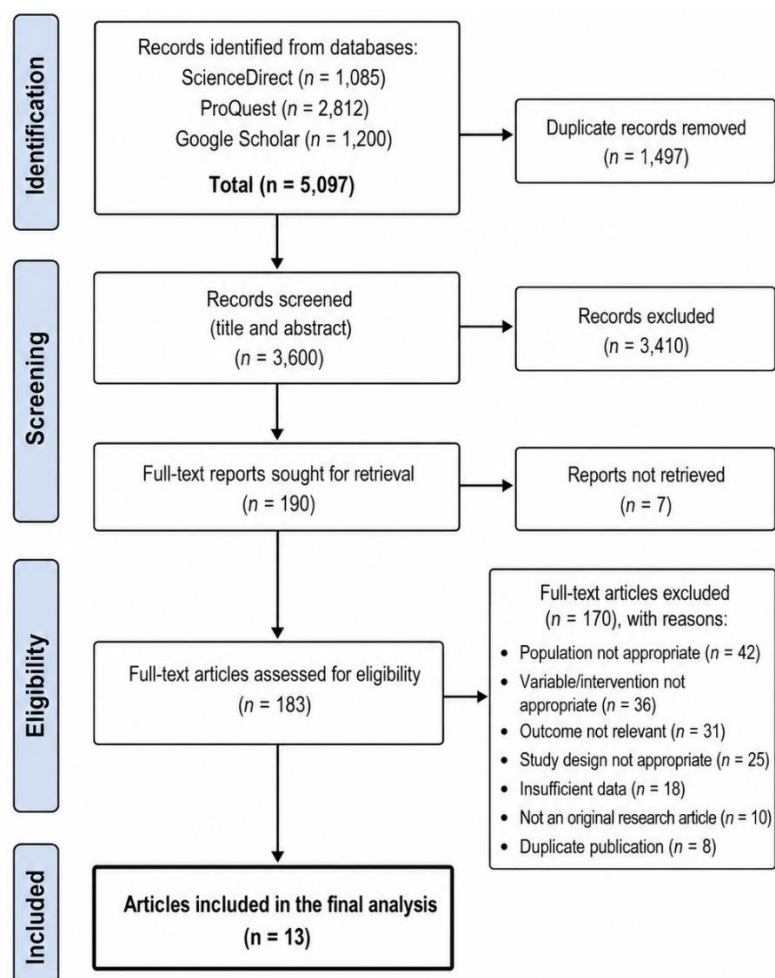


Figure 1. PRISMA 2020 Flow Diagram

Articles were included if they (1) were empirical studies published in peer-reviewed journals between 2020 and 2026; (2) were conducted in formal or nonformal education, learning, teacher training, higher education, professional education, or learning evaluation; (3) involved the development, adaptation, validation, evaluation, or refinement of educational measurement instruments such as tests, scales, questionnaires, rubrics, and performance assessments; (4) used the Rasch Model or a Rasch family model, including the Rating Scale Model, Partial Credit Model, or Many-Facet Rasch Model; and (5) reported at least one psychometric finding on validity, reliability, model fit, unidimensionality, response category function, DIF, or separation.

Articles were excluded if they only discussed the concepts of the Rasch Model without empirical results, used an item response theory approach other than Rasch without a clear Rasch analysis, or focused on clinical and health instruments unrelated to education. Editorials, opinion pieces, conference proceedings and abstracts, books, book chapters, theses, dissertations, and articles without available full text were also excluded. The reason for exclusion at each screening stage was recorded and is reported in the PRISMA flow diagram (Figure 1).

Result

The database search identified a total of 5,097 records: 1,085 from ScienceDirect, 2,812 from ProQuest, and 1,200 from Google Scholar (Table 1). After removing duplicates and screening titles and abstracts against the criteria above, 183 records were retained for full-text assessment; 170 of these were subsequently excluded for the reasons summarized in the PRISMA flow diagram (Figure 1). A final total of 13 articles met all inclusion criteria and were retained for narrative synthesis.

Table 1. Database Search Strategy and Initial Records Retrieved (2020–2026)

Database	Search Keywords/String	Initial Records Retrieved
Science Direct	"validity" AND "reliability" AND "Rasch model"	1,085 articles
Proquest	"validity" AND "reliability" AND "Rasch model" AND "learning"	2,812 articles
Google Scholar	"validity" AND "reliability" AND "Rasch model" AND "education" AND "learning"	1,200 articles

Consistent with the criteria in the Method section, only peer-reviewed journal articles were retained at every screening stage. The resulting set of included studies is summarized in Table 2.

1. RQ1 Publication Characteristics, Educational Context, and Diversity of Rasch Model-Based Instruments.

The first finding indicates that the use of the Rasch Model in validating educational instruments has expanded across a wide range of constructs, educational levels, and instrument formats. The corpus spans 2020–2025, with a broadening focus from learning assessment, teacher–student relationships, creativity, and learning environments toward contemporary constructs such as digital literacy, artificial intelligence literacy, critical thinking, academic involution, and mental health. This shift shows that the Rasch Model is increasingly used to meet the need for measurement that is more dynamic, multidimensional, and sensitive to technological developments (Alhadabi & Aldhafri, 2021; Avinç & Doğan, 2024; Dong et al., 2025; Yang et al., 2023).

Table 2 summarizes the key characteristics of the studies identified in the corpus for this theme, including the country, educational level, construct or instrument, and psychometric indicators reported in this review.

Table 2. Summary of Studies Included in the Review

Author(s) & Year	Country	Educational Level	Construct / Instrument	Key Psychometric Indicators Reported in This Review
Hayat (2022)	Indonesia	Undergraduate	Basic Statistics Test	Unidimensionality with Principal Component Analysis Residual (PCAR); item parameter estimates and fit statistics; person ability estimates and reliability; test information function
Alhadabi & Aldhafri (2021)	Oman	Middle school	Student–Teacher Relationship Scale	Unidimensionality with Rasch Principal Component Analysis (PCA), Infit-Outfit MNSQ, ZSTD, Pt-Mean Correlation, Item separation and reliability, Person separation and reliability.
Astuti et al. (2024)	Indonesia	Teachers (in-service)	Ethnomathematics-based numeracy teaching belief	Item/person fit (MNSQ 0.4–1.5, ZSTD –2/+2); person reliability .89; item reliability .98; person separation 4.19; item separation 6.55
Avinç & Doğan (2024)	Turkey	Secondary	Digital literacy scale	Unidimensionality (EFA+CFA+Rasch); local independence ($r < 0.70$); category function (undecided category malfunction)
Bui et al. (2020)	Vietnam	Higher education	Learning-quality questionnaire for creativity development	Mean scores; item difficulty; Infit-outfit MNSQ; Pt-mean correlation; reliability; cronbach’s alpha; separation index

Author(s) & Year	Country	Educational Level	Construct / Instrument	Key Psychometric Indicators Reported in This Review
Che Lah et al. (2022)	Malaysia	Higher education	Online Social Learning Environment Instrument (OSLEI)	Item/person fit (MNSQ 0.4–1.5, ZSTD –2/+2)
Dong et al. (2025)	China	Higher Education	Multidimensio nal AI literacy scale	Item fit (MNSQ 0.6–1.4); Wright map; response probability curves
Kassiaver a et al. (2024)	Indonesia	Secondary (physics)	Two-tier critical thinking test	Infit-outfit MNSQ; standard error item and person; standard deviation item and person; reliability item and person; chi square ($p < 0.5$)
Davis & Boone (2021)	United States	Higher education	Other- directed, Lighthearted, Intellectual, and Whimsical (OLIW) adult playfulness scale	Unidimensionality with Principal Component Analysis Residual (PCAR); Infit-outfit MNSQ; Pt-Mean Correlation, Item separation and reliability, Person separation and reliability.
Wang et al. (2025)	Spain and China	Primary	Depression, Anxiety, and Stress Scale- 21 (DASS-21)	Rasch (difficulty, point biserial, infit- outfit MNSQ), Confirmatory Factor Analysis (X^2 , CFI, NNFI, RMSEA, SRMR).
Sari et al. (2024)	Indonesia	Secondary	Numeracy learning- environment instrument	Confirmatory Factor Analysis combined with Rasch analysis
Noben et al. (2021)	Netherla nds	Higher Education	International Comparative Analysis of Learning and Teaching (ICALT) instrument	Unidimensionality, infit-outfit MNSQ, Confirmatory Factor Analysis (CFA)
Yang et al. (2023)	China	Higher education	Academic involution scale	Item fit (MNSQ 0.75–1.33); item reliability .99; person reliability .90; item separation 8.32; person separation 3.05; unidimensionality (PCAR 47.3%/12.8%); 5-point category function; DIF (2 items, education background)

Rasch-based research in the corpus spans Indonesia, Malaysia, China, Türkiye, Vietnam, the United States, the Netherlands, Spain, and Oman, and is widely used to adapt or develop instruments suited to local language, culture, and curriculum. Studies on basic statistics in Indonesia (Hayat, 2022), academic involution in China (Yang et al., 2023), online social learning environments in Malaysia (Che Lah et al., 2022), and digital literacy in Türkiye (Avinç & Doğan, 2024) show that Rasch validation is closely tied to contextualized measurement. By educational level, the corpus spans primary, secondary, higher, and in-service teacher education. Primary education is represented by the cross-national validation of the DASS-21 among primary schoolteachers in Spain and China (Wang et al., 2025); secondary education by a student-teacher relationship scale in Oman (Alhadabi & Aldhafri, 2021), a digital literacy scale in Türkiye (Avinç & Doğan, 2024), a two-tier critical thinking test in Indonesian physics classrooms (Kassiavera et al., 2024), and a numeracy learning-environment instrument in Indonesian secondary schools (Sari et al., 2024). Higher education is the most prominent context, covering a basic statistics test (Hayat, 2022), a creativity-oriented teaching-quality questionnaire in Vietnam (Bui et al., 2020), an online social learning environment instrument in Malaysia (Che Lah et al., 2022), an adult playfulness scale in the United States (Davis & Boone, 2021), an AI literacy scale in China (Dong et al., 2025), the ICALT observation instrument among university lecturers in the Netherlands (Noben et al., 2021), and an academic involution scale in China (Yang et al., 2023).

Astuti et al. (2024) extend the model to in-service teachers, measuring ethnomathematics-based numeracy teaching beliefs among Indonesian mathematics teachers. Together, the thirteen studies span constructs from statistics achievement and critical thinking to teacher-student relationships, teaching beliefs, digital and AI literacy, learning environments, adult playfulness, mental health, and teaching quality, underscoring the broad applicability of the Rasch model across educational levels and countries.

Likert-type scales and questionnaires still dominate, particularly for affective, perceptual, and dispositional constructs such as digital literacy (Avinç & Doğan, 2024), adult playfulness (Davis & Boone, 2021), mental health (Wang et al., 2025), teaching beliefs (Astuti et al., 2024), and academic involution (Yang et al., 2023). The Rasch Model is also applied to cognitive tests, such as a basic statistics test (Hayat, 2022) and a two-tier critical thinking test in physics (Kassiavera et al., 2024), and to classroom-observation protocols such as ICALT (Noben et al., 2021), confirming that it is cross-format across polytomous rating scales, dichotomous or graded-response tests, and observer-rated instruments. Sample sizes varied widely, from around 45 respondents (Astuti et al., 2024) to 995 students (Alhadabi & Aldhafri, 2021), with larger and methodologically different frames elsewhere: the DASS-21 study drew on 1,350 Spanish and 2,580 Chinese primary schoolteachers (Wang et al., 2025), and the university teaching-quality study on 203

observed lessons (Noben et al., 2021). Sample size must thus be considered against the purpose of the analysis, the number of items, respondent heterogeneity, and generalization needs: small samples can still provide preliminary diagnostic information on item functioning, while larger samples allow testing of parameter stability and cross-group differences (Alhadabi & Aldhafri, 2021; Astuti et al., 2024; Dong et al., 2025).

Overall, the corpus of studies summarized in Table 2 reveals three main patterns. First, the Rasch model has been applied to a variety of constructs spanning the cognitive, affective, social, digital, and pedagogical domains. Second, its application extends to instruments such as scales, questionnaires, objective tests, and two-level tests. Third, studies tend to be oriented toward local contexts and specific populations; therefore, the validity of the results must be understood as evidence related to the characteristics of the respondents and the educational environment in which the instruments are used.

2. RQ2. Psychometric Indicators in Testing Validity and Reliability Using the Rasch Model.

Rasch-based validation studies generally use a set of psychometric indicators to assess item quality, response patterns, measurement reliability, and the instrument's ability to separate respondents and items. Item and person fit statistics, Infit and Outfit Mean-Square (MNSQ) together with their standardized counterparts (ZSTD), are the most consistently reported indicator, appearing in virtually every Rasch-based study (Alhadabi & Aldhafri, 2021; Astuti et al., 2024; Bui et al., 2020; Che Lah et al., 2022; Davis & Boone, 2021; Dong et al., 2025; Hayat, 2022; Noben et al., 2021; Sari et al., 2024; Wang et al., 2025). The acceptance band varies: MNSQ of 0.5–1.5 with ZSTD between –2 and +2 is the most common convention (Che Lah et al., 2022; Davis & Boone, 2021; Noben et al., 2021), Dong et al. (2025) tighten the range to 0.6–1.4, Sari et al. (2024) use 0.4–1.5, and Alhadabi and Aldhafri (2021) apply a more liberal 0.6–1.7 (and even up to 2.0) band, reflecting the polytomous, affective nature of their scale. The point-measure correlation is the near-universal companion to fit statistics, serving as the Rasch analogue of item discrimination, with thresholds typically >0.3 (Astuti et al., 2024; Che Lah et al., 2022); Bui et al. (2020) report a range of 0.312–0.757 across their 43-item instrument.

Person and item separation and reliability (Wright & Stone, 1979) form the second most prevalent indicator family, benchmarked against Fisher's (2007) or Bond and Fox's (2015) criteria (person separation ≥ 2.0 /reliability $\geq .80$; item separation ≥ 3.0 –4.0/reliability $\geq .90$). Reported values vary widely: Astuti et al. (2024) report item separation of 6.55 (reliability .98), Sari et al. (2024) 10.67 (.99), and Alhadabi and Aldhafri (2021) 3.79/.94 versus 7.23/.98 across two subscales, while Davis and Boone (2021) flag person separation below the 2.0 benchmark (1.22–1.67) as an unresolved limitation of the OLIW scale, showing that this indicator is also used diagnostically to justify revision. Cronbach's alpha is still co-reported in most Rasch-dominant studies (Bui et al., 2020, $\alpha = .963$; Che Lah et al., 2022, $\alpha = .95$; Astuti et al., 2024, $\alpha = .81$; Dong et al., 2025, $\alpha = .971$; Avinç &

Doğan, 2024, split-half/Guttman coefficients of .971/.962), apparently as a bridging statistic for readers less familiar with Rasch conventions. Only Bui et al. (2020) report the "number of strata" (H) index ($H = 6.442-6.979$).

Unidimensionality testing via Principal Component Analysis (PCA) of standardized residuals is the third major indicator family, present in nearly every multi-item application, typically anchored to raw variance explained by measures $>40-50\%$ and unexplained variance in the first contrast $<15\%$ (or an eigenvalue $<2.0-3.0$) (Fisher, 2007; Bond & Fox, 2015). Reported strength varies: Sari et al. (2024) report 54.4% explained variance with only 1.3% in the first contrast, Astuti et al. (2024) improve from $56.1\%/10.8\%$ to $63.8\%/7.9\%$ after item removal, and Che Lah et al. (2022) from $39.2\%/9.7\%$ to $43.4\%/7.5\%$. By contrast, Noben et al. (2021) accept 24.78% explained variance against a 50% benchmark, and Dong et al. (2025) flag a borderline first-contrast eigenvalue of 5.8 despite acceptable item fit, indicating that PCA of residuals functions in practice less as a strict pass/fail gate than as an interpretive diagnostic.

For polytomous instruments, category-functioning diagnostics (Andrich threshold ordering, category probability curves, and minimum observations per category) form a fourth layer of evidence (Davis & Boone, 2021; Alhadabi & Aldhafri, 2021). The Wright (item-person) map recurs across nearly the entire Rasch subset (Davis & Boone, 2021; Noben et al., 2021; Hayat, 2022; Dong et al., 2025) as the primary visual device for communicating item-difficulty hierarchy, person-item targeting, and floor/ceiling behavior, complementing rather than substituting for the numeric fit and separation indices.

DIF and hybrid CTT/CFA/SEM indices form a less universal but increasingly visible layer, particularly in studies concerned with cross-group or cross-cultural comparability: Noben et al. (2021) apply a DIF-contrast threshold of >0.64 logits, Avinç and Doğan (2024) the Swaminathan and Rogers (1990) >0.60 criterion, Sari et al. (2024) test bias by gender and school status at a $<5\%$ probability threshold, and Wang et al. (2025) combine Rasch-based DIF with multi-group CFA to establish configural, metric, and scalar measurement invariance between Spanish and Chinese teacher samples (ΔCFI , $\Delta RMSEA$, $\Delta SRMR$ criteria). Studies that pair Rasch with CFA/SEM (Avinç & Doğan, 2024; Noben et al., 2021; Sari et al., 2024; Wang et al., 2025) additionally report a classical battery (KMO and Bartlett's test, factor loadings, χ^2/df , CFI, TLI/NNFI, RMSEA, SRMR, GFI/AGFI/NFI, composite reliability, and average variance extracted), indicating a hybrid trend in which Rasch indicators supplement rather than replace classical evidence. Hayat's (2022) guessing adjustment parameter is a unique outlier tied to the cognitive-test context, showing that indicator selection is shaped by instrument type as well as methodological convention.

3. RQ3. Completeness of Psychometric Evidence.

A qualitative synthesis of representative studies indicates that the depth of psychometric reporting still varies. Yang et al. (2023) illustrate that reliability should be interpreted together with other evidence: their scale has an item reliability of 0.99, a person reliability of 0.90, an item separation of 8.32, and a person separation of 3.05, complemented by results on unidimensionality, response category function, and DIF. This is more robust than studies reporting only reliability values or fit statistics.

Unidimensionality is a key assumption of the Rasch Model, requiring that an instrument primarily measure a single latent construct, and it is reported more prominently for Likert-scale instruments than for dichotomous tests. Yang et al. (2023) used PCA of residuals to show that 47.3% of the variance was explained by the Rasch dimension, with 12.8% in the first contrast, supporting the interpretation that the academic involution scale measures a single main construct. Avinç and Doğan (2024) combined EFA, CFA, and Rasch analysis to strengthen unidimensionality evidence for the digital literacy scale, and standardized residual correlations, all below 0.70, supported local independence. Following conventional benchmarks (Boone, 2016; Fisher, 2007), at least 40% variance explained with less than 15% unexplained variance in the first contrast (or an eigenvalue below 2.0–3.0) supports unidimensionality, and residual correlations below 0.70 support local independence. The variance values reported by Yang et al. (2023) meet the first benchmark, and the residual correlations reported by Avinç and Doğan (2024) meet the second. Local independence matters because an instrument may appear unidimensional overall yet contain item pairs too closely related through similar wording, content, or response context.

Studies of cognitive tests in the corpus, such as Hayat (2022) and Kassiavera et al. (2024), place greater emphasis on item and person parameter estimation and reliability (Table 2), reflecting the response characteristics of test items rather than rating scales; internal-structure analysis nevertheless remains important to ensure that ability variation is not driven by item format, content domain, or irrelevant response strategies. Response category function is particularly relevant for Likert or polytomous scales, and not all categories are used consistently. Avinç and Doğan (2024) found that the “undecided” category of the digital literacy scale did not function optimally, with a difference of only 0.99 logits between categories 2 and 3, below their criterion, and recommended merging or removing it so that the scale would have four response categories.

In contrast, Yang et al. (2023) demonstrated better category functioning on a five-point scale: each category had sufficient frequency, the mean score increased monotonically, category cutoff values were below 2.0, and thresholds increased sequentially. Response categories are thus part of a measurement structure whose functionality must be demonstrated, not merely administrative answer choices.

DIF assesses whether an item is harder for groups of respondents with the same level of the latent construct. It has not yet been used consistently, but studies that apply it contribute to understanding instrument equity: Noben et al. (2021) and Avinç and Doğan (2024) applied explicit DIF-contrast criteria, Sari et al. (2024) tested bias by gender and school status, and Wang et al. (2025) combined Rasch-based DIF with multi-group CFA to compare Spanish and Chinese teachers. Yang et al. (2023) identified two items with DIF by educational background (undergraduate versus graduate students), both concerning final-exam strategies that can substantively carry different meanings for the two groups. DIF therefore does not always mean that an item must be removed; researchers must consider whether the difference is substantive, reflects group characteristics, or constitutes bias that threatens score interpretation.

Some studies, particularly single-institution instrument-development studies with limited samples, do not report DIF. Dong et al. (2025), for example, evaluated fit, reliability, the Wright map, and response probability curves but reported neither DIF nor invariance across disciplines, cultures, and demographic groups, and recommended testing generalizability in a broader population. Overall, three levels of reporting completeness emerge: basic reporting focused on fit and reliability; a second level adding unidimensionality, response category functions, or person-item maps; and comprehensive reporting integrating fit, reliability, internal structure, response categories, and DIF. Yang et al. (2023) is the most comprehensive example at the third level, while Avinç and Doğan (2024) also provide strong evidence by combining classical test theory and Rasch analysis.

Discussion

1. The Expanding Reach of Rasch Measurement Across Constructs, Levels, and Instrument Formats

The first research question shows that the Rasch Model is no longer confined to the achievement tests for which it was historically developed but has become a general-purpose tool for constructing measures across the cognitive, affective, social, pedagogical, and digital domains. The same model family supports a basic statistics test (Hayat, 2022), a two-tier critical thinking test in physics (Kassiavera et al., 2024), attitudinal scales such as adult playfulness (Davis & Boone, 2021) and academic involution (Yang et al., 2023), belief measures among in-service teachers (Astuti et al., 2024), and observer-rated protocols such as ICALT (Noben et al., 2021). This flexibility follows from the logic of the model: because persons and items are placed on a common logit scale irrespective of whether responses are dichotomous, polytomous, or observer-generated, it offers a single analytic language for instrument families that classical test theory would treat separately. The chronology is equally informative. Earlier studies concentrate on established constructs such as teaching quality and teacher-student relationships,

whereas the most recent address constructs that barely existed as measurement targets a decade ago, notably artificial intelligence literacy (Dong et al., 2025) and digital literacy (Avinç & Doğan, 2024). For such newly articulated constructs, whose item hierarchies are least well understood, the Rasch Model is attractive because it yields a testable ordering of item difficulty that can be inspected against construct theory rather than merely a coefficient of internal consistency.

A second pattern concerns geography. The corpus is dominated by Asian settings, particularly Indonesia, China, Malaysia, and Türkiye, with a smaller European and North American presence. This is best interpreted not as a claim about where Rasch measurement is most rigorous but as evidence that the model is used for contextualization: studies adapt instruments to local language, curriculum, and culture and then use Rasch diagnostics to verify that the adapted version still functions as intended, as in the DASS-21 comparison between Spanish and Chinese primary schoolteachers (Wang et al., 2025). Read through validity as an accumulation of context-bound evidence (Messick, 1995; AERA, APA, & NCME, 2014), this pattern is methodologically coherent: validation is most active where imported instruments meet unfamiliar populations and parameter invariance cannot simply be presumed.

The variation in sample size, from 45 respondents (Astuti et al., 2024) to combined samples of several thousand teachers (Wang et al., 2025), also deserves careful interpretation rather than a simple ranking by respondent count. Small samples can legitimately support formative, diagnostic work on item wording, category functioning, and gross misfit, whereas parameter stability, measurement invariance, and DIF require substantially larger and more heterogeneous samples. The corpus therefore contains two partially distinct genres, instrument development and instrument confirmation, which should not be judged against a single evidentiary standard. Problems arise only when conclusions from the first genre are reported in the language of the second, that is, when an instrument validated on a small single-institution sample is described as generally valid and reliable without qualifying the population to which that claim extends.

2. Convergence on a Core Indicator Set and Divergence in Decision Criteria

The second research question reveals a literature that has converged on what to report far more successfully than on how to judge what is reported. Infit and outfit mean-squares with their standardized counterparts, point-measure correlations, person and item reliability, and separation indices constitute a de facto core battery in virtually every study. Beneath this uniformity, acceptance bands differ: from the conservative 0.6–1.4 of Dong et al. (2025) through the conventional 0.5–1.5 to the liberal 0.6–1.7, and occasionally 2.0, of Alhadabi and Aldhafri (2021). Wider bands are defensible for polytomous affective scales, where some variability reflects genuine individual differences rather than noise, whereas high-stakes cognitive tests warrant tighter control, but the rationale for a band is rarely stated. The same item may thus be retained in one study and deleted in another,

so comparability depends on decision rules that often remain implicit, echoing the heterogeneity and incompleteness of Rasch reporting documented in adjacent fields (Aryadoust et al., 2021; Stolt et al., 2022).

A similar pattern appears in unidimensionality. PCA of standardized residuals is nearly universal, but its output ranges from a strict admission gate to an interpretive aid. Some studies report values well beyond conventional benchmarks, such as 54.4% of variance explained with 1.3% in the first contrast (Sari et al., 2024), or document improvement after item removal (Astuti et al., 2024; Che Lah et al., 2022); yet Noben et al. (2021) accept 24.78% against a 50% benchmark, and Dong et al. (2025) retain a scale despite a first-contrast eigenvalue of 5.8. Rather than lapses, these are better understood as evidence that PCA of residuals is a diagnostic to be weighed against item content and theoretical structure, particularly for observational and multidimensional instruments where strict unidimensionality is neither expected nor desirable. What the literature lacks is not stricter thresholds but explicit argumentation linking residual structure to construct theory, in line with Boone's (2016) view that Rasch indices are interpretive devices rather than automatic verdicts.

Two further observations follow from the indicator profile. First, the persistence of Cronbach's alpha alongside Rasch reliability (Bui et al., 2020; Che Lah et al., 2022; Astuti et al., 2024; Dong et al., 2025) indicates a transitional reporting culture in which classical coefficients bridge to audiences unfamiliar with logit-based evidence. This is pragmatic, but risky when a high alpha reassures readers about an instrument with weak person separation, as Davis and Boone (2021) transparently acknowledge for the OLIW scale. Second, indicator selection is shaped by instrument type rather than convention alone: rating-scale diagnostics appear only for polytomous formats, and a guessing adjustment parameter only in the cognitive-test context of Hayat (2022). Any reporting standard must therefore be conditional on instrument format rather than a single undifferentiated checklist.

3. A Three-Level Gradient in the Completeness of Psychometric Evidence

The third research question exposes the review's central methodological finding. Under the completeness rubric described in the Method section, the corpus distributes across three levels: basic reporting limited to fit and reliability; intermediate reporting adding unidimensionality, response category functioning, or a Wright map; and comprehensive reporting integrating fit, reliability, internal structure, category functioning, and DIF. Yang et al. (2023) exemplify the comprehensive level, while Avinç and Doğan (2024) reach comparable strength by combining classical and Rasch evidence, including a local-independence check and a documented decision to collapse a non-functioning undecided category. Most studies cluster at the basic and intermediate levels, and the omissions are systematic rather than random: local independence is the least

frequently tested assumption, followed by DIF, followed by formal response category diagnostics.

This gradient matters because the omitted categories are the conditions under which the reported statistics can be believed. Acceptable fit computed on data with locally dependent item pairs overstates reliability, because redundant items inflate consistency without adding measurement information; separation indices computed on disordered or underused thresholds describe a category structure respondents are not actually using; and a logit scale established without a DIF check may not carry the same meaning for the subgroups a study intends to compare, which is precisely the inference most educational applications require. The phrase “valid and reliable,” used without specifying which assumptions were tested, thus conveys less than it appears to. When performed, these checks are far from ritualistic: DIF analysis produced substantive rather than merely statistical insight in Yang et al. (2023), where flagged items concerned exam strategies that plausibly carry different meanings for undergraduate and graduate respondents. The appropriate response was interpretation rather than automatic deletion, which is itself an argument for reporting DIF as a routine diagnostic rather than a pass-fail gate.

Incompleteness should also be read contextually. Studies that omit DIF are disproportionately single-institution development studies with modest samples, for which subgroup analyses would be underpowered and potentially misleading, and Dong et al. (2025) explicitly frame the absence of invariance testing as an agenda for subsequent work. The problem is less that every study should report everything than that the field lacks a shared convention for stating which evidence was not obtained and why, so that readers can calibrate the scope of a validity claim rather than infer it.

4. Implications for Reporting Practice and Instrument Use in Education

Taken together, the three findings support a modest but concrete proposal: Rasch-based validation studies in education would benefit from a format-conditional minimum reporting set accompanied by explicit decision rules. For all instruments, this would include item and person fit with the acceptance band and its justification, person and item reliability and separation, evidence on internal structure with the criterion applied, and a person-item map; for polytomous instruments, category frequencies, average measures, and threshold ordering, together with any decision to collapse categories and its consequences; and for instruments intended to compare groups, administration modes, or cultures, a DIF or invariance analysis with the effect-size criterion stated. Studies should also report what was examined and rejected, including misfitting items removed, iterations performed, and the effect of removal on the remaining parameters, since presenting only a clean final solution conceals the analytic path. Where an element is not feasible, a brief statement of the reason and of the resulting limit on generalization would allow readers to interpret the evidence appropriately.

For instrument users in schools, universities, and educational agencies, the label “valid and reliable” should be read as a question rather than a credential: which population supplied the calibration sample, whether item difficulty is well targeted to the intended respondents, whether the response categories functioned as designed, and whether item functioning has been examined across the groups to be compared. The corpus shows that these questions are answerable within the Rasch framework and frequently lead to instrument revision, category collapsing, or explicit restriction of the intended population. For journals and reviewers, the marginal value of an additional Rasch validation study lies less in reporting another set of acceptable fit statistics than in documenting the assumptions tested and the decisions made, which is what makes cumulative synthesis possible.

5. Limitations

Several limitations qualify these conclusions. First, the search was restricted to three databases and to English- and Indonesian-language publications, which plausibly under-represents Rasch work published in Chinese, Spanish, Portuguese, and Arabic and reinforces the Asian concentration observed. Second, the corpus of thirteen studies supports narrative synthesis but not frequency claims or statistical comparison across subgroups; the patterns reported characterize a purposively screened literature rather than prevalence. Third, the completeness rubric evaluates what studies reported, not what they did, so analyses conducted but unreported are indistinguishable from omitted ones, and the incompleteness documented is at least partly a reporting phenomenon shaped by journal word limits. Fourth, restriction to peer-reviewed journal articles may under-represent instruments with poorer psychometric performance. Fifth, extraction and rubric classification were interpretive, and no formal inter-rater agreement statistic was computed. Future research could address these limits through multilingual and grey-literature searches, meta-analytic synthesis of separation and reliability indices for comparable instruments, and the development and empirical evaluation of a consensus reporting checklist for Rasch analysis in educational research, comparable to those emerging in adjacent fields (Aryadoust et al., 2021; Stolt et al., 2022).

Conclusion

This PRISMA-guided review of thirteen empirical studies identified through a 2020–2026 search shows that the Rasch Model has become a versatile framework for validating educational instruments across cognitive, affective, pedagogical, and digital constructs, converging on a core battery of fit, point-measure correlation, reliability, and separation indices, yet diverging markedly in the decision criteria applied to them and distributing unevenly across three levels of reporting completeness, with local independence, response-category functioning, and differential item functioning most often left untested. Because these less frequently reported elements are the assumptions on which the

credibility of the more commonly reported statistics depends, the field's principal need is not additional single-instrument validation studies but a format-conditional minimum reporting standard with explicit, justified decision rules, so that claims of validity and reliability specify the population, the assumptions tested, and the analytic decisions through which they were reached.

In answer to the three research questions, first, the studies analysed span nine countries, four educational levels from primary schooling to in-service teacher education, and three instrument formats (polytomous rating scales, dichotomous and graded cognitive tests, and observer-rated protocols), with the most recent studies concentrated on emerging constructs such as digital and artificial intelligence literacy. Second, the most frequently used indicators are infit and outfit mean-squares with their standardised counterparts, point-measure correlation, person and item reliability, and the separation index, typically accompanied by principal component analysis of standardised residuals and a person-item map, while Cronbach's alpha persists as a co-reported classical statistic. Third, reporting is uneven rather than uniformly complete: fit and reliability are close to universal and internal-structure evidence is common, but local independence, category diagnostics, and differential item functioning appear only in a minority of studies, most fully in those designed from the outset for cross-group or cross-cultural comparison.

The contribution of this review is therefore diagnostic rather than merely cumulative. Instead of adding a further demonstration that the Rasch Model can validate an educational instrument, it documents how the model is currently used and where the evidentiary chain is thinnest, and proposes a minimum reporting set conditional on instrument format, with explicit and justified decision rules and a transparent account of the analytic path, including items removed and the effect of their removal. Adopting such a convention would make validity claims interpretable across studies and allow future syntheses to compare instruments rather than merely count them.

For researchers, a Rasch analysis should be reported as an argument about score meaning for a specified population, not as a sequence of coefficients that clear conventional thresholds. For those who select and use instruments, the phrase "valid and reliable" warrants four follow-up questions concerning the calibration sample, the targeting of item difficulty, the functioning of the response categories, and the comparability of item functioning across groups. Subsequent work can extend these findings through multilingual and grey-literature searches, meta-analytic synthesis of separation and reliability indices for comparable instruments, and a consensus reporting checklist for Rasch analysis in educational research.

References

- Aghekyan, R. (2019). Measuring High School Students' Science Identities, Expectations of Success in Science, Values of Science and Environmental Attitudes: Development and Validation of the SIEVEA Survey. *Science Education International*, 30(4), 342–353. <https://doi.org/10.33828/sei.v30.i4.12>
- Alhadabi, A., & Aldhafri, S. (2021). A Rasch model analysis of the psychometric properties of the student-teacher relationship scale among middle school students. *European Journal of Educational Research*, 10(2), 957–973. <https://doi.org/10.12973/eu-er.10.2.957>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive Review of Rasch measurement in Language Assessment: Recommendations and Guidelines for Research. *Language Testing*, 38(1), 6–40. <https://doi.org/10.1177/0265532220927487>
- Astuti, E. P., Wijaya, A., & Hanum, F. (2024). Teachers' belief in ethnomathematics-based numeracy learning scale: A Rasch model analysis. *TEM Journal*, 13(2), 992–1006. <https://doi.org/10.18421/TEM132-14>
- Avinç, E., & Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the Rasch model. *Education and Information Technologies*, 29(17), 22895–22941. <https://doi.org/10.1007/s10639-024-12662-7>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315814698>
- Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, 15(4), rm4. <https://doi.org/10.1187/cbe.16-04-0148>
- Bui, T. L. T., Kazarenkov, V. I., & Tran, V. D. (2020). Application of Rasch model to Develop a Questionnaire for Evaluating the Quality of Teaching for Students' Creativity Development. *International Journal of Learning, Teaching and Educational Research*, 19(8), 278–296. <https://doi.org/10.26803/ijlter.19.8.15>
- Che Lah, N. H., Tasir, Z., & Jumaat, N. F. (2022). An evaluation of the online social learning environment instrument (OSLEI) using Rasch model analysis. *SAGE Open*, 12(2), 21582440221104084. <https://doi.org/10.1177/21582440221104083>
- Davis, D. R., & Boone, W. J. (2021). Using Rasch analysis to evaluate the psychometric functioning of the other-directed, lighthearted, intellectual, and whimsical (OLIWI) adult playfulness scale. *International Journal of Educational Research Open*, 2, Article 100054. <https://doi.org/10.1016/j.ijedro.2021.100054>
- Dong, Y., Xu, W., Huang, J., & Yann, K. (2025). Validating and refining a multi-dimensional scale for measuring AI literacy in education using the Rasch model. *Humanities and Social Sciences Communications*, 12, 1317. <https://doi.org/10.1057/s41599-025-05670->

- Fisher, W. P., Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, 21(1), 1095. <https://www.rasch.org/rmt/rmt211m.htm>
- Hayat, B. (2022). Adjustment for guessing in a basic statistics test for Indonesian undergraduate psychology students using the Rasch model. *Cogent Education*, 9(1), Article 2059044. <https://doi.org/10.1080/2331186X.2022.2059044>
- Indihadi, D., Suryana, D., & Ahmad, A. B. (2022). The analysis of construct validity of Indonesian creativity scale using Rasch model. *Creativity Studies*, 15(2), 560–576. <https://doi.org/10.3846/cs.2022.15182>
- Kassiavera, S., Suparmi, A., Cari, C., & Sukarmin, S. (2024). Application of Rasch model in two-tier test for assessing critical thinking in physics education. *Journal of Baltic Science Education*, 23(6), 1227–1242. <https://doi.org/10.33225/jbse/24.23.1227>
- 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>
- Noben, I., Maulana, R., Deinum, J. F., & Hofman, W. H. A. (2021). Measuring university teachers' teaching quality: a Rasch modelling approach. *Learning Environments Research*, 24(1), 87–107. <https://doi.org/10.1007/s10984-020-09319-w>
- Nurussaniah, N., Setyosari, P., Kuswandi, D., & Ulfa, S. (2025). Psychometric validation of an analytical skills test in physics using the Rasch model. *Journal of Baltic Science Education*, 24(3), 522–537. <https://doi.org/10.33225/jbse/25.24.522>
- 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. *Systematic Reviews*, 10(1), 89. <https://doi.org/10.1186/s13643-021-01626-4>
- Peeters, M. J., Cor, M. K., Castleberry, A. N., & Gonyeau, M. J. (2025). Comparing Holistic and Mixed-Approach Rubrics for Academic Poster Quality. *American Journal of Pharmaceutical Education*, 89(4). <https://doi.org/10.1016/j.ajpe.2025.101379>
- Prasetyo, W. H., Sari, B. I., Patmisari, Sa'diyah, H., Naidu, N. B. M., & Prasetyo, E. (2022). Using Rasch model analysis for assessing psychometric properties of digital citizenship in Indonesian students. *Journal of E-Learning and Knowledge Society*, 18(3), 96–110. <https://doi.org/10.20368/1971-8829/1135684>
- Sari, C. M., Ridwan, A., & Sastrawijaya, Y. (2024). Learning Environment of Numeracy: Development and Validation With Confirmatory Factor Analysis and Rasch Model. *Pedagogika*, 153(1), 185–198. <https://doi.org/10.15823/p.2024.153.9>
- Soeharto, S., & Csapó, B. (2022). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, 46, 101132. <https://doi.org/10.1016/j.tsc.2022.101132>

- Sovey, S., Osman, K., & Mohd Matore, M. E. E. (2022). Rasch analysis for disposition levels of computational thinking instrument among secondary school students. *Eurasia Journal of Mathematics, Science and Technology Education*, 18(3), em2088. <https://doi.org/10.29333/ejmste/11794>
- Stolt, M., Kottorp, A., & Suhonen, R. (2022). The Use and Quality of Reporting of Rasch Analysis in Nursing Research: A Methodological Scoping Review. *International Journal of Nursing Studies*, 132, 104244. <https://doi.org/10.1016/j.ijnurstu.2022.104244>
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. <https://doi.org/10.1111/j.1745-3984.1990.tb00754.x>
- Wang, X., Cao, C.-H., Liao, X.-L., Jiang, X.-Y., Griffiths, M. D., Chen, I.-H., Lin, C.-Y., & Malas, O. (2025). Comparing the psychometric evidence of the Depression, Anxiety, and Stress Scale-21 (DASS-21) between Spanish and Chinese primary schoolteachers: insights from classical test theory and Rasch analysis. *BMC Psychology*, 13, Article 450. <https://doi.org/10.1186/s40359-025-02728-7>
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.
- Yang, Y., Peng, Y., Li, W., Lu, S., Wang, C., Chen, S., & Zhong, J. (2023). Psychometric evaluation of the academic involution scale for college students in China: An application of Rasch analysis. *Frontiers in Psychology*, 14, 1135658. <https://doi.org/10.3389/fpsyg.2023.1135658>