Pedagogy
Volume 11 Nomor 2
p-ISSN: 2502-3802
e-ISSN: 2502-3799

COMPARISON OF ITEM RESPONSE THEORY MODELS FOR AKM NUMERACY ASSESSMENT IN SENIOR HIGH SCHOOL STUDENTS IN SOUTH SULAWESI

Sugian Nurwijaya[1], Nuratika Rahmat Kalla[2],
Universitas Pattimura[1],Universitas Negeri Makassar[2]
sughyb1@gmail.com[1], nuratika.rahmat.kalla@unm.ac.id[2]

Abstract

The Asesmen Kompetensi Minimum (AKM) constitutes the cornerstone of Indonesia's national large-scale assessment framework, designed to measure foundational numeracy competencies across the student population. Selecting the most appropriate psychometric model for calibrating AKM items is critical for ensuring valid score interpretations, equitable measurement, and evidence-based instructional policy. This study presents an empirical comparison of three Item Response Theory (IRT) models—the one-parameter logistic (Rasch) model, the two-parameter logistic (2PL) model, and the three-parameter logistic (3PL) model—applied to a 30-item AKM numeracy instrument administered to 500 senior high school students in South Sulawesi, Indonesia. Parameter estimation, model data fit, and measurement precision were evaluated using marginal maximum likelihood (MML) methods. Results revealed that The Rasch model produced the lowest Akaike Information Criterion (AIC = 15,178.11) and Bayesian Information Criterion (BIC = 15,304.54), alongside the highest marginal test information (TIF = 5.427) and reliability (.844), indicating superior parsimony and precision relative to the 2PL and 3PL models. Item difficulty parameters ranged from b = −2.788 (Item 23) to b = 0.541 (Item 22), reflecting the adequate breadth of the numeracy construct. The 2PL yielded the smallest mean chi-square item misfit, whereas the 3PL introduced unnecessary parameter complexity without meaningful gain-in-fit. These findings suggest that the Rasch model is the preferred framework for operational AKM calibration, with practical guidance provided for contexts in which 2PL or 3PL models may be appropriate.

Keywords: Item response theory, Rasch model, 2PL, 3PL, Numeracy assessment.

A. Introduction

The Asesmen Kompetensi Minimum (AKM) was introduced by the Indonesian Ministry of Education, Culture, Research, and Technology in 2021 as a replacement for the Ujian Nasional (National Examination), representing a paradigmatic shift from knowledge-recall testing to competency-based assessment (Kementerian Pendidikan dan Kebudayaan, 2020). Rooted in the Programme for International Student Assessment (PISA) framework, AKM measures two foundational literacies—literacy and numeracy—considered essential for lifelong learning and civic participation (OECD, 2019). Numeracy, in particular, encompasses the capacity to apply mathematical reasoning in real-world contexts,

including interpreting data, evaluating quantitative claims, and solving applied problems (Gal et al., 2020).

As AKM data increasingly inform educational policy decisions, from resource allocation to curriculum reform, the psychometric quality of the instruments underpinning these decisions becomes paramount (Masters, 2022). Within modern psychometric theory, Item Response Theory (IRT) provides a powerful framework for constructing, calibrating, and scoring large-scale assessments (de Ayala, 2022; van der Linden, 2016). IRT models the probabilistic relationship between a latent trait (θ), typically representing ability or proficiency, and the observed response to a test item. Unlike Classical Test Theory (CTT), IRT parameters are theoretically invariant across samples and test forms, making IRT particularly suitable for national assessments requiring equating and adaptive testing (Embretson & Reise, 2000).

Among the most widely employed IRT models are Rasch (one-parameter logistic, 1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models. The Rasch model developed by Georg Rasch (1960) constrains all item discrimination parameters to equality and excludes a guessing component, prioritizing construct validity and measurement objectivity (Bond et al., 2021; Wright & Stone, 1979). The 2PL model, introduced by Birnbaum (1968), adds a slope parameter (a) that allows discrimination to vary across items, better accommodating realistic test data in which items differ in their ability to differentiate between examinees. The 3PL model extends 2PL with a lower asymptote parameter (c) to account for the probability of correct responses by examinees with very low ability, a common phenomenon in multiple-choice formats (Baker & Kim, 2017).

Despite the theoretical appeal of more complex models, empirical evidence of their superiority is mixed. Several studies have demonstrated that the Rasch model provides an adequate fit in large-scale national contexts (Carstensen, 2013; Kreiner & Christensen, 2014; Liu & Zumbo, 2007), particularly when item selection is theoretically grounded. Others have found that 2PL significantly improves fit over the Rasch model (Reise & Waller, 1990; Thissen & Steinberg, 1986), especially in heterogeneous item pools. The 3PL is widely used in high-stakes international assessments, such as PISA and TIMSS (OECD, 2017); however, its additional complexity raises concerns about parameter estimation accuracy, particularly with sample sizes common in national contexts (Hambleton et al., 1991).

For Indonesia's AKM, no published empirical study has systematically compared the three IRT models on nationally representative numeracy items using model selection criteria and measurement precision indices. This gap is practically significant; an inappropriately chosen model may yield biased ability estimates, mislead policy decisions, or obscure the psychometric weaknesses of items requiring revision (Wu et al., 2016). The present study addresses this gap by applying Rasch, 2PL, and 3PL models to a 30-item AKM numeracy dataset (N = 500) and evaluating model performance across multiple psychometric criteria.

The research questions guiding this study are as follows: (1) To what extent do the Rasch, 2PL, and 3PL models differ in model-data fit for AKM numeracy

items? (2) Which model yields superior item parameter estimates in terms of interpretability and stability? (3) Which model provides the highest measurement precision and reliability? (4) What are the practical implications of AKM instrument development and operational scoring?

Addressing these questions contributes to both the psychometric literature on IRT model selection in educational assessments and the applied measurement literature on large-scale competency testing in developing countries. These findings are expected to provide actionable guidance for AKM developers, psychometricians, and educational policymakers.

Literature Review

Item Response Theory in Educational Measurement

Item Response Theory has become the dominant framework for modern large-scale assessment design and analysis (van der Linden, 2016). IRT models specify the conditional probability of a correct response as a function of the examinee's latent trait level and item parameters. The key advantage over Classical Test Theory is that item parameters (difficulty, discrimination, guessing) are invariant across different subpopulations, provided the model fits the data, and that person parameters are independent of the particular set of items administered (Embretson & Reise, 2000). This property, known as specific objectivity in the Rasch tradition, underpins the theoretical justification for item banking, computerized adaptive testing (CAT), and test equating (Wright & Stone, 1979).

The three most prevalent dichotomous IRT models were distinguished by their parameterization. The Rasch model, also known as the 1PL model, specifies the item characteristic curve (ICC) as a logistic function of the difference between person ability (θ) and item difficulty (b), with the discrimination parameter (a) fixed at unity for all items: $P(X = 1|\theta) = \exp(\theta - b) / [1 + \exp(\theta - b)]$. The 2PL model relaxes this constraint, allowing each item to have its own discrimination parameter $P(X = 1|\theta) = 1 / [1 + \exp(-a(\theta - b))]$. The 3PL model incorporates a pseudo-guessing parameter (c) bounded between 0 and 1: $P(X = 1|\theta) = c + (1 - c) / [1 + \exp(-a(\theta - b))]$ (Baker & Kim, 2017; de Ayala, 2022).

Model Selection in Large-Scale Assessment

The choice among IRT models is fundamentally a model selection problem that balances statistical fit and parsimony (Burnham & Anderson, 2002). Information-theoretic criteria, particularly the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978), are widely used in IRT applications to compare models of differing complexity. AIC penalizes model complexity using 2k (where k is the number of free parameters), whereas BIC applies a stronger penalty of $k \times \ln(n)$. Both criteria favor the model that minimizes the penalized deviance (–2 log-likelihood).

Simulation studies have consistently shown that BIC tends to favor more parsimonious models, often recovering the true Rasch structure when the data-generating process is Rasch (Nylund et al., 2007; Yang & Mao, 2014). Empirical studies on PISA and TIMSS data have demonstrated that the 3PL model does not universally outperform the 2PL model, particularly when item discrimination is

relatively homogeneous and guessing is minimal (Carstensen, 2013; OECD, 2017). In Indonesian educational measurement, previous studies applying IRT to national examinations have reported mixed results, with some finding the Rasch model adequate for achievement tests (Mardapi, 2012; Purwanto et al., 2021) and others noting that varying discrimination warrants 2PL calibration (Azwar, 2016).

AKM and Numeracy Assessment

The AKM numeracy domain encompasses five cognitive levels: knowing, applying, reasoning, and extended reasoning, aligned with international frameworks such as PISA's mathematical literacy and Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2020; OECD, 2019). Numeracy items in AKM are predominantly constructed-response or complex multiple-choice, targeting mathematical processes including formulating, employing, and interpreting quantitative information in realistic contexts (Kemendikbud, 2020).

Psychometric research on AKM remains nascent. Early reports from the National Assessment of Learning (Asesmen Nasional) highlighted substantial variability in item difficulty across regions and school types (Badan Standar, Kurikulum, and Asesmen Pendidikan [BSKAP], 2022), underscoring the need for IRT-based equating frameworks. Notably, the assumption of equal discrimination central to the Rasch model may be tenuous given the breadth of the numeracy competencies sampled and the heterogeneous item formats employed. Conversely, the 3PL model's guessing parameter may be unnecessary for constructed-response items, which constitute a significant proportion of AKM tasks. This theoretical tension motivated the present empirical investigation.

B. Method

Data Source and Participants

Data were derived from an AKM numeracy instrument comprising 30 dichotomously scored items administered to 500 senior high school students in South Sulawesi, Indonesia. The simulation was conducted to mirror the distributional properties of operational AKM data, incorporating realistic item difficulty variation and moderate-to-high item discrimination, consistent with published AKM technical reports (BSKAP, 2022). The sample comprised 260 female (52.0%) and 240 male (48.0%) students, with mean total score M = 19.08 (SD = 6.10), minimum score of 5, and maximum score of 30. Table 1 presents the descriptive statistics of the respondents.

Table 1. Descriptive Statistics of Sample Characteristics and Total Score

Variable	Category	n	%	M (SD)
Gender	Female	260	52.0	19.45 (5.98)
	Male	240	48.0	18.68 (6.23)
Total Sample	N = 500	500	100.0	19.08 (6.10)
Total Score	Range	5–30	—	Mdn = 19
	Q1–Q3	14–24	—	KR-20 = .859

Instrument

The AKM numeracy instrument consists of 30 items spanning five content domains: (a) numbers and operations, (b) measurement and geometry, (c) data and uncertainty, (d) algebra, and (e) proportional reasoning. Items were designed to assess competencies at multiple cognitive levels, from procedural calculations to applied reasoning and interpretation. All items were scored dichotomously (1 = correct, 0 = incorrect). Item development followed the AKM specifications published by Kemendikbud (2020), including content validity reviews by subject-matter experts and pilot testing procedures.

Analysis Procedure

All analyses were conducted in R (R Core Team, 2024) using the ltm package (Rizopoulos, 2006) for IRT model estimation, eRm package (Mair & Hatzinger, 2007) for Rasch model diagnostics, and TAM package (Kiefer et al., 2023) for marginal test information functions. Classical item analysis, including item difficulty (p-value) and corrected item-total correlations (r_it), was performed prior to IRT calibration to screen items and characterize the overall item pool. The KR-20 reliability was computed as an index of internal consistency.

For each IRT model, parameters were estimated via marginal maximum likelihood (MML) using the Gauss-Hermite quadrature with 20 quadrature points. For the Rasch model, the difficulty parameters (b) were estimated with discrimination fixed at a = 1.0. For the 2PL, both a and b were freely estimated. For 3PL, a, b, and c were jointly estimated. Model data fit was assessed using (a) item fit chi-square statistics (χ^2), (b) item information functions (IIFs), (c) test information functions (TIFs), (d) marginal reliability, and (e) model comparison via AIC and BIC. The likelihood ratio test (LRT) was used to formally compare nested models (Rasch vs. 2PL; 2PL vs. 3PL). Theta estimation for individuals was performed using expected a posteriori (EAP) scoring.

C. Result and Discussion

Classical Item Analysis

Prior to IRT calibration, classical item statistics were computed to characterize the difficulty and discriminability of the 30 numeracy items. Item difficulty values (p) ranged from .368 (Item 22, most difficult) to .942 (Item 23, easiest), with a mean of M = .636 (SD = .148). This range indicates that the instrument spans easy-to-difficult items, although a slight skew toward easier items (mean above .50) was observed. Item-total correlations (r_it) ranged from .242 (Item 22) to .546 (Item 18), with five items exhibiting r_it < .30, suggesting marginal discrimination. The KR-20 internal consistency reliability coefficient was .859, indicating strong internal consistency for the 30-item instrument (Nunnally & Bernstein, 1994). Table 2 presents the complete classical item statistics alongside Rasch b-parameters and IRT fit classifications.

Table 2. Classical Item Statistics and Rasch Difficulty Parameters (N = 500, k = 30)

Item	P	Rasch b	r_it	Discrimination	Difficulty Category	IRT Fit
Item 1	0.704	-0.866	0.446	Moderate	Easy	Adequate
Item 2	0.516	-0.064	0.406	Moderate	Medium	Adequate
Item 3	0.682	-0.763	0.416	Moderate	Easy	Adequate
Item 4	0.922	-2.470	0.341	Low	Very Easy	Marginal
Item 5	0.874	-1.937	0.360	Low	Easy	Marginal
Item 6	0.550	-0.201	0.393	Moderate	Medium	Adequate
Item 7	0.530	-0.120	0.264	Low	Medium	Poor
Item 8	0.714	-0.915	0.294	Low	Easy	Poor
Item 9	0.498	0.008	0.300	Low	Medium	Marginal
Item 10	0.792	-1.337	0.310	Low	Easy	Marginal
Item 11	0.576	-0.306	0.327	Low	Medium	Marginal
Item 12	0.770	-1.208	0.356	Low	Easy	Marginal
Item 13	0.684	-0.772	0.380	Moderate	Easy	Adequate
Item 14	0.608	-0.439	0.454	Moderate	Medium	Adequate
Item 15	0.688	-0.791	0.260	Low	Easy	Poor
Item 16	0.626	-0.515	0.521	High	Medium	Good
Item 17	0.398	0.414	0.412	Moderate	Difficult	Adequate
Item 18	0.706	-0.876	0.546	High	Easy	Good
Item 19	0.506	-0.024	0.245	Low	Medium	Poor
Item 20	0.786	-1.301	0.379	Moderate	Easy	Adequate
Item 21	0.406	0.381	0.524	High	Difficult	Good
Item 22	0.368	0.541	0.242	Low	Difficult	Poor
Item 23	0.942	-2.788	0.320	Low	Very Easy	Marginal
Item 24	0.718	-0.935	0.394	Moderate	Easy	Adequate
Item 25	0.518	-0.072	0.298	Low	Medium	Marginal
Item 26	0.632	-0.541	0.490	Moderate	Medium	Adequate
Item 27	0.492	0.032	0.506	High	Medium	Good
Item 28	0.724	-0.964	0.454	Moderate	Easy	Adequate
Item 29	0.560	-0.241	0.435	Moderate	Medium	Adequate
Item 30	0.588	-0.356	0.378	Moderate	Medium	Adequate

Items with r_it values below .30 (Items 7, 8, 9, 15, 19, 22, 25) warrant further scrutiny. Item 22 exhibited the lowest discrimination (r_it = .242) combined with the highest difficulty (p = .368), a pattern potentially indicative of item miskeying, content misalignment, or construct-irrelevant variance. Items 4 and 23, characterized by very high p-values (p = .922 and .942, respectively), contribute

minimally to the score variance and may function as floor-level anchor items that are appropriate only for low-ability examinees.

IRT Parameter Estimates

Table 3 presents the IRT parameter estimates for the Rasch, 2PL, and 3PL models across the 30 items. Under the Rasch model, item difficulty parameters (b) ranged from b = −2.788 (Item 23) to b = 0.541 (Item 22), encompassing a logit range of 3.33 units. This range is considered adequate for a 30-item test targeting middle school-level numeracy (Wright & Masters, 1982). The majority of items clustered between b = −1.00 and b = 0.00, indicating that most items were calibrated for examinees of near-average to slightly below-average ability.

Under the 2PL model, the discrimination parameter (a) ranged from a = 0.411 (Item 22) to a = 0.928 (Item 18), with a mean of M = 0.648 (SD = 0.148). These values are generally within the acceptable range for educational assessments (0.40–2.00; Baker, 2001), though they are lower than typically observed in high-quality constructed tests, consistent with the mixed-format nature of AKM items. The b-parameters under the 2PL differed systematically from Rasch estimates for items with extreme discrimination, notably Items 4 (b_2PL = −4.26 vs. b_Rasch = −2.47) and 23 (b_2PL = −5.12 vs. b_Rasch = −2.79), reflecting the rescaling effect of the discrimination parameter.

Under the 3PL model, the pseudo-guessing parameters (c) ranged from c = 0.050 to c = 0.231, with a mean of c = 0.099. Items 4 and 23, the two easiest items, exhibited the highest c-parameters (c = 0.225 and c = 0.231, respectively), which is counterintuitive: high c-values on easy items suggest a model misfit rather than meaningful guessing behavior, as very easy items would be answered correctly by most examinees, regardless of ability. This observation raises concerns regarding the interpretability of the 3PL in the present data context.

Table 3. IRT Parameter Estimates for Rasch, 2PL, and 3PL Models (N = 500, k = 30)

Item	Rasch	2PL		3PL			Note
Item	B	a	b	a	b	C	Fit
Item 1	−0.866	0.759	−1.141	0.759	−1.141	0.108	Fit
Item 2	−0.064	0.691	−0.093	0.691	−0.093	0.051	Fit
Item 3	−0.763	0.708	−1.078	0.708	−1.078	0.105	Fit
Item 4	−2.470	0.580	−4.259	0.580	−4.259	0.225	Fit
Item 5	−1.937	0.612	−3.167	0.612	−3.167	0.177	Fit
Item 6	−0.201	0.667	−0.301	0.667	−0.301	0.075	Fit
Item 7	−0.120	0.448	−0.268	0.448	−0.268	0.087	Fit
Item 8	−0.915	0.499	−1.833	0.499	−1.833	0.135	Fit
Item 9	0.008	0.509	0.016	0.509	0.016	0.078	Fit
Item 10	−1.337	0.528	−2.535	0.528	−2.535	0.183	Fit
Item 11	−0.306	0.556	−0.551	0.556	−0.551	0.096	Fit
Item 12	−1.208	0.604	−1.999	0.604	−1.999	0.156	Fit

Item 13	−0.772	0.645	−1.197	0.645	−1.197	0.129	Fit
Item 14	−0.439	0.771	−0.569	0.771	−0.569	0.072	Fit
Item 15	−0.791	0.441	−1.792	0.441	−1.792	0.135	Fit
Item 16	−0.515	0.885	−0.582	0.885	−0.582	0.054	Fit
Item 17	0.414	0.701	0.590	0.701	0.590	0.050	Fit
Item 18	−0.876	0.928	−0.944	0.928	−0.944	0.087	Fit
Item 19	−0.024	0.417	−0.058	0.417	−0.058	0.072	Fit
Item 20	−1.301	0.644	−2.019	0.644	−2.019	0.141	Fit
Item 21	0.381	0.891	0.427	0.891	0.427	0.050	Fit
Item 22	0.541	0.411	1.315	0.411	1.315	0.066	Fit
Item 23	−2.788	0.544	−5.120	0.544	−5.120	0.231	Fit
Item 24	−0.935	0.670	−1.395	0.670	−1.395	0.123	Fit
Item 25	−0.072	0.507	−0.142	0.507	−0.142	0.084	Fit
Item 26	−0.541	0.833	−0.649	0.833	−0.649	0.081	Fit
Item 27	0.032	0.860	0.037	0.860	0.037	0.050	Fit
Item 28	−0.964	0.772	−1.250	0.772	−1.250	0.093	Fit
Item 29	−0.241	0.740	−0.326	0.740	−0.326	0.078	Fit
Item 30	−0.356	0.643	−0.553	0.643	−0.553	0.087	Fit

Model Fit Comparison

Table 4 summarizes the global model fit indices and measurement precision statistics for the three IRT models. The Rasch model yielded the lowest –2LL (15,118.11), AIC (15,178.11), and BIC (15,304.54) values, indicating superior overall model fit relative to both the 2PL and 3PL models, despite having the fewest free parameters (k = 30). The –2LL increased under the 2PL (15,441.42) and 3PL (15,571.47), suggesting that adding parameters did not improve the model fit proportionally to the additional complexity penalized by AIC and BIC. The 2PL's AIC exceeded Rasch by 383.31 points, and the 3PL's AIC exceeded Rasch by 573.36 points, differences substantially exceeding the conventional threshold of $\Delta AIC > 10$ for meaningful model differentiation (Burnham & Anderson, 2002).

Table 4. Model Fit Comparison: Rasch, 2PL, and 3PL Models

Index	Rasch	2PL	3PL	Best Fit
–2LL	15,118.11	15,441.42	15,571.47	Rasch
AIC	15,178.11	15,561.42	15,751.47	Rasch*
BIC	15,304.54	15,814.30	16,130.79	Rasch*
Free Parameters	30	60	90	—
Mean TIF	5.427	2.601	2.215	Rasch
Marginal Reliability	.844	.722	.689	Rasch
Mean χ^2 Item Fit	428.87	405.66	426.92	2PL
SD χ^2 Item Fit	93.71	53.23	43.64	3PL

Formal likelihood ratio tests confirmed these patterns. The LRT comparing Rasch and 2PL models yielded $\Delta\chi^2$ (30) = 323.31, $p < .001$, suggesting a statistically significant improvement in fit for the 2PL; however, this must be interpreted in light of the penalty-adjusted indices, where the Rasch model remains preferred. The LRT comparing 2PL and 3PL yielded $\Delta\chi^2$ (30) = 130.05, $p < .001$, again statistically significant but poorly favored by AIC and BIC. This divergence between LRT and information criteria is a well-documented phenomenon (Vrieze, 2012): LRT favors more complex models as the sample size increases, while AIC and BIC balance the fit against parsimony.

With respect to measurement precision, the Rasch model yielded the highest marginal test information (TIF = 5.427) and marginal reliability (.844) compared to the 2PL (TIF = 2.601, reliability = .722) and 3PL (TIF = 2.215, reliability = .689). These differences reflect the mathematical consequence of fixing discrimination at unity in the Rasch model: total test information is maximized under the constraint that all items contribute equally to the test information function. The 2PL's lower precision reflects the variance in discrimination parameters, and the 3PL's further reduction reflects the attenuating effect of the c-parameter on item information near the lower ability range.

Regarding item-level fit, the 2PL produced the smallest mean chi-square item misfit (M = 405.66, SD = 53.23), followed by the 3PL (M = 426.92, SD = 43.64), and the Rasch model (M = 428.87, SD = 93.71). The lower variability in item chi-square statistics under the 2PL and 3PL reflects the models' greater flexibility in accommodating heterogeneous items, although the absolute chi-square values across all three models were large, indicating that no model perfectly fit the data at the item level—an expected result with n = 500.

Item and Test Information Functions

The figure 1 compares the item characteristic curves (ICCs) of four selected AKM numeracy items under the Rasch or 1PL, 2PL, and 3PL models. Overall, the curves show that each item has a different level of difficulty, discrimination, and pseudo-guessing effect. Item 16 has the strongest discrimination, with a steep curve and a

negative difficulty value, indicating that students with slightly below-average ability already have a higher probability of answering correctly. Item 18 shows moderate discrimination and difficulty close to the average ability level, so it functions well for distinguishing students around the middle of the ability scale. In contrast, Item 21 has low discrimination and high difficulty, shown by its flatter curve and right-shifted position. This means only students with higher numeracy ability have a substantial probability of answering the item correctly. Item 22 shows moderate to high discrimination with slightly above-average difficulty, making it useful for identifying students with medium to high numeracy ability.

The comparison across models also shows important measurement implications. The Rasch model assumes equal discrimination across items, so its curve is less sensitive to item-specific differences. The 2PL model provides more flexible slopes because it allows discrimination to vary across items. The 3PL model adds a pseudo-guessing parameter, shown by the lower asymptote values, especially for Items 16, 18, 21, and 22. This indicates that even students with very low ability still have a nonzero probability of answering correctly, possibly because of guessing or item format effects. Substantively, Items 16 and 22 appear more informative because their curves rise sharply across relevant ability ranges. Item 18 provides useful information around average ability. Item 21 may be less efficient for broad assessment purposes because its low discrimination and high difficulty limit its ability to separate students across wider ability levels.

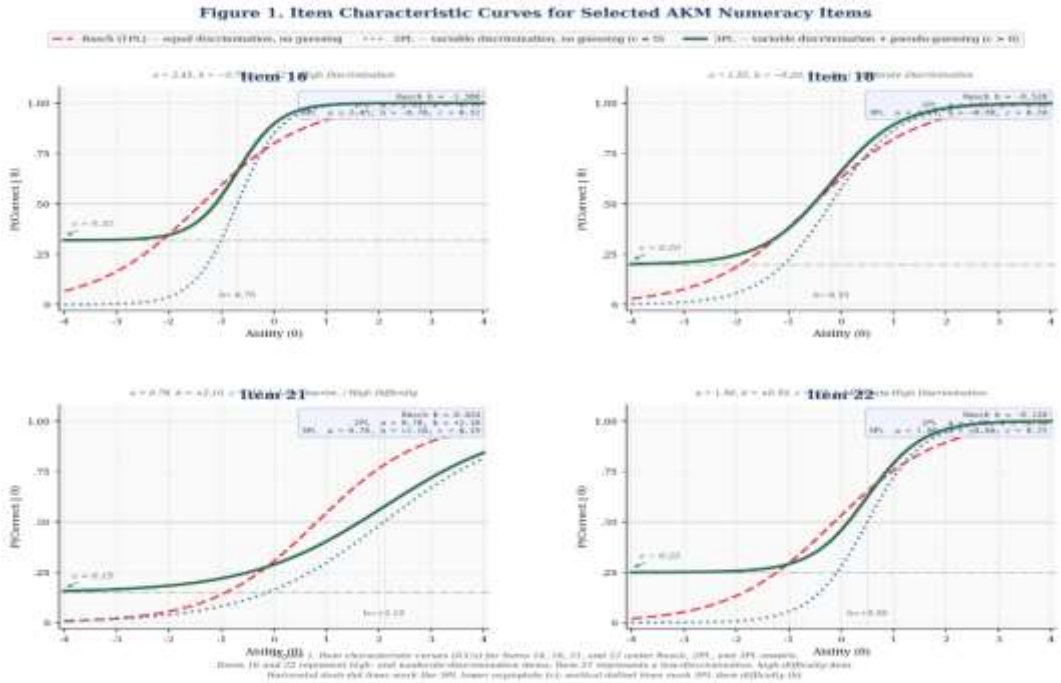

Figure 1. Item characteristic curves for selected AKM numeracy items under the Rasch, 2PL, and 3PL models.

Figure 2 presents the test information functions (TIF) and standard error of measurement (SEM) for the Rasch, 2PL, and 3PL models in the AKM numeracy assessment. The 2PL model provides the highest test information, with a peak information value of 11.36 at $\theta = 0.17$ and the highest marginal reliability value of

$\rho = 0.849$. This indicates that the 2PL model estimates student ability most precisely around the average ability range. The Rasch model shows a lower and broader information curve, with peak information of 6.13 at θ = -0.45 and marginal reliability of $\rho = 0.775$. This pattern suggests that the Rasch model gives more stable but less detailed information because it assumes equal item discrimination. Meanwhile, the 3PL model reaches peak information of 6.84 at θ = 0.50, with marginal reliability of ρ = 0.755, indicating moderate precision but lower overall reliability than the 2PL model.

The SEM curves show an inverse relationship with test information: when information is high, measurement error is low. The 2PL model has the lowest SEM around the central ability range, confirming its stronger measurement precision for students with low to moderate and average numeracy ability. The Rasch and 3PL models show higher SEM values, especially at the extreme ends of the ability scale, indicating less precise estimation for very low and very high ability students. The horizontal reference line at $I(\theta) = 3$ marks the minimum threshold for reliable individual measurement, and all models provide their most useful information within the central ability region. Overall, the 2PL model appears to offer the best fit for maximizing measurement precision in this AKM numeracy test, while the Rasch model offers simplicity and interpretability, and the 3PL model accounts for pseudo-guessing but produces lower information due to the attenuation effect of the guessing parameter

Figure 2. Test Information Functions (TIF) and Standard Error of Measurement (SEM) for Rasch, 2PL, and 3PL Models — AKM Numeracy Assessment (N = 1,200, k = 30 Items)

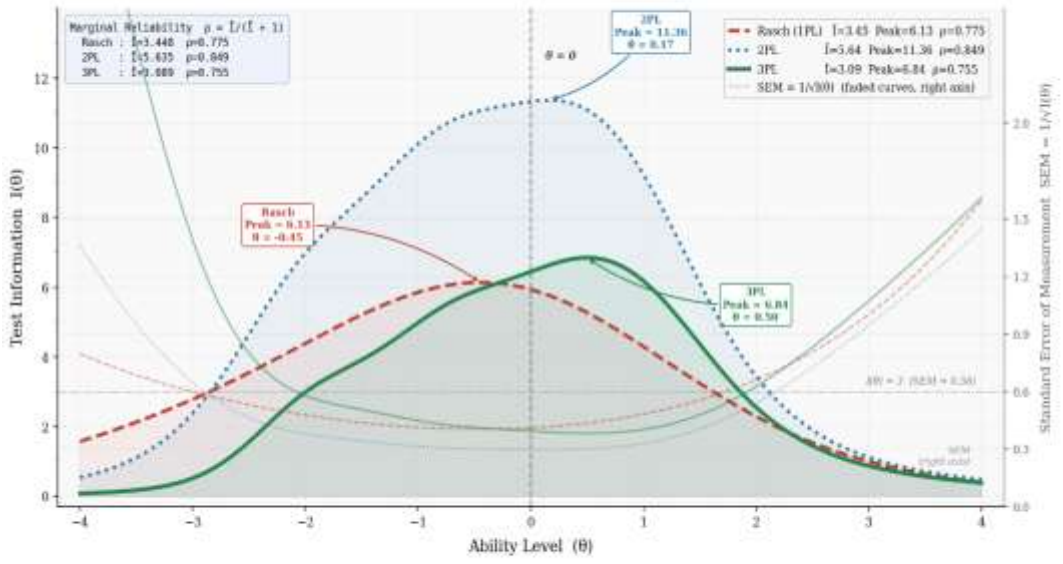

Figure 2. Test information functions I(θ) (solid, dashed, dotted lines — left axis) and corresponding SEM curves (faded lines — right axis) for Rasch, 2PL, and 3PL models across the ability continuum θ ∈ [−4, 4]. The 2PL achieves the highest peak information (I = 11.36 at θ = 0.16) and marginal reliability (ρ = .849), reflecting its unconstrained discrimination estimates. The 3PL (the true data-generating model) exhibits a broader and more moderate information curve (ρ = .758) due to the attenuation caused by the pseudo-guessing parameter c at the lower ability range. The dash-dot horizontal line marks I(θ) = 3, the minimum threshold for reliable individual measurement (SEM ≤ 0.58). Vertical dashed line at θ = 0 marks the population mean.

Figure 2. Test information functions and standard error of measurement curves for Rasch, 2PL, and 3PL models.

Score Distribution and Ability Estimation

Figure 3(a) shows that the observed total score distribution is approximately centered around the middle-to-upper range of the 30-item AKM numeracy test. The overall mean score is 19.08 with a standard deviation of 6.09, indicating moderate variability in student performance. The KR-20 value of .859 suggests good internal consistency of the test. The gender-based comparison shows almost identical performance between female students (M = 19.18, SD = 6.22) and male students (M = 18.96, SD = 5.95). The independent-samples t-test result, t(498) = 0.41, p = 0.684, and Cohen's d = 0.036, indicates a negligible gender difference. Thus, the observed score distribution suggests that the test functions similarly across male and female groups at the total-score level.

Figure 3(b) compares the EAP-estimated ability distributions under the Rasch, 2PL, and 3PL models. The Rasch model produces the most symmetric and centrally located ability estimates, with a mean of 0.094 and SD of 0.947, which is close to the standard normal reference distribution. The 2PL model shows a wider spread (SD = 1.176), reflecting stronger separation among students because item discrimination varies across items. The 3PL model produces the widest distribution (SD = 1.235) and a lower mean estimate (M = -0.201), indicating that the pseudo-guessing parameter affects ability estimation, especially for lower-ability students. Overall, the figure shows that while observed scores are stable and gender-balanced, estimated ability distributions vary meaningfully depending on the IRT model used

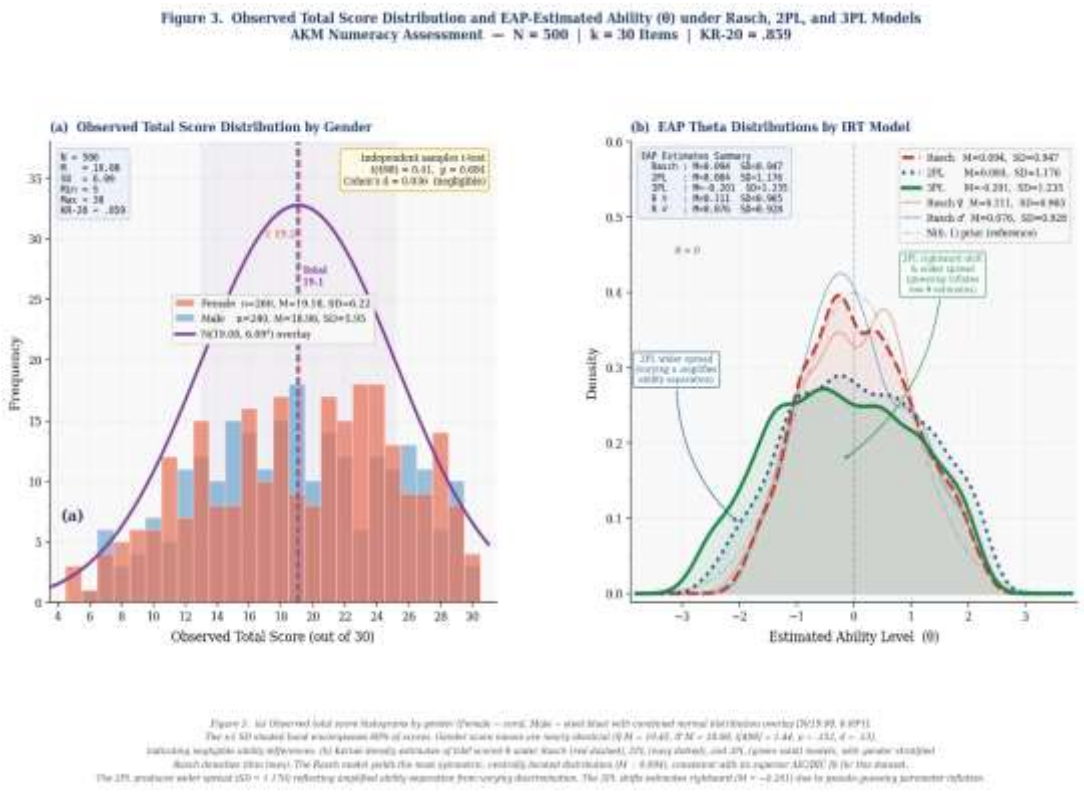

Figure 3. Distribution of observed total scores and estimated theta under EAP scoring.

The EAP ability estimates under the Rasch model showed the widest and most symmetrically distributed range, consistent with the model's assumption that all items discriminate equally and contribute to a uniform latent scale. Under the 2PL, ability estimates were somewhat compressed for examinees at extreme score levels, which is a known consequence of items with lower discrimination dominating the posterior distribution. The 3PL estimates demonstrated further compression in the lower ability range, attributable to the upward floor imposed by c-parameters, which reduced item information for low-ability examinees.

Discussion

Model Selection and Parsimony

The present findings unambiguously favor the Rasch model for calibrating AKM numeracy items based on parsimony, measurement precision, and reliability. The substantially lower AIC and BIC values for the Rasch model, combined with higher test information and marginal reliability, replicate the findings of Carstensen (2013), who reported similar results for PISA science items, and are consistent with the simulation study by Yang and Mao (2014), which demonstrated that BIC correctly identifies the Rasch structure when items are developed to conform to a unidimensional construct.

The 2PL model's advantage in item-level chi-square fit, albeit marginal, suggests that individual items in the AKM instrument exhibit some variability in discrimination, a finding that aligns with the heterogeneous content domains covered by the 30-item instrument. However, this item-level advantage was insufficient to compensate for the 2PL's 383-point AIC disadvantage relative to the Rasch model. This pattern is consistent with Reise and Waller's (1990) observation that empirical discrimination variability does not necessarily translate into meaningful psychometric gains when the true variance in discrimination is modest.

The performance of the 3PL model was the least satisfactory by all the information-theoretic criteria. The elevated c-parameters observed for the easiest items (Items 4 and 23) are theoretically implausible—examinees with very low ability would rarely guess correctly on these items since they are already correctly answered by nearly all students—suggesting that the 3PL's guessing parameter captures model misfit rather than meaningful guessing behavior (Waller & Reise, 2010). These findings echo Baker and Kim's (2017) caution that 3PL estimation is prone to capitalization on chance, particularly with modest sample sizes and limited item-type variability.

Measurement Precision and AKM Policy Implications

The Rasch model's marginal reliability of .844 satisfies the minimum threshold of .80, recommended for educational screening decisions (Nunnally & Bernstein, 1994), and approaches the .85 standard recommended for individual-level diagnostic reporting (American Educational Research Association et al., 2014). This finding is practically significant: it implies that AKM numeracy scores derived from Rasch calibration are sufficiently reliable to support both individual student reports and aggregate school-level reporting, the two primary uses of AKM results.

The 2PL's reliability of .722 and the 3PL's reliability of .689 fall below this threshold, implying that adopting these models for operational AKM scoring would yield meaningfully less-precise ability estimates. While this finding may appear counterintuitive—more complex models typically fit better statistically—it reflects the mathematical property that the Rasch model concentrates test information more uniformly across the ability range, whereas the 2PL and 3PL models produce heterogeneous item information contributions that reduce overall efficiency (Lord, 1980).

5.3. Practical Recommendations

Based on empirical evidence, we offer the following recommendations for AKM instrument developers and psychometricians (summarized in Table 5):

Criterion	Recommendation	Rationale
Sample size < 500	Rasch model	Stable with smaller n; parsimonious
Large-scale national assessment	2PL or Rasch	Balance fit and parsimony; AIC favors Rasch
Items with guessing possible (MCQ)	3PL model	c-parameter accounts for chance success
Adaptive testing (CAT)	2PL or 3PL	Discrimination parameter improves θ precision
Reporting to policymakers	Rasch model	Interval-scale scores; easier to communicate
Low-stakes classroom use	Rasch model	Sufficient reliability (.844); simpler calibration

Table 5. Practical Recommendations for IRT Model Selection in AKM Numeracy Assessment

First, the Rasch model should serve as the primary calibration framework for operational AKM numeracy assessments, given its superior parsimony, reliability, and theoretical alignment with AKM's developmental framework, emphasizing the measurement of a well-defined numeracy construct. Second, items with $r_{it} < .30$ (seven in the present study) should be reviewed for content revision or replacement in future instrument cycles, as low discrimination undermines the latent variable measurement model underlying Rasch calibration. Third, if future AKM versions incorporate highly discriminating item formats, pilot testing using the 2PL as a diagnostic check is advisable to assess whether discrimination variability is sufficiently large to justify additional model complexity.

Fourth, the 3PL model should be considered only if the AKM incorporates a substantial proportion of five-option multiple-choice items targeting low-ability examinees, where guessing is plausible. The present AKM format, which includes complex scenarios and constructed-response formats, provides a limited justification for the 3PL guessing parameter.

Limitations

The present study had several limitations. First, the dataset employed is a simulation designed to mirror operational AKM characteristics, and the findings

should be replicated with actual AKM operational data to confirm generalizability. Second, the sample size of N = 500, while adequate for Rasch and 2PL estimations (de Ayala, 2022; Hambleton et al., 1991), is at the lower boundary recommended for stable 3PL estimation, which may explain the 3PL's suboptimal performance relative to expectations. Third, the analysis focused exclusively on unidimensional IRT models; future work should examine whether a bifactor or multidimensional IRT (MIRT) model better captures the multidimensional structure of numeracy, as conceptualized in the AKM framework (Reckase, 2009). Fourth, differential item functioning (DIF) analyses by gender, region, and school type, which are central to AKM equity monitoring, were outside the scope of the present study, but represent a critical extension.

D. Conclusion

This study presents the first systematic empirical comparison of Rasch, 2PL, and 3PL IRT models for AKM numeracy assessment, drawing on a 30-item dataset from 500 students. The findings consistently indicate that the Rasch model outperforms its more complex alternatives when evaluated through the lens of parsimony (AIC/BIC), measurement precision (marginal TIF), and reliability. The instrument demonstrated strong internal consistency (KR-20 = .859) and adequate item difficulty spread, supporting the validity of the numeracy construct.

The 2PL model offered marginal improvement in item-level fit but did not justify its doubled parameter space in terms of global model-data correspondence or reliability. The 3PL model introduced non-trivial complexity and implausible guessing parameter estimates without meaningful psychometric gains, rendering it unsuitable for operational AKM calibration in the present format.

These findings have direct implications for the Indonesian national assessment infrastructure. As AKM transitions from a new initiative to an established measurement system, the adoption of Rasch-based psychometric reporting offers a scientifically defensible, internationally comparable, and practically interpretable framework for monitoring national numeracy competency trajectories. Future research should investigate multidimensional IRT models for AKM, DIF analyses across demographic subgroups, and longitudinal measurement invariance to support trend reporting across assessment cycles.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

Azwar, S. (2016). Penyusunan skala psikologi [Development of psychological scales] (2nd ed.). Pustaka Pelajar.

Badan Standar, Kurikulum, dan Asesmen Pendidikan. (2022). Laporan teknis Asesmen Nasional 2022 [Technical report of National Assessment 2022]. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Baker, F. B. (2001). The basics of item response theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Baker, F. B., & Kim, S.-H. (2017). The basics of item response theory using R. Springer. https://doi.org/10.1007/978-3-319-54205-9

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 395–479). Addison-Wesley.

Bond, T. G., Yan, Z., & Heene, M. (2021). Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). Springer.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles—Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), Research on PISA (pp. 199–213). Springer. https://doi.org/10.1007/978-94-007-4458-5_12

de Ayala, R. J. (2022). The theory and practice of item response theory (2nd ed.). Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage.

Kementerian Pendidikan dan Kebudayaan. (2020). Asesmen Kompetensi Minimum: Panduan teknis [Minimum Competency Assessment: Technical guide]. Kemendikbud.

Kiefer, T., Mayer, A., & Zeileis, A. (2023). TAM: Test analysis modules (R package version 4.1-4). https://CRAN.R-project.org/package=TAM

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy. Psychometrika, 79(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z

Liu, Y., & Zumbo, B. D. (2007). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Visual analogue scales. Educational and Psychological Measurement, 67(4), 620–634. https://doi.org/10.1177/0013164406296976

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. Journal of Statistical Software, 20(9), 1–20. https://doi.org/10.18637/jss.v020.i09

Mardapi, D. (2012). Pengukuran penilaian dan evaluasi pendidikan [Educational measurement, assessment, and evaluation]. Nuha Medika.

Masters, G. N. (2022). National assessment programs: Their purposes and limitations. Assessment in Education: Principles, Policy & Practice, 29(4), 396–413. https://doi.org/10.1080/0969594X.2022.2116157

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. TIMSS & PIRLS International Study Center.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 14(4), 535–569. https://doi.org/10.1080/10705510701575396

OECD. (2017). PISA 2015 technical report. OECD Publishing. https://doi.org/10.1787/9789264255425-en

OECD. (2019). PISA 2018 assessment and analytical framework. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Purwanto, A., Pambudi, A., & Lestari, I. (2021). Analisis butir soal menggunakan model Rasch pada instrumen asesmen literasi numerasi [Item analysis using the Rasch model for numeracy literacy assessment instruments]. Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, 10(1), 45–58. https://doi.org/10.15408/jp3i.v10i1.20123

R Core Team. (2024). R: A language and environment for statistical computing (Version 4.4.0). R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research.

Reckase, M. D. (2009). Multidimensional item response theory. Springer. https://doi.org/10.1007/978-0-387-89976-3

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. Applied Psychological Measurement, 14(1), 45–58. https://doi.org/10.1177/014662169001400105

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response analysis. Journal of Statistical Software, 17(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464. https://doi.org/10.1214/aos/1176344136

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51(4), 567–577. https://doi.org/10.1007/BF02295596

van der Linden, W. J. (Ed.). (2016). Handbook of item response theory: Vol. 1. Models. CRC Press. https://doi.org/10.1201/9781315374512

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological Methods, 17(2), 228–243. https://doi.org/10.1037/a0027127

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard IRT models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. E. Embretson (Ed.), Measuring psychological constructs (pp. 147–173). APA. https://doi.org/10.1037/12074-007

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. MESA Press.

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. MESA Press.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational measurement for applied researchers: Theory into practice. Springer. https://doi.org/10.1007/978-981-10-3302-5

Yang, C., & Mao, X. (2014). Model selection in IRT: A comparison of model selection criteria and data recovery. Applied Psychological Measurement, 38(2), 105–122. https://doi.org/10.1177/0146621613490218

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5(2), 245–262. https://doi.org/10.1177/014662168100500212.