AFRICAN JOURNAL OF THEORY AND PRACTICE OF EDUCATIONAL ASSESSMENT (AJTPEA)

Hard Print ISSN: 2545-5834
e-Print ISSN: 2992-362X
VOL. 15, JUNE, 2026

COMPARING POLYTOMOUS ITEM RESPONSE THEORY MODELS FOR COGNITIVE LOAD: EVIDENCE FROM PARTIAL CREDIT AND GENERALIZED PARTIAL CREDIT MODELS

https://doi.org/10.83151/g0ry-ck77

Christopher Adah Ocheni[1] (https://orcid.org/0000-0001-6007-2362), **John Joseph Agah**[2] (https://orcid.org/0000-0001-9076-0887), **Anayo David Nnaji**[3] (https://orcid.org/0009-0002-1635-7525), **Basil Chinecherem Ezennadi Oguguo**[2] (https://orcid.org/0000-0002-0723-2248), and **Samuel Uchenna Nwani**[2] (https://orcid.org/0000-0003-3335-2013)

[1]Department of Educational Studies in Psychology, University of Alabama, Tuscaloosa, USA. [2]Department of Science Education, Faculty of Education, University of Nigeria, Nsukka. [3]Department of Educational Evaluation and Counseling Psychology, University of Benin, Benin City, Edo State, Nigeria.

Abstract

The focus of the study was to empirically compare the sensitivities of the partial credit and generalized partial credit polytomous item response theory models for cognitive load. Methodologically, a psychometric analysis research design was adopted in this study. Three research questions guided the study. This study used secondary data collected from a sample of 307 (186 males and 121 females) Nigerian secondary school students through an online survey. The data were collected using the Cognitive Load Rating Scale (CLRS). The data was analysed using descriptive statistics such as mean, standard deviation, kurtosis, and skewness, while marginal and empirical reliability indices were used to establish the internal structure and consistency of the items. The Partial Credit Model (PCM) and Generalized Partial Credit Model (GPCM) were used for the analyses of data. The findings of the study showed, among others, that both models provided acceptable fit estimates; however, the GPCM was a better fit to the data. It was concluded that the CLRS shows sound psychometric properties, with ordered categories, acceptable reliability, and adequate fits for most items. Based on the findings, it was recommended that educational researchers and practitioners should rely on the GPCM for assessing the model fit and other psychometric properties of the CLRS.

Keywords: Cognitive Load, Rating Scale, Partial Credit Model, Generalized Partial Credit, Model Fit, Polytomous IRT, Rasch model

Introduction

The growing social, political, and economic dynamics have steered an increasing role of digital tools and skills in the twenty-first century learning environment in a bid to help students manage their learning demands. Particularly, the density of information in the learning environment, which requires students' perpetual attention, has somehow eroded educators' need to understand the mental demands of instructional content. This challenge could have been exposed due to the sensitivity of the limitations of the human working memory (Miller, 1956; Cowan, 2010), and is heralding a growing research interest in the significance of the cognitive load theory (CLT) among twenty-first-century educators in support of differentiated instruction and adaptive learning technologies (Paas & Ayres, 2014). According to Mittal and Banti (2025), cognitive load reflects the brain's limited ability to handle multiple elements simultaneously, making it essential to manage information presentation to avoid overload. To help students maximize their learning, there is a need to understand students' mental efforts in making useful progress with the learning materials.

Students require different levels of mental effort to successfully utilize information that is available in the learning environment. These mental efforts highlight a crucial psychological construct, described as cognitive load (Sweller, 1988). Sweller defined cognitive load as the total amount of mental effort being used in the working memory during learning tasks, and this is grounded in the limited capacity and duration of the working memory. According to Sweller, the CLT posits that the human cognitive system has limited capacity, and instructional efforts must consider this constraint to optimize learning outcomes. Ouwehand, Lespiau, Tricot, and Paas (2025) defined cognitive load as a dynamic construct influenced by task complexity, learner expertise, and instructional design. In this study, cognitive load refers to the mental effort required by students to process educational information for effective learning.

As a crucial component of educational information, cognitive load presents a unique construct that focuses on the information itself and the way it is presented. Sweller (1988) distinguished between three types of cognitive loads. The intrinsic load, which is the complexity inherent in the material itself; the extraneous load, which is the way information is presented, which can either hinder or help learning; and the germane load, which is the mental effort devoted to processing, constructing, and automating schemas. The cognitive load explains why students often struggle in fast-paced or very complex lessons (Society for Education and Training, 2025), and the role of digital technology in supporting learning. According to Sweller (2020), cognitive load hinges on human cognitive architecture, distinguishing between biologically primary and secondary knowledge, in which the latter requires explicit instruction. Oguguo et al. (2025) argued that the focus is on minimizing extraneous loads while enhancing germane load by taking steps to neutralize the noise components of instruction.

The significance of understanding students' cognitive load in the classroom can help educators to enhance student engagement, improve deeper understanding, and reduce confusion for improved academic performance. Understanding students' cognitive loads can foster the sequencing of curriculum content from simple to complex, so that intrinsic load can be effectively managed (Houichi & Sarnou, 2020). Also, Uzun (2021) suggested that it can significantly impact the choice of teachers' instructional strategies by aligning teaching methods with cognitive architecture to reduce extraneous load. According to the Society for Education and Training (2025), in assessment, the understanding of students' cognitive load can help teachers to design tasks that reflect cognitive processing rather than rote memorization.

Ample studies in the literature suggest the existence of a positive relationship between cognitive load and students' learning outcomes. Al-Sarry et al. (2022) revealed that cognitive load is positively correlated with academic achievement. Achor et al. (2022) revealed a significant impact of cognitive load on basic eight students' achievement in social studies. Evans et al. (2024) revealed that teachers' load-reducing instructional strategies indicated lower cognitive load and were positively associated with relative autonomous motivation, engagement, and achievement. Baiduri et al. (2024) revealed a statistically insignificant positive relationship between cognitive load and students' achievement in mathematics. Akbulut et al. (2025) showed a significant influence of pre-training on cognitive load and achievement in a computer-based learning environment. AL-Nasraween et al. (2025) showed a positive and statistically significant correlation between cognitive load, academic achievement, and test wisdom. This evidence shows that cognitive load plays a major role in the achievement of students. Hence, the crucial need for the present study is to draw attention to the psychometric effectiveness of students' cognitive load measures.

The measures of cognitive load provide a veritable tool for assessing students' cognitive load. Such assessment is a critical step in understanding how students respond to the items of cognitive load measures for valid inferences. As a psychological construct, cognitive load is often measured using the Likert-type scales, as in the present study, which makes the polytomous item response theory (IRT) models increasingly relevant for analysing such data. The polytomous item response theory models have been a veritable tool for providing evidence for analysing measures from a wide range of psychological constructs, such as students' cognitive loads, which take on ordinal responses. According to Dai et al. (2021), polytomous IRT models, like the Partial Credit Model (PCM) and the Generalized Partial Credit Model (GPCM), can be effective in analysing multiple response categories and for providing nuanced insights into item and person-level parameters. Wind (2023) highlighted that polytomous IRT models examine the degree to which ordinal rating scales function in psychometrically useful ways. Such psychometric functions are crucial for understanding and interpreting scores of data generated from such instruments. However, studies have highlighted the importance of selecting an appropriate IRT model when evaluating polytomous item response scales, such as cognitive load scales, as model choice can significantly influence parameter estimates, item functioning, and the interpretation of results (Toland, 2014; Liu et al., 2022).

The Partial Credit Model (PCM) is a fundamental and more prominent polytomous IRT model for analysing the model fit of data obtained from a multi-response scale. The PCM was developed by Masters (1982). It is a Rasch-family model, an extension of the Rasch model to the polytomous response scales. It assumes equal discrimination, which is fixed, usually at 1 across all items. This is particularly useful when the items of an instrument are designed to be equally informative and when simplicity in interpretation is desired. However, the assumption of equality of discrimination may not hold in practice, especially in cognitive load assessments where items may vary in relevance and complexity.

The Partial Credit Model (PCM) is advantageous for its simpler model with fewer parameters, making it easier to estimate and interpret (Embretson & Reise, 2000); suitability for assessments where items are designed to be equally discriminating (Bond & Fox, 2015); and ideal for Rasch-based measurement, where invariance and specific objectivity are desired (Linacre, 2021). The major disadvantages of the Partial Credit Model (PCM) include its assumption of equal discrimination across items, which may not reflect real-world data (Muraki, 1992), and it may underfit data when item discrimination varies significantly (Reise & Waller, 2009). By implication, PCM is more suitably used when the items of the instrument are designed to be equally discriminating; simplicity and interpretability are prioritized; and when the assessment follows Rasch measurement principles.

An attempt to address the limitation of the PCM motivated Muraki (1992) to develop the Generalized Partial Credit Model (GPCM) by relaxing the PCM's assumption of uniform item discrimination for polytomous-item scales. The GPCM allows each item to have its own discrimination parameter, offering greater flexibility and potentially better model fit. This model is especially valuable when items differ in their ability to distinguish between levels of the latent trait being considered, such as cognitive load. The GPCM introduces additional complexity in parameter estimation and interpretation. It requires larger sample sizes and more sophisticated estimation techniques, such as the Expectation-Maximization (EM) algorithm, to produce stable and reliable results (Muraki, 1992). Interestingly, empirical evidence by Muraki (1992) and Furr (2017) suggests that the GPCM may outperform the PCM in terms of fit and predictive accuracy, particularly in assessments involving heterogeneous item structures. However, the increased number of parameters raises concerns about overfitting and model parsimony, especially in small-scale studies.

Consolidating on the weaknesses of the Partial Credit Model (PCM), the Generalized Partial Credit Model (GPCM) has the advantages of being more flexible to account for varying item discrimination (Muraki, 1992); it often provides better model fit and more accurate ability estimates (de Ayala, 2009); and it is suitable for complex assessments with diverse item types and scoring rubrics (Thissen & Steinberg, 1986). The main disadvantages of the Generalized Partial Credit Model (GPCM) include its requirement of additional parameter estimation as a result of its complexity (Embretson

& Reise, 2000), as well as the greater computational demands and potential for overfitting with small sample sizes (Wallmark et al., 2023). As implied, GPCM is best used when the items of the instrument have varying discrimination, when a more flexible model is needed for better fit, and when the test includes a mix of item types and scoring schemes.

Recent applications of PCM and GPCM in analysing the fitness of different measurement instruments have shown a major shift to the need to psychometrically assess the quality of the measurement instrument. Furr (2017) used latent regression techniques to compare PCM and GPCM in educational assessments and found that GPCM provided more accurate person ability estimates when item discrimination varied. Wind (2023) used a real data analysis and a simulation study to systematically explore the sensitivity of rating scale analysis techniques based on the partial credit model (PCM) and the generalized partial credit model (GPCM), and found that both models provided valuable information about the rating scale threshold ordering and precision, yet, the GPCM showed a higher model sensitivity. Harsana et al. (2024) found that the GPCM produced the best fit for reading literacy items when compared with other polytomous Item Response Theory models, such as the Graded Response Model (GRM), Partial Credit Model (PCM), and Nominal Reasons Model (NRM), in Indonesia. These studies suggest the superiority of the GPCM in assessing the fitness of polytomous-response rating scales.

However, some evidence suggests that with some conditions, it is possible to disfavour the GPCM in fitness to polytomous data sets. Contrary to the strengths of the GPCM, Yılmaz (2019) employed the GPCM for assessing the fit of a mixed-format elementary school science test in Turkey, in comparison with the Graded Response Model (GRM) for the open-ended items. The result revealed that the GPCM did not produce better data fit statistics compared to the Graded Response Model. But, Dai et al. (2021) in a simulation study found that the GPCM performed more stably across the lengths of the instrument and it revealed more accurate person parameters in the presence of a large amount of missing data, while the stability of the GRM improves notably as the instrument length increases and the GRM revealed more accurate person parameters when the proportion of missing data is small.

In all educational settings, the accurate measurement of cognitive load is essential for evaluating instructional effectiveness and optimizing learning environments. This is especially important since the model choice can significantly influence the validity and reliability of the polytomous response results. While PCM remains a useful tool for uniform item sets, GPCM can provide enhanced flexibility and precision in more complex measurement scenarios. Despite their widespread use, limited empirical evidence for comparing the performance of PCM and GPCM in the context of cognitive load measurement has become a serious problem that needs to be addressed. More so, this gap raises concerns about model selection and its implications for interpreting cognitive load data. Without a clear understanding of which model offers superior psychometric properties under varying conditions, researchers and practitioners may risk drawing inaccurate conclusions about students' cognitive experiences. These raise the researchers' curiosity about the behaviour of the cognitive load rating scale items, and the performance of the PCM compared to the GPCM on the cognitive load research, where the trade-offs between model simplicity and flexibility are highlighted. This study builds on existing research to comparatively evaluate the psychometric properties of PCM and GPCM in a cognitive load context, aiming to inform best practices in educational measurement. Therefore, the study sought to determine:

1. What is the reliability and internal structure of the cognitive load scale?
2. How well do PCM and GPCM represent the item and person response patterns of the cognitive load rating scale?
3. Which model (PCM or GPCM) provides a more accurate and informative representation of the cognitive load scale based on model-fit indices and item information function?

Methodology

This study employed a methodological psychometric analysis research design to evaluate the measurement properties of the cognitive load rating scale. Psychometric analysis is a research design in which the quantitative properties of tests are systematically analysed to guarantee reliable and valid measurements (Yang & Yao, 2014; Torabi et al, 2020). Specifically, a model comparison approach within the Item Response Theory (IRT) framework was used to compare the Partial Credit Model (PCM) and the Generalized Partial Credit Model (GPCM). This study used secondary data collected from a sample of 307 (186 males and 121 females) Nigerian secondary school students through an online survey. The data were responses to the Cognitive Load Rating Scale (CLRS) designed by Ocheni et al. (under review). The scale comprises 10 items structured on a 4-point Likert scale of strongly agree (SA=4) to strongly disagree (SD=1). The scale has two sections, A and B: While section A captured demographic information such as gender and age, section B contains item statements measuring cognitive load. The scale has three items measuring intrinsic cognitive load, three items on extrinsic cognitive load, and four items measuring germane cognitive load.

Analytical Framework

To evaluate the internal structure of the scale, the descriptive analysis involved mean, standard deviation, kurtosis, and skewness. The internal consistency reliability was estimated using the marginal and empirical reliability indices based on the IRT model. To substantiate the reliability estimate, the Cronbach alpha method was used to evaluate the internal consistency estimate of the CLRS. Generally, reliability estimates between 0.70 and 1.00 are interpreted as acceptable (Tavakol & Dennick, 2011).

Partial Credit Model

The partial credit model (PCM) is a polytomous item response theory (IRT) model. The PCM is employed when the item response categories differ across items, such as the Likert-type items with several response categories (Ocheni & Wind, 2026; Niyirinda et al., 2026). Lopez-Pina et al. (2016) report that while the rating scale model (RSM) permits equal transitions between categories for each item, the PCM allows each item to have different unordered thresholds. Dhyaaldian et al. (2022) noted that in PCM, thresholds do not require the same order as response options (a unique set of thresholds is estimated for each item). This is an indication that in PCM, thresholds are not on the same scale, unlike the RSM. Since the cognitive load rating scale (CLRS) has a Likert-type item structure, this study adopts the PCM, an IRT framework, to evaluate the psychometric properties of CLRS. The PCM uses a log-odds form expressed as:

$$In\left[\frac{P_{ni(xi=k)}}{P_{ni(xi=k-1)}}\right] = \theta_n - \delta_i(-\tau_{ik})$$

Where θ_n = location of person *n* on the construct (i.e., person's ability), δ_i is the difficulty of item *i*, and τ_{ik} denote the difficulty of moving from category *k-1* to category *k* for item *i*.

Generalized Partial Credit Model

While the PCM is a polytomous IRT based on the Rasch model, the generalized partial credit model (GPCM) is a non-Rasch polytomous IRT model for item responses in three or more ordered categories. It is similar to the PCM, but each item has a unique slope parameter (α_1) in addition to the unique thresholds. It is expressed as:

$$P_{ni(x=k)} = \frac{\exp\left(\sum_{k=0}^{x}[\alpha_i(\theta_n-\delta_{ik})]\right)}{\sum_{j=0}^{m}\sum_{k=0}^{j}[\alpha_i(\theta_n-\delta_{ik})]}$$

Where α_i is the discrimination parameter for item *i*, θ_n = location of person *n* on the construct (i.e., person's ability), δ_{ik} is the step difficulty or threshold for moving from category *k-1* to category *k* on item *i*.

The GPCM has the same general interpretation as the PCM, where higher locations generally indicate more-difficult items, and needs to be interpreted together with thresholds and slopes. For persons, the higher locations generally indicate more stable or higher-level persons. The values of the slope can range from negative to positive infinity, where larger absolute values indicate stronger discrimination, and a positive slope suggests that the probability for a response in a given category increases as participant locations increase.

Model Fit Evaluation

Model fit evaluation was established using both item-level and global fit statistics. The i*tem-level fits* were evaluated using the infit and outfit mean square error (MSE) statistics, the signed chi-square test (S-X^2), and the root mean square (RMSEA) item fit statistics. The outfit MSE for an item is the unweighted average of the squared standardized residuals for response to that item.

$$Outfit\ MSE = \frac{\sum_{n=1}^{N} Z_{ni}^2}{N}$$

Where Z^2_{ni} is the squared standardized residual for person *n*'s response to item *i*, and *N* is the number of responses to item *i.* The infit *MSE*, by contrast, is a weighted average of the squared standardized residuals, with weights based on response variance (W_{ni}).

$$Infit\ MSE = \frac{\sum_{n=1}^{N} W_{ni}\, Z_{ni}^2}{\sum_{n=1}^{N} W_{ni}}$$

Values of infit and outfit MSE close to 1.00 indicate acceptable fit, while standardized values near 0 suggest good fit (Linacre, 2002). The signed chi-square test

($S\text{-}X^2$) on the other hand, compares observed and expected responses across subgroups with different total scores. This is interpreted as a chi-square statistic using *p-values*. Non-significant values signify a good fit (Orlando & Thissen, 2003). The root mean square error of approximation (RMSEA) summarizes the residuals across subgroups. It is reported alongside the $S\text{-}X^2$, and smaller values (<.05) indicate a good fit.

The *person fit* was evaluated using the standardized person-fit statistics for polytomous data (Zh), and positive values of Zh indicate an acceptable fit (Drasgow et al., 1985). For the *global model fit* evaluation, the Akaike information criterion (AIC), Bayesian information criterion (BIC), sample-size adjusted BIC (SABIC), and log-likelihood ratio. The AIC evaluates model fit while penalizing model complexity (i.e., the number of parameters. The BIC measures the model fit, but imposes a stronger penalty for model complexity than AIC, and it is especially useful when sample sizes are large because it discourages unnecessary parameters. The SABIC is a modified version of the BIC that adjusts the penalty term based on sample size to avoid over-penalizing when *N* is very large. Lower values of AIC, BIC, and SABIC indicate a better fit and parsimony. On the other hand, the log-likelihood (LL) quantifies how well the model reproduces the observed the higher (less negative) it is, the better the model fit.

Results

Descriptive Statistics and Internal Structure of the Scale

The result in Table 1 shows that the mean responses of students on the 10-item cognitive load scale ranged from 2.82 to 3.03 (*SD* = 0.89 to 1.03). This shows a fairly balanced response distribution, with relatively higher category usage by the students. The marginal and empirical reliability were .89 and .78, respectively, indicating strong measurement precision across the latent trait continuum. The internal consistency for the scale is acceptable, with Cronbach's alpha (α) = 0.79 (95% CI [0.76, 0.83]). The results also show that the items demonstrate modest negative *skew* (-0.32 to -0.53) and relatively moderate *kurtosis* (-1.01 to -0.59), which is an indication that most of the responses are in the upper categories. The corrected item-total correlations range from 0.32 to 0.59, indicating moderate to adequate homogeneity in responses. The response heatmap and category usage plots in Figure 1 confirm this pattern, showing that most of the responses are in the higher category, with relatively sparse use of the lowest category.

Table 1

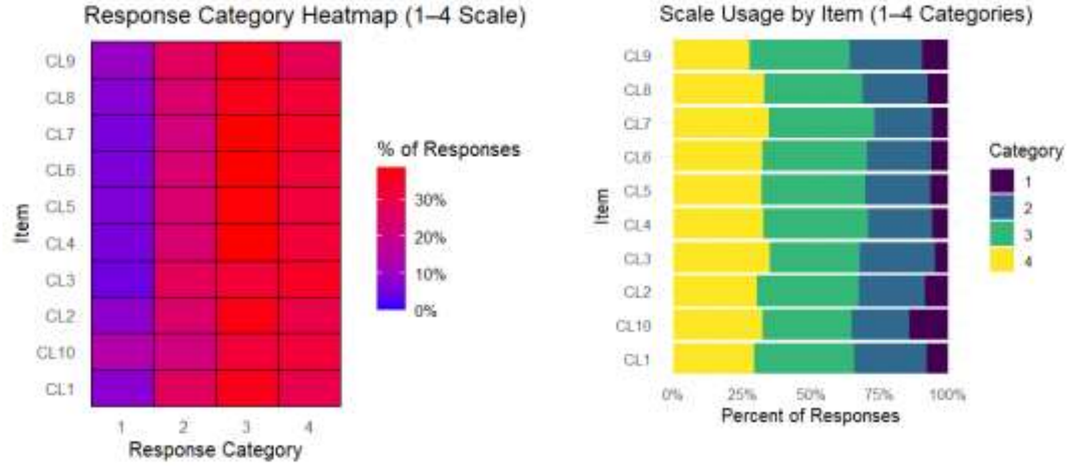
Descriptive statistics for the cognitive load scale.

Items	n	mean	sd	Alpha-if deleted	Corrected item-total	skew	kurtosis	se
1	307	2.88	0.92	0.79	0.32	-0.33	-0.84	0.05
2	307	2.90	0.93	0.79	0.34	-0.40	-0.78	0.05
3	307	2.99	0.90	0.77	0.50	-0.35	-0.96	0.05
4	307	2.98	0.89	0.77	0.51	-0.44	-0.69	0.05
5	307	2.96	0.90	0.77	0.53	-0.43	-0.71	0.05
6	307	2.97	0.89	0.76	0.59	-0.44	-0.69	0.05

7	307	3.03	0.89	0.78	0.41	-0.53	-0.59	0.05
8	307	2.95	0.92	0.77	0.52	-0.44	-0.77	0.05
9	307	2.82	0.94	0.77	0.53	-0.32	-0.86	0.05
10	307	2.83	1.03	0.78	0.41	-0.42	-1.01	0.06

Figure 1

Response heatmap and category usage plot for the cognitive load scale



The results of the PCM in Table 2 show that the thresholds are ordered for all items, with $b_1 < b_2 < b_3$, which supports monotonic category functioning. For instance, Item 1 has b = -1.88, -0.48, 0.50; Item 7 has b = -2.02, -0.84, -0.31. The *Signed Chi-Square* ($S\text{-}X^2$) for item-fit shows that most of the items are fitting well (e.g., item 3, p=.81; item 5, p=.87; item 7, p=.74), while some items have fit indices at the borderline (Item 2 and 10, p = 0.07-0.08), and one clear misfit (Item 8, p < 0.001) with the largest item-level *RMSEA* among the items based on the PCM results. Overall, the PCM captures the category ordering, with good model fit estimates, except for item 8, with misfitting estimates.

Table 2

Model evaluation results using PCM

items	a	b1	b2	b3	$S\text{-}X^2$	$df.S\text{-}X^2$	RMSEA $S\text{-}X^2$	$p.S\text{-}X^2$
1	1	-1.88	-0.48	0.50	44.26	36.00	0.03	0.16
2	1	-1.71	-0.60	0.47	51.36	38.00	0.03	0.07
3	1	-2.45	-0.41	0.16	24.98	32.00	0.00	0.81
4	1	-2.10	-0.71	0.40	37.33	35.00	0.01	0.36
5	1	-1.99	-0.68	0.42	26.83	36.00	0.00	0.87
6	1	-2.04	-0.71	0.41	41.43	35.00	0.02	0.21
7	1	-2.02	-0.84	0.31	30.20	36.00	0.00	0.74
8	1	-1.84	-0.62	0.34	63.15	37.00	0.05	0.00
9	1	-1.61	-0.47	0.57	33.19	38.00	0.00	0.69
10	1	-0.97	-0.58	0.29	51.75	39.00	0.03	0.08

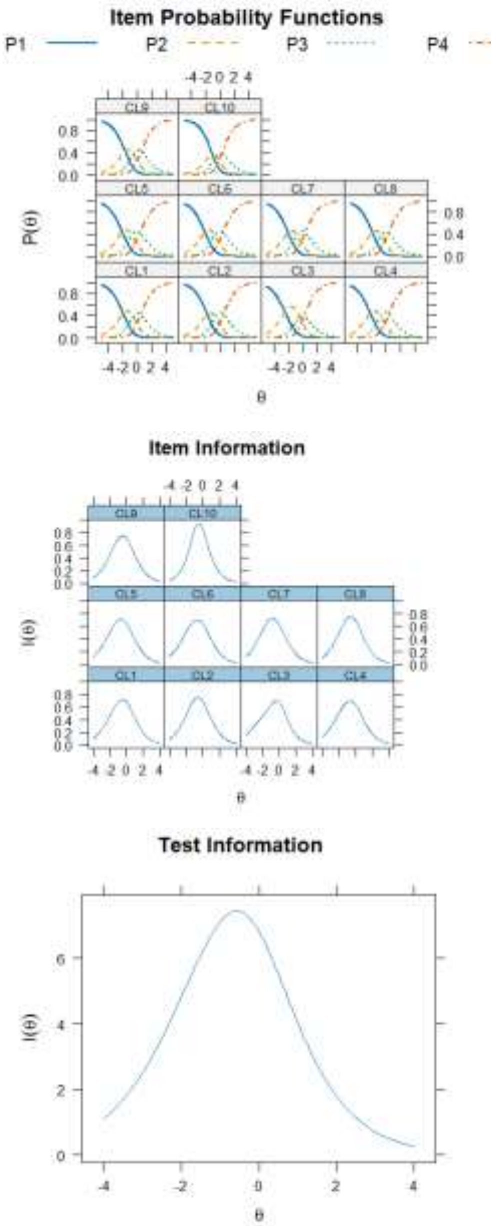
The results of the GPCM in Table 3 show that allowing the discrimination parameter (a) to vary improves data-model fit. The slopes ranged widely from very low (Item 1, a=0.33) to high (Item 6, a=1.59), indicating differential contribution of items to measurement precision. The thresholds remained ordered for all items. The *Signed Chi-Square* ($S\text{-}X^2$) shows reasonable improvement for the majority of the items, relative to the PCM. Based on the GPCM, most of the items fit well, with only item 8 still misfitting ($p = 0.01$) and Item 10 at the borderline ($p = 0.10$). This pattern suggests that varying the discrimination parameter is an important feature in exploring this data, although item 8 remains problematic across both models, suggesting that the item needs further investigation.

Table 3

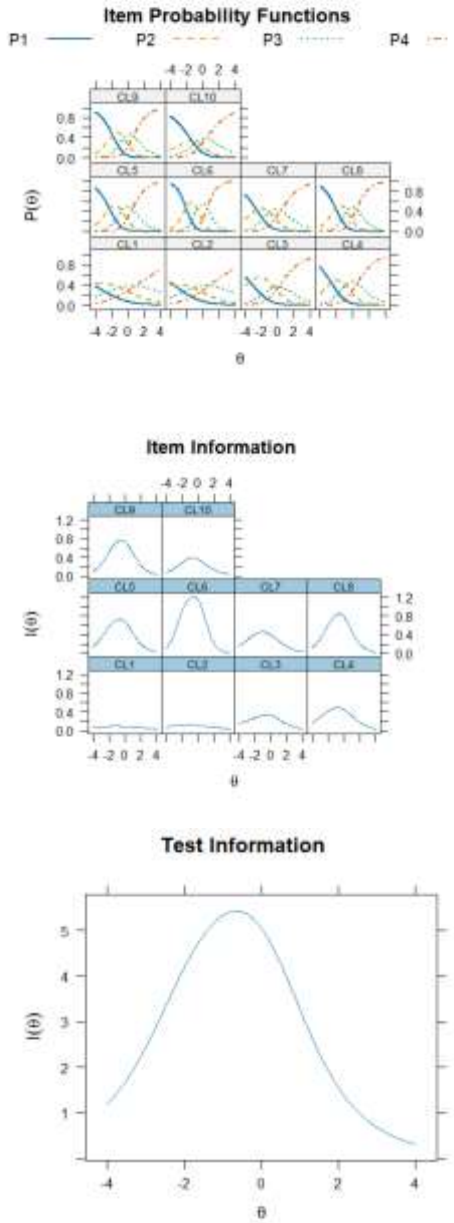
Model evaluation results using GPCM

Items	a	b1	b2	b3	$S\text{-}X^2$	$df.S\text{-}X^2$	RMSEA $S\text{-}X^2$	$p.S\text{-}X^2$
1	0.33	-4.30	-1.06	0.86	44.17	44.00	0.00	0.46
2	0.36	-3.49	-1.28	0.76	52.58	44.00	0.03	0.18
3	0.69	-3.41	-0.53	0.18	27.10	33.00	0.00	0.76
4	0.86	-2.59	-0.87	0.51	34.21	33.00	0.01	0.41
5	1.09	-2.23	-0.75	0.53	22.33	30.00	0.00	0.84
6	1.57	-2.03	-0.69	0.52	32.51	28.00	0.02	0.25
7	0.80	-2.57	-1.05	0.40	27.43	35.00	0.00	0.82
8	1.15	-2.03	-0.67	0.46	53.35	32.00	0.05	0.01
9	1.08	-1.83	-0.51	0.70	34.31	37.00	0.00	0.60
10	0.63	-1.29	-0.84	0.34	52.00	40.00	0.03	0.10

The results in Figure 2 revealed that the item probability curves for both PCM and GPCM show ordered thresholds that align with the observed category use. The item information functions indicate different peaks across items, with the GPCM emphasizing items with varied and higher (e.g., item 6) discriminations compared to the PCM with fixed discrimination estimates. The total test information curves for both PCM and GPCM peak near the trait mean, which also aligns with the descriptive findings and person-ability distribution. Practically, the precision is strongest for respondents around the average cognitive load level, and narrower toward the low and higher ends. These results demonstrate that, on average, most items will provide adequate measures of students' cognitive load. Figure 2 shows the item probability function and test information functions for both PCM and GPCM.



Item and test information plots for PCM



Item and test information plots for GPCM

The results of person fit statistics under the PCM and GPCM were evaluated using the outfit and infit *MSE*, standardized residual (*Zstd*), and standardized person-fit index (*Zh*), presented in Table 4. From the results, the mean *infit* and *outfit* MNSQ are close to 1.00 (PCM: $Infit_{mean} = 0.89$, $Outfit_{mean} = 0.89$; GPCM: $Infit_{mean} = 0.90$, $Outfit_{mean} = 0.89$), with mean standardized residuals (*Zstd*) slightly negative but within the acceptable range of ±2.00 for both models. The average standardized person-fit indices (Z_h) for both models are positive (Zh_{PCM} =0.30, Zh_{GPCM} =0.30), indicating acceptable fit for the models. The wide gap between the minimum and maximum outfits and infits for both PCM ($Outfit_{min}$ = 0.01, $Outfit_{max}$ = 2.81; $Infit_{min.}$ = 0.01, $Outfit_{max}$ = 2.79) and GPCM ($Outfit_{min}$ = 0.01, $Outfit_{max}$ = 3.08; $Infit_{min.}$ = 0.01, $Outfit_{max}$ = 3.21), indicating a few outlying respondents, which may be due to aberrant or idiosyncratic response behavior. The mean central tendency near 0 suggests that the scale is most informative for students with typical or average levels of cognitive load.

Table 4 *Summary of Person fit statistics for both PCM and GPCM*

	F1	SE_F1	Outfit	Zstd	Infit	Zstd	Zh
Min	-1.76	0.34	0.01	-6.40	0.01	-6.37	-5.28
1st Quartile	-0.43	0.34	0.53	-1.22	0.53	-1.25	-0.21
Median	0.04	0.35	0.75	-0.59	0.74	-0.58	0.64
Mean	-0.00	0.36	0.89	-0.43	0.89	-0.42	0.30
3rd Quartile	0.57	0.38	1.17	0.55	1.17	0.54	1.01
Max	1.69	0.50	2.81	3.52	2.79	3.50	2.37
Generalized Partial Credit Model (GPCM)							
	F1	SE_F1	Outfit	Zstd	Infit	Zstd	Zh
Min	-2.74	0.40	0.01	-6.26	0.01	-6.23	-5.46
1st Quartile	-0.70	0.40	0.52	-1.26	0.53	-1.22	-0.24
Median	-0.01	0.42	0.74	-0.59	0.76	-0.51	0.69
Mean	-0.00	0.44	0.89	-0.43	0.90	-0.39	0.30
3rd Quartile	0.69	0.46	1.14	0.49	1.18	0.58	1.04
Max	2.10	0.63	3.08	3.72	3.21	3.86	2.31

The model comparison was evaluated using global model fit statistics such as AIC, BIC, SABIC, and the log-likelihood ratio (Table 5). Based on the results, there is a significant and clear difference in the overall model fit between the PCM and the GPCM. From the results, the GPCM achieved a better fit relative to the PCM, with lower *AIC* (7265.77 vs. 7308.40), lower *SABIC* and *HQ*, and lower *BIC* (7414.84 vs.7423.93), and higher *log-likelihood* (-3592.88 vs. -3623.20). The likelihood-ratio test, $X^2_{(9)} = 60.63$, $p<0.001$, evidently supports having free slopes (discrimination parameters). In all, the model selection criteria consistently endorse the GPCM for the data over the PCM.

Table 5 *Model comparison results between PCM and GPCM*

	AIC	SABIC	HQ	BIC	logLik	X^2	*df*	P
PCM	7308.40	7325.61	7354.60	7423.93	-3623.20			
GPCM	7265.77	7287.98	7325.38	7414.84	-3592.88	60.63	9	0

Discussion and Practical Implications

The findings from this study show acceptable internal consistency and coherent functioning of the items of the cognitive load scale. The item-total correlation suggests that all items contribute positively to the overall functioning of the scale. Although statistics such as skewness and kurtosis, including the category usage and heatmap plots, show that the respondents often endorse higher categories of cognitive load, which should be taken into account when reporting or interpreting cognitive load scores obtained using the scale. The category function showed ordered thresholds across items, supporting the current 4-category structure. There was no evidence of disordered category thresholds; however, the high endorsement of the higher categories suggests an increase in measurement precision for students with moderate to high cognitive load. This finding aligns with the findings of Wind (2023) to the extent that both the PCM and the GPCM showed sensitivity to the cognitive load rating scale; both models provided valuable information about the cognitive load scale.

The item-level analysis shows that Item 8 has a notable misfit in both models and requires further investigation. A close examination of Item 8 (*I find it difficult to manage my assignments and homework at a time*) shows that the item is double-barreled, which could be the reason for its misfit. Therefore, Item 8 would require rewording, replacement or must be reported and interpreted with caution. Items 2 and 10 are borderline under PCM, but acceptable under GPCM, which further supports adopting the GPCM. In terms of the person fit estimates, the average person fit estimates are adequate, though a few respondents showed notable misfit. This may be due to inconsistent responses or data entry anomalies.

The results further show that both models provided the highest item information near the mean trait level, demonstrating that the scale is most suitable for the measurement of moderate to high cognitive load, as most of the items may not precisely measure low cognitive load. On the basis of model choice, the significant improvement in global fit indices and the substantive plausibility of different discriminations show that the GPCM is the best model for the cognitive load scale. It captures the uneven strength of items and reduces misfit without sacrificing ordered thresholds or interpretability. Thus, GPCM estimates should be reported as primary, while PCM can be used to check for robustness. The findings of this study support those of Furr (2017); Dai et al. (2021); Wind (2023), and Harsana et al. (2024), who found that GPCM provided more accurate person ability estimates and fit for the cognitive load rating scale due to the varied item discrimination.

Conclusion and Recommendations

In conclusion, this scale shows sound psychometric properties, with ordered categories, acceptable reliability, and adequate fits for most items. The model comparison strongly supports the GPCM, reflecting meaningful differences in item discrimination that improve item fit and test information. Based on the findings of this study, it was recommended that:

1. Educational researchers and practitioners should rely on the GPCM for assessing the model fit and other psychometric properties of the cognitive load rating scale.
2. The government should sponsor training programs that equip teachers on the use of the GPCM for assessing the psychometric properties of cognitive load rating scales.

References

Achor, S. Z., Zaria, L. I., & Achor, E. E. (2022). Perceived cognitive load and students' performance in social studies. *Journal of Research in Instructional, 2*(2), 129 140. https://doi.org/10.30862/jri.v2i2.74

Akbulut, Y., Dönmez, O., Ceylan, B., & Firat, T. (2025). Effect of pre-training on cognitive load and achievement in a computer-based learning environment. *J. Comput. High Educ., 37*, 1225–1242. https://doi.org/10.1007/s12528-024-09420-6

AL-Nasraween, M. S., AL-Karamneh, M., & Alsoudi, S. (2025). Modeling the Causal Structural Relationship Between Test Wisdom, Cognitive Load, and Academic Achievement Among University Students. *Journal of Educational and Social Research, 15*(1), 78. https://doi.org/10.36941/jesr-2025-0007

Al-Sarry, M. J. S., Majeed, B. H., & Kareem, S. K. (2022). Cognitive Load of University Students and its Relationship to their Academic Achievement. *TEXAS Journal of Philosophy, Culture and History, 3*, 1-13.

Baiduri, Holisin, I., Inganah, S., & Hidayati, W. S. (2024). The impact of cognitive load on learning achievement and semester level in mathematics education students. *Journal of Hunan University (Natural Sciences), 51*(8), 19-35. https://doi.org/10.55463/issn.1674-2974.51.8.2

Bond, T. G., & Fox, C. M. (2015*). Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. *Front. Educ. 6,* 721963. https://doi.org/10.3389/feduc.2021.721963

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford Press.

Dhyaaldian, S. M. A., Kadhim, Q. K., Mutlak, D. A., Neamah, N. R., Kareem, Z. H., Hamad, D. A., Tuama, J. H., & Qasim, M. S. (2022). A comparison of polytomous Rasch models for the analysis of C-tests. *Tabaran Institute of Higher Education, 12*(2), 107-117.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates.

Evans, P., Vansteenkiste, M., Parker, P., Kingsford-Smith, A., & Zhou, S. (2024). Cognitive load theory and its relationships with motivation: a Self-Determination

Theory Perspective. *Educational Psychology Review, 36*(7), 1-25. https://doi.org/10.1007/s10648-023-09841-2

Furr, D. C. (2017). *Partial credit and generalized partial credit models with latent regression.* Stan Case Studies. Retrieved from https://mc-stan.org/learn-stan/case-studies/pcmandgpcm.html

Harsana, F., Retnawati, H., Dewanti, S., Lumenyela, R., Sotlikova, R., Adzima, M., & Septiana, A. (2024). Comparison of item characteristic analysis models of reading literacy test with polytomous Item Response Theory. *REID (Research and Evaluation in Education), 10*(2), 214-226. https://doi.org/10.21831/reid.v10i2.77852

Houichi, A., & Sarnou, D. (2020). Cognitive Load Theory and its Relation to Instructional Design: Perspectives of Some Algerian University Teachers of English. *Arab World English Journal, 11* (4) 110-127. https://dx.doi.org/10.24093/awej/vol11no4.8

Linacre, J. M. (2002). What do infit and outfit, mean square and standardized mean? *Rasch Measurement Trans, 168,* 878.

Linacre, J. M. (2021). *Winsteps Rasch Measurement Software*. Winsteps.com.

Liu, Y., Wang, C., & Paek, I. (2022). A comparison of polytomous item response theory models in measuring cognitive load. *Educational and Psychological Measurement, 82*(1), 45–67. https://doi.org/10.1177/00131644211031234

Lopez-Pina, J., Meseguer-Henarejos, A., Gascon-Canovas, J., Navarro-Villalba, D., Sinclair, V. G., & Wallston, K. A. (2016). Measurement properties of the brief resilient coping scale in patients with systemic lupus erythematosus using Rasch analysis. *Health and Quality of Life Outcomes, 14*:128. https://doi.org/10.1186/s12955-016-0534-3

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Mittal, P., & Banti. (2025). Cognitive load theory: Applications and implications in education. *International Journal of Advanced Scientific Research, 10*(3), 201–205. https://www.allscientificjournal.com/assets/archives/2025/vol10issue3/10098.pdf

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Niyirinda, T., Wind, S., & Ocheni, C. (2026). Evaluating the psychometric properties of the Revised Math Anxiety Rating Scale (MARS) among Ugandan high schools. *International Journal of Assessment Tools in Education*, Advanced Online Publication. https://doi.org/10.21449/ijate.1738209

Ocheni, C. A., & Wind, S. A. (2026). Exploring the psychometric properties of the Testwiseness Rating Scale using a measurement modelling approach.

International Journal of Assessment Tools in Education, 13(2), 626-649. https://doi.org/10.21449/ijate.1735942

Ocheni, C. A., Okeke, A. O., Oyeniran, D. A., Dadzie, J., Ikeh, F. E., Ene, C. U., Agugoesi, O. J., Onyeagba, M. C. N., Nnaji, A. D., Oguguo, B. C. E., Agah, J. J., Ugwuanyi, C. S., Onu, W., Onoja, E. A., & Nwani, S. U. (under review). Multidimensional Rasch model and Rasch tree analysis: A psychometric evaluation of the cognitive load rating scale.

Oguguo, B. C. E., Ejembi, S. O., & Nnaji, A. D. (2025). Appraisal of the interaction effects of two forms of assessment for learning techniques on student's achievement and retention in mathematics. *School Science and Mathematics*,1–18. https://doi.org/10.1111/ssm.18359

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298. https://doi.org/10.1177/0146621603027004004

Ouwehand, K., Lespiau, F., Tricot, A., & Paas, F. (2025). Cognitive load theory: Emerging trends and innovations. *Education Sciences, 15*(4), Article 458. https://www.mdpi.com/2227-7102/15/4/458

Paas, F., & Ayres, P. (2014). Cognitive load theory: A broader view on the role of memory in learning and education. *Educational Psychology Review, 26*(2), 191–195. https://doi.org/10.1007/s10648-014-9263-5

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48.

Society for Education and Training. (2025). The importance of cognitive load theory. https://set.et-foundation.co.uk/resources/the-importance-of-cognitive-load-theory

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development, 68(*1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577.

Thompson, N. (2019). *The Generalized Partial Credit Model (GPCM)*. Assessment Systems. https://assess.com/what-is-the-generalized-partial-credit-model/

Toland, M. D. (2014). Practical guide to conducting confirmatory factor analysis for instrument development. *The Journal of Human Lactation, 30*(3), 335–341. https://doi.org/10.1177/0890334414540430

Torabi, M., Safdari, R., Nedjat, S., Mohammad, K., Goodarzi, M., & Dargahi, H. (2020). Design and Psychometrics of the Assessment Instrument for Innovation Capabilities of Medical Sciences Universities Using the Cube Model Approach. *Iranian journal of public health, 49*(2), 323–331. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC7231711/

Uzun, K. (2021). Cognitive load theory and its educational implications. *6TH International Congress on Life, Social, and Health Sciences in a Changing World, Congress Proceedings Book*, 433-437.

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2023). Analyzing polytomous test data: A comparison between an information-based IRT model and the generalized partial credit model. *Journal of Educational and Behavioral Statistics*.

Wind, S. A. (2023). Detecting rating scale malfunctioning with the partial credit model and generalized partial credit model. *Educational and Psychological Measurement, 83*(5), 953–983. https://doi.org/10.1177/00131644221116292

Yang, J. J., & Yao, G. (2014). *Psychometric Analysis*. In: Michalos, A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_2313

Yılmaz, H. B. (2019). A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test. *International Electronic Journal of Elementary Education, 11*(5), 539-545. https://doi.org/10.26822/iejee.2019553350