Psychology, Evaluation, and Technology in Educational Research
5 (1), 2022, 1-13

Available Online: http://petier.org/index.php/PETIER

Polytomous scoring correction and its effect on the model fit: A case of item response theory analysis utilizing R

Agus Santoso [1, a] *, Timbul Pardede [1, b], Ezi Apino [2, c], Hasan Djidu [2, 3, d], Ibnu Rafi [2, e], Munaya Nikma Rosyada [2, f], Heri Retnawati [2, g], Gulzhaina K. Kassymova [4, 5, h]

[1] Universitas Terbuka. Cabe Raya, Pondok Cabe, Pamulang, Tangerang Selatan 15437, Indonesia
[2] Universitas Negeri Yogyakarta. Jl. Colombo No. 1 Karangmalang, 55281, Yogyakarta, Indonesia
[3] Universitas Sembilanbelas November Kolaka. Jl. Pemuda, Kolaka 93561, Sulawesi Tenggara, Indonesia
[4] Satbayev University. Satpaev St 22, Almaty 050000, Kazakhstan
[5] Abai Kazakh National Pedagogical University. Dostyk Ave 13, Almaty 050010, Kazakhstan
[a] aguss@ecampus.ut.ac.id; [b] timbul@ecampus.ut.ac.id; [c] apinoezi@gmail.com; [d] hasandjidu@gmail.com; [e] ibnurafi789@gmail.com; [f] munayanikma38@gmail.com; [g] heri_retnawati@uny.ac.id; [h] g.kassymova@satbayev.university
* Corresponding Author.

Received: 22 November 2021; Revised: 29 November 2021; Accepted: 3 November 2022

Abstract: In item response theory, the number of response categories used in polytomous scoring has an effect on the fit of the model used. When the initial scoring model yields unsatisfactory estimates, corrections to the initial scoring model need to be made. This exploratory descriptive study used response data from Take Home Exam (THE) participants in the Statistical Methods I course organized by the Open University, Indonesia, in 2022. The stages of data analysis include coding the rater's score; analyzing frequency; analyze the fit of the model based on graded, partial, and generalized partial credit models; analyze the characteristic response function (CRF) curve; scoring correction (rescaling); and re-analyze the fit of the model. The fit of the model is based on the chi-square test and the root mean square error of approximation (RMSEA). All model fit analyzes were performed by using R. The results revealed that scoring corrections had an effect on model fit and that the partial credit model (PCM) produced the best item parameter estimates. All results and their implications for practice and future research are discussed.

Keywords: Model Fit; GRM; PCM; GPCM; R Programming; Polytomous Scoring Approach

How to Cite: Santoso, A., Pardede, T., Apino, E., Djidu, H., Rafi, I., Rosyada, M. N., Retnawati, H., & Kassymova, G. K. (2022). Polytomous scoring correction and its effect on the model fit: A case of item response theory analysis utilizing R. *Psychology, Evaluation, and Technology in Educational Research*, *5*(1), 1-13. https://doi.org/10.33292/petier.v5i1.148

INTRODUCTION

Assessment is used to determine learning outcomes at a certain level of education (Rafi et al., 2023; Retnawati, Kartowagiran, et al., 2017; Reynolds, 2010), including at the higher education level. The assessment uses measurement tools, one of which is a test (Brookhart & Nitko, 2019; Retnawati, Hadi, et al., 2017). One form of test that is frequently used is a constructed-response test (Brookhart & Nitko, 2019; Retnawati, Hadi, et al., 2017). Constructed-response tests are used to ensure that the tests used measure what they are supposed to measure (Retnawati, Hadi, et al., 2017), guarantee the honesty of the test takers, and reduce the potential for cooperation between test takers in working on the test. Based on the advantages of the

10.33292/petier.v5i1.148

constructed-response test, the Indonesia Open University (Universitas Terbuka) has developed a test system called the Take Home Exam (THE) which uses constructed-response items used to measure student learning outcomes.

THE at the Open University is an examination that is held at the end of the semester. THE is administered online, then students are given the opportunity to work on the test 12 hours after the test is administered or available in the system. After that, each subject lecturer would assess student work by assigning a score on each item that students work on. The score of each item is then entered into the system. The final student score is obtained by adding up the scores for each item. Estimating student competence or ability in this way still uses the classical test theory approach.

The classical test theory approach has several limitations (see Hambleton & Jones, 2005; Retnawati, 2016; Zanon et al., 2016). In the classical test theory approach, the participant's score obtained from a test is limited to that test, so that the test results are not possible to be generalized beyond the test (Hambleton & Jones, 2005; Retnawati, 2016; Zanon et al., 2016). Thus, the scores obtained by test takers are highly dependent on the choice of test, not on their abilities. As an illustration, when participants choose a test with an easy level of difficulty, the score will be high, but conversely if the test chosen has a high level of difficulty, the score will be low. Accordingly, another approach is needed to overcome the limitations of the classical test theory approach.

There is another approach that can be used to estimate student abilities known as the modern approach. This modern approach is often referred to item response theory (IRT). For constructed-response (essay) items, the scoring used is polytomous. There are several mathematical models for polytomous scoring to choose from: graded response model (GRM), partial credit model (PCM), and generalized partial credit model (GPCM) (Auné et al., 2020; Hambleton et al., 1991; Retnawati, 2014; Yılmaz, 2019). The analysis using these three models has not been applied to the scoring and estimation of students' abilities in THE developed by the Open University.

In the GRM, the response of a participant to the item j is categorized into $m + 1$ sequential categories, where $k = 0, 1, 2, ..., m$ and m is the number of steps in completing the item j correctly (Retnawati, 2014; Samejima, 1970). In this model, the index of difficulty in each step is also ordered. The relationship between item parameters and participant abilities in the GRM for homogeneous cases (a_j or item discrimination is the same for each step) by Samejima (1970) is expressed in Formula 1 and 2.

$$P_{jk}(\theta) = P^{*}_{jk}(\theta) - P^{*}_{jk+1}(\theta) \tag{1}$$

$$P_{jk}(\theta) = \frac{\exp\left[a_j\left(\theta - b_{jk}\right)\right]}{1 + \exp\left[a_j\left(\theta - b_{jk}\right)\right]} \tag{2}$$

where $P^{*}_{j0}(\theta) = 1$ and $P^{*}_{jm+1}(\theta) = 0$; θ is the level of individual's ability; $P_{jk}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of category k on the item j; $P^{*}_{jk}(\theta)$ is the probability of an individual with the level of ability θ to obtain a score of k or more on the item j; a_j is item discrimination index of the item j; and b_{jk} is difficulty index of category k for the item j.

The PCM is an extension of the Rasch model on dichotomous data (Masters, 1988; Retnawati, 2014). PCM assumes that each item has the same discriminating power. PCM is similar to GRM for items scored in tiered categories, but the index of difficulty in each step does not have to

be ordered (Retnawati, 2014). In this case, a step can have a higher level of difficulty than the next step. The PCM model by Masters (1988) is Formulated 3:

$$P_{jk}(\theta)=\frac{\exp\left[\sum_{v=0}^{k}\left(\theta-b_{jv}\right)\right]}{\sum_{h=0}^{m}\exp\left[\sum_{v=0}^{k}\left(\theta-b_{jv}\right)\right]};k=1,2,3,\ldots,m \quad (3)$$

where θ represents the level of ability of a test taker; $P_{jk}(\theta)$ is the probability of a test taker whose level of ability is θ to obtain a score of category k on the item j; and b_{jv} represents difficulty index of category v on the item j.

GPCM is an advance development of polytomous scoring (Muraki, 1992; Retnawati, 2014). Muraki (1992) suggested that GPCM is a general form of PCM which is known as the item category response function. The GPCM is mathematically Formulated 4 (Muraki, 1992).

$$P_{jk}(\theta)=\frac{\exp\left[\sum_{v=0}^{k}a_{j}\left(\theta-b_{jv}\right)\right]}{\sum_{h=0}^{m}\exp\left[\sum_{v=0}^{k}a_{j}\left(\theta-b_{jv}\right)\right]};k=1,2,3,\ldots,m \quad (4)$$

where θ denotes the level of ability of a test taker; $P_{jk}(\theta)$ is the probability of a test taker whose level of ability is θ to obtain a score of category k on the item j; a_j denotes item discriminating power of the item j; and b_{jv} represents difficulty index of category v on the item j.

The GRM, PCM, and GPCM have been widely used in the previous studies to analyze the quality of constructed-response tests and to estimate the ability of test takers (e.g., Abdelhamid et al., 2021; Safitri & Retnawati, 2020; Sun et al., 2021). In the present study, we strives to apply the same to THE developed by the Open University. The contribution of this study is expected to provide constructive input for institutions in improving the quality of tests and their scoring. On a broader scale, this study is expected to provide guidance for researchers and test developers in designing scoring rubrics and choosing the right polytomous scoring model. Thus, this study aims to analyze the effect of the polytomous scoring model and its scoring corrections on the model fit.

METHODS

Study Design

This descriptive study focuses on exploring the effect of the polytomous scoring model on the test instrument used in the Take Home Exam (THE) at the Open University and its scoring corrections on the fit of the polytomous IRT models. Three polytomous scoring models were tested in this study: GRM, PCM and GPCM. In order to perform the polytomous IRT models fit test, first, we rescaled each test taker's raw score to six response categories (i.e., 0, 1, 2, 3, 4, 5). The new scores were then estimated using item response theory using the GRM, PCM, and GPCM . Furthermore, the scores with the six response categories were corrected by considering the results of item parameter estimation and the categorical response function (CRF) curve. The scoring correction produces a new score with three response categories (0, 1, and 2). Then this score is estimated again using the GRM, PCM, and GPCM. Finally, this study will test which model produces the best model fit parameter estimates after scoring corrections on the instrument used in THE.

Participants

This study involved Open University students who took the Take Home Exam (THE) for the Statistical Methods I course which was held in 2022. A total of 1173 students participated in this study. Participants took THE online and they were given a maximum of 12 hours to complete the test. Participants could access test items and upload their answers to the system provided by the Open University. They cannot upload their answers if they take more than 12 hours to complete the test.

Data Collection

This study used student response data in answering the THE. The instruments used in the THE were used by the Open University to determine learning outcomes in courses offered in the certain academic year. In this study, we only focused on the of student response on Statistical Methods I test held in 2022. The test consists of four constructed-response (essay) items. Student's response or work on the test then was assessed by each lecturer in charge of the course online. Each item was assigned a maximum score of 25. The raw score for each item was input in real-time by the lecturer on the system.

The data analyzed in this study was the raw score data per item for each student. We applied documentation techniques to obtain the student response data. In this case, the data was retrieved directly from the examination system developed by the Open University. The data obtained consisted of student ID, raw score per item, and total score. In this study, however, we only focused on raw scores per item, while student ID and total scores were not analyzed. Thus, the final data analyzed was the raw score per item for 1173 students taking THE.

Data Analysis

Data analysis begins with coding the rater's scoring of the constructed-response items. This coding produces a polytomous score with six response categories (i.e., score of 0, 1, 2, 3, 4, and 5). Afterwards, we conducted a frequency analysis to describe the distribution of each response category. The next step was to analyze the fit of the model with GRM, PCM, and GPCM. At this stage, we focused on investigating the model fit parameters, namely the chi-square test and the Root Mean Square Error of Approximation (RMSEA). In addition, we also analyzed the characteristic response function (CRF) curve for each test item in the GRM, PCM, and GPCM. We used the results of this analysis to make corrections to previous scoring. Based on the information on the CRF curve, we corrected the score from six response categories to three response categories (i.e., 0, 1 and 2). After that, we re-analyzed the data using three response categories to obtain model fit information. We also investigated changes in the value of the model fit parameters before and after scoring corrections were made. The best model was obtained by considering the *p*-value and RMSEA. All analyzes were performed with RStudio using the 'mirt' package (Chalmers, 2012; The R Development Core Team, 2013).

RESULTS AND DISCUSSION

Distribution of Polytomous Scoring Results

The polytomous scoring on THE consists of six score categories, namely 0, 1, 2, 3, 4, and 5. A score of 0 represents the test taker's performance in answering THE is low and a score of 5 represents that the test taker's performance is high. The distribution of the test takers' performance in answering the four THE items for each score category is presented in Table 1. In the item 1, the students' performance was relatively even for each category. In addition, in the item 1, the most test takers obtained a score of 1 (21.5%) and a score of 2 (21.3%) and the

least test takers obtained a score in category 0 (12.4%). In the item 2, most test takers obtained a score of 5 (36.7%) and the least participants obtained a score of 0 (7.3%). In the item 3, the most participants received a score of 3 (30%) and the fewest participants received a score of 5 (4.5%). In the item 4, the distribution of student performance was relatively even, where the most test takers received a score of 4 (21.3%) and the lowest test taker received a score of 0 (12.2%).

Table 1. Distribution of Polytomous Scoring Results on THE (N = 1173)

Score	Item 1		Item 2		Item 3		Item 4	
	n	%	*n*	%	*n*	%	*n*	%
0	145	12.4	86	7.3	154	13.1	143	12.2
1	252	21.5	99	8.4	160	13.6	197	16.8
2	250	21.3	123	10.5	262	22.3	185	15.8
3	138	11.8	188	16.0	352	30.0	229	19.5
4	157	13.4	247	21.1	192	16.4	250	21.3
5	231	19.7	430	36.7	53	4.5	169	14.4

Comparison of the Fit of Polytomous IRT Models

Three polytomous scoring models, i.e., GRM, PCM, and GPCM, were used to estimate the item and test taker's ability parameters. The value of RMSEA was used to identify the fit of the items with the scoring model used. If $p < 0.05$, then an item does not fit the scoring model. The RMSEA and p-values of the four items for each polytomous scoring model are presented in Table 2. Based on Table 2, for GRM, there is only one item, namely item 1, which fits the model. Meanwhile, for scoring with PCM and GPCM, none of the items fit the model. These results indicate that the scoring using six categories (i.e., score of 0 to 5) does not fit the polytomous scoring models used: GRM, PCM, and GPCM. Categorical response function (CRF) curves for each item for the GRM, PCM, and GPCM respectively are presented in Figure 1. The response for each item in the GRM, PCM, and GPCM does not form a specific pattern, which makes it difficult to interpret. Furthermore, the CRF curve for each scoring category for each item does not work properly. This further confirms that polytomous scoring with six categories does not fit the GRM, PCM, and GPCM. Thus, the polytomous scoring category used in the THE for the Statistical Methods I test needed to be corrected.

Table 2. Comparison of Fit of Polytomous Scoring Models in the THE

Item	GRM		PCM		GPCM	
	RMSEA	*p*	RMSEA	*p*	RMSEA	*p*
1	0.015	0.147	0.025	0.002*	0.019	0.040*
2	0.042	0.000*	0.036	0.000*	0.036	0.000*
3	0.046	0.000*	0.030	0.000*	0.029	0.000*
4	0.024	0.011*	0.024	0.004*	0.020	0.023*

Note. * $p < 0.05$, item does not fit the model

Correction on THE Scoring Category

The polytomous scoring of the instrument used in the Statistical Methods I test which originally consisted of six categories was revised into three categories (i.e., score 0, score 1, and score 2). This correction was performed to obtain better estimation of item and ability parameters. The results of the correction in the THE polytomous scoring category are presented in Table 3. The first category on the old scoring (i.e., a score of 0) for each item was not corrected. In item 1, the second category (i.e., score 1) and third category (i.e., score 2) on the

old scoring are combined into a new score of 1. This is based on the CRF curve (see Figure 2) which shows that the curves of participants' responses for score 1 and score 2 tend to coincide. Furthermore, in the item 1, the scores 3, 4, and 5 on the old version of scoring were combined into a new score, namely score 2. This combination was carried out after observing the curves of participants' responses for the scores of 3, 4, and 5 which also tended to coincide (see Figure2). For the item 2, item 3, and item 4, scores 1, 2, 3, and 4, based on the CRF curves which tended to coincide, were combined into a new score of 2, while a score of 5 on the old version of scoring was changed to a score of 2. Therefore, in the new version of polytomous scoring, the lowest score is 0 and the highest is 2. A score of 0 represents a test taker with low performance, while a score of 2 represents a test taker with high performance.

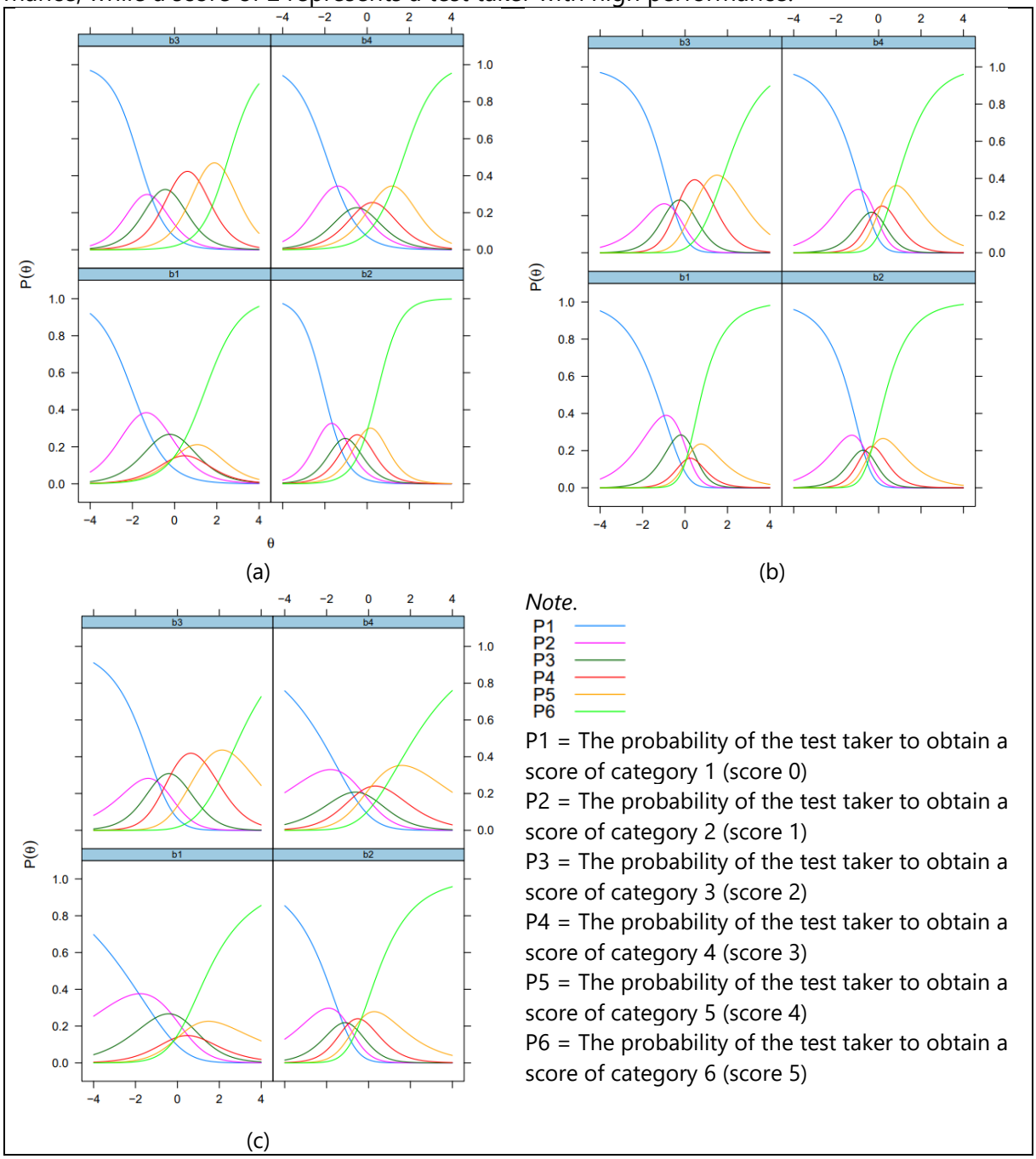

Figure 1. CRF Curve of Four Constructed-Response Items Used in the Statistical Methods I Test Before Scoring Corrections Were Made: (a) GRM; (b) PCM; and (c) GPCM

Table 3. The Results of Correction on THE Scoring Category

The old scoring (6 categories)	The new scoring			
	Item 1 (3 categories)	Item 2 (3 categories)	Item 3 (3 categories)	Item 4 (3 categories)
0	0	0	0	0
1	1			
2		1	1	1
3				
4	2			
5		2	2	2

Distribution of New Version Polytomous Scoring Results

The new version of polytomous scoring, which is an improvement on the old version of polytomous scoring in terms of the number of score categories, consists of three score categories, namely score 0, score 1, and score 2. Score 0 represents the test taker's performance in answering THE is low and score 2 represents that the test taker's performance is high. The distribution of the test takers' performance in answering the four constructed-response items based on the new scoring category is presented in Table 4. In the item 1, most students scored in category 2 (44.9%) and the least scored in category 0 (12.4%)). In the item 2, most test takers obtained a score of 1 (56%) and the least test takers obtained a score in category 0 (7.3%). In the item 3, most test takers received a score of category 1 (82.3%) and the least participants received a score of category 2 (4.5%). In the item 4, most participants obtained a score of category 2 (73.4%) and the least number of participants scored in category 0 (12.2%). In general, with the new polytomous scoring, most of the test takers' performance in the THE was in the moderate category.

Table 4. Distribution of New Version Polytomous Scoring (N = 1173)

New score	Item 1		Item 2		Item 3		Item 4	
	n	%	*n*	%	*n*	%	*n*	%
0	145	12.4	86	7.3	154	13.1	143	12.2
1	502	42.8	657	56.0	966	82.3	861	73.4
2	526	44.9	430	36.7	53	4.5	169	14.4

Comparison of the Fit of Polytomous IRT Models After Scoring Correction

In order to determine the best polytomous scoring model or polytomous IRT model, the results of scoring correction were estimated using the GRM, PCM, and GPCM to obtain the RMSEA value and the fit of each item with the model. The RMSEA and p-values of the four constructed-response items used in the THE which were obtained after scoring correction are presented in Table 5. When the scoring correction was made so that there are three score categories, the estimation results with the GRM show that there are three items that fit the model and one item (i.e., item 2) did not fit the model (RMSEA = 0.089; p = 0.017). The same thing can be found in the estimation results with the GPCM, where three items fit the model, while one item (i.e., item 2) does not fit the model (RMSEA = 0.078, p = 0.039). However, when the results of scoring correction were estimated with the PCM, all items fit the model (RMSEA ranges from 0 to 0.067). Thus, the THE instrument with the new version of polytomous scoring that consists of three score categories is best estimated using the PCM.

Table 5. THE New Polytomous Scoring Model's Fit Comparison

Item	GRM		PCM		GPCM	
	RMSEA	*p*	RMSEA	*p*	RMSEA	*p*
1	0.000	0.463	0.049	0.187	0.007	0.408
2	0.089	0.017*	0.067	0.074	0.078	0.039*
3	0.006	0.698	0.000	0.624	0.000	0.645
4	0.062	0.096	0.065	0.081	0.060	0.108

Note. * $p < 0.05$, item does not fit the model

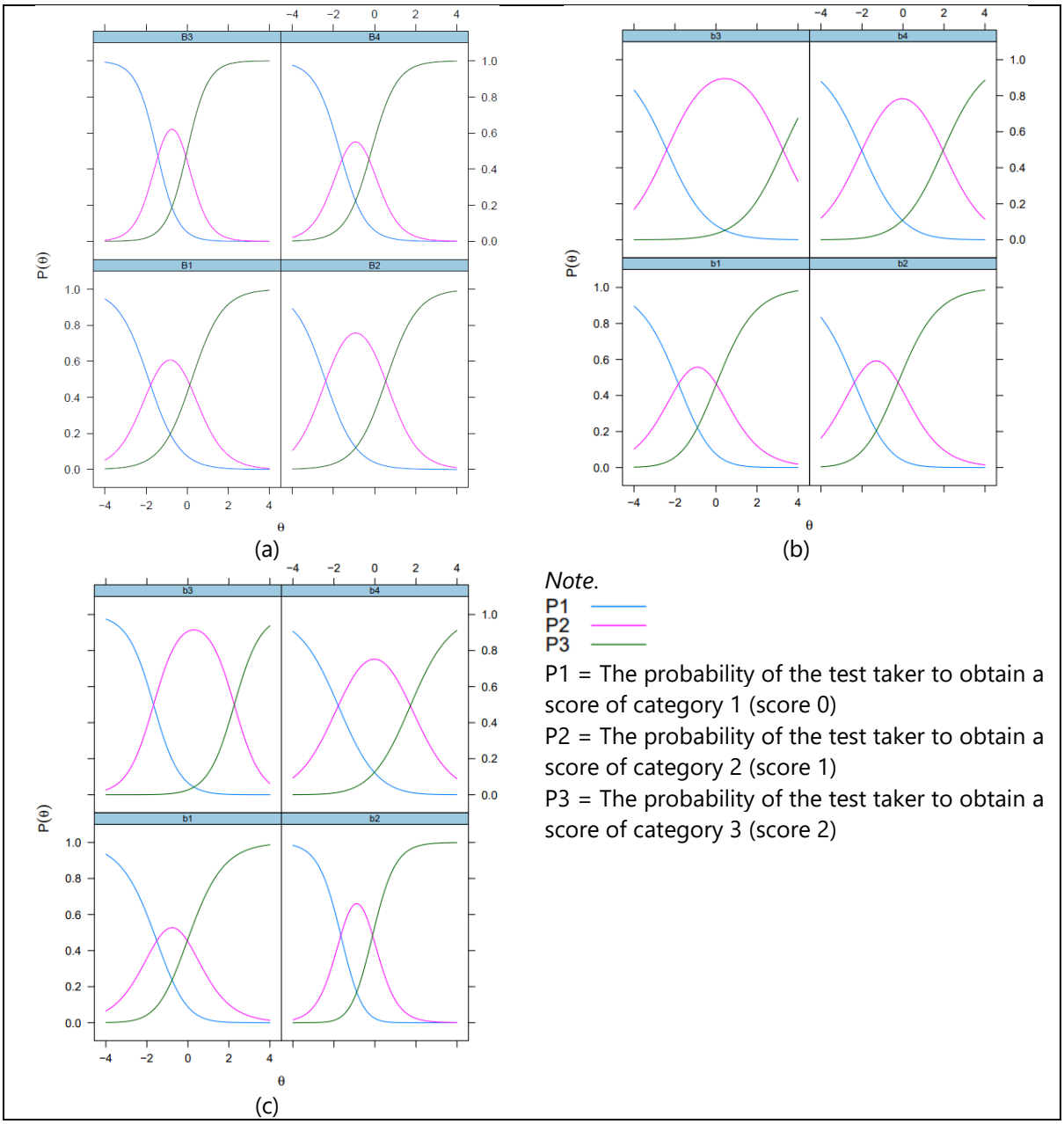

Figure 2. CRF Curve of Four Constructed-Response Items Used in the Statistical Methods I Test After Scoring Corrections Were Made: (a) GRM; (b) PCM; and (c) GPCM

If we look at the CRF curves of each item (see Figure 2), the curves for each category work well so that the information provided by the CRF curves is meaningful. An example of interpreting the CRF curve for item 1 using the estimated PCM model (see Figure 2b): test takers with an ability of less than –1.9 have a higher probability of obtaining a 0 score with a

probability up to 1; test takers with an ability of –1.9 to 0.1 have a higher probability of obtaining a 1 score with a maximum probability of 0.58; whereas participants with an ability of more than 0.1 have a higher probability of obtaining a 2 score with a probability up to 1. Thus, in item 1 the test taker requires a minimum ability of –1.9 to obtain a score of 1, while in order to obtain a score of 2, a test participant is required with a minimum ability of 0.1. Likewise with item 2 (see Figure 2b), to obtain a score of 1 the test taker requires a minimum ability of –2.2 and a score of 2 requires a minimum ability of –0.2. The same analogy used for interpreting the CRF curves item 3 and item 4.

Reliability and IRT Assumptions

In order to reveal the reliability of the THE instrument, reliability estimation was conducted using Cronbach's α formula. The estimation was performed using the scoring correction results (three response categories: 0, 1, and 2). The estimation results with Cronbach's α formula yield a reliability coefficient of 0.634. Although not excessively high, the reliability coefficient adequately describes the quality of THE instrument. Thus, the THE instrument has relatively good consistency in estimating test takers' ability.

One of the crucial assumptions in item response theory is that the instrument must only measure a single dimension (unidimensional assumption). To put these assumptions to the test, we conducted factor analysis to obtain the eigenvalues for each factor formed using the help of the jamovi program. This factor analysis is preceded by verifying the adequacy of the sample and the correlation matrix is not the identity matrix. It was found that KMO measure of sampling adequacy (MSA) is 0.68 overall, suggesting the sample was adequate for factor analysis. Bartlett's test of sphericity demonstrates that the correlation matrix is not an identity matrix (χ^2 (6) = 686.33, p < 0.01). After that, factor analysis was performed to obtain the eigenvalues of the four components or factors based on the number of items in the THE instrument. The eigenvalues for the four factors that comprise THE instrument is presented in Table 6. Table 6 shows that there is only one dominant factor (eigenvalue > 1), which is factor 1. This result indicates that the THE instrument only measures one dominant dimension, thus fulfilling the unidimensional assumption. This is confirmed by the scree plot in Figure 3 which shows only one steep slope is formed, indicating that the instrument meets the unidimensional assumption.

Table 6. Eigenvalues from Four Factors

Factor	Eigenvalues
1	1.3667
2	0.0901
3	0.0524
4	–0.1434

The next assumption in IRT that must be satisfied is local independence. Local independence indicates a condition in which the test taker's response to a test item does not affect their response to other test items (Retnawati, 2014). This condition could happen because the abilities or factors that can affect the performance of test takers in a test are constant. Retnawati (2014) suggests that support for fulfilling the assumption of local independence is obtained when it can be shown that there is the same in magnitude between the probability that test takers with a certain ability level can answer all the test items correctly and the product of the probabilities of test takers with that particular level of ability can answer each test item correctly. In this study, the assumption of local independence is said to be satisfied by investigating that in terms of content, the test items used in THE are independent of one another. In other words,

the test takers' responses to each test item used in THE do not affect their responses to other test items.

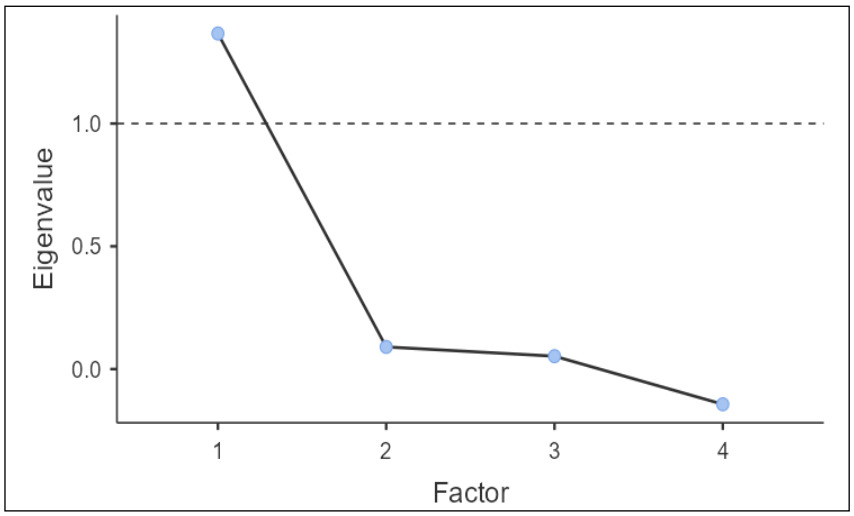

Figure 3. Scree Plot of the Eigenvalues of the Factors Formed

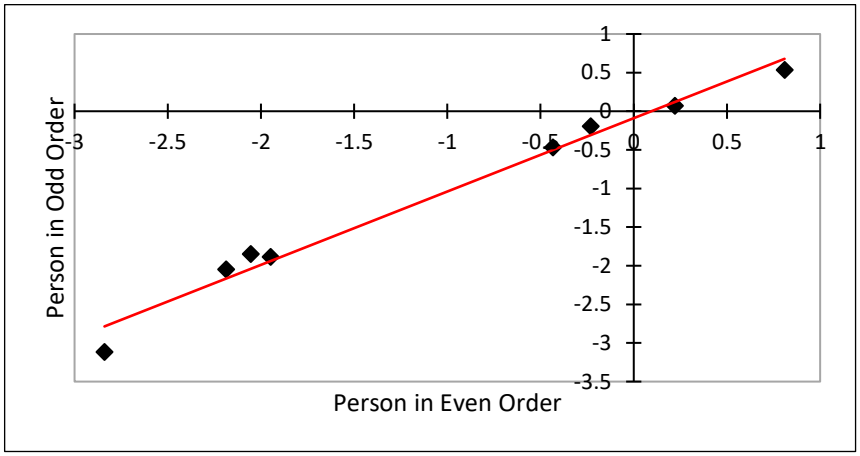

Figure 4. Plot of THE Item Parameter Invariance (Steps 1 and 2)

Another assumption that must be satisfied in IRT is parameter invariance. In this study, we only focused on examining the invariance of the item parameters because ability parameter invariance is impossible to be assessed because the THE instrument consists of only four items. In order to test the invariance of the item parameters, we first split the data into two parts based on the order of test takers. The first half was the data of test takers who are in odd order, while the second half was the data of test takers who are in even order. We then estimated the item parameters for every split using the PCM. The PCM was chosen because it produces the best model fit compared to the GRM and GPCM. The results of item parameter estimation (i.e., the value of step 1 and step 2) of each item in the first and second halves were correlated. The correlation coefficient obtained is 0.991, indicating that the item parameter invariance assumption is satisfied. The plot of the step 1 and step 2 parameters for each item is presented in Figure 4. Because the distribution of the dots in Figure 4 tends to be close to a straight line, there is no issue that needs to be addressed regarding THE item parameter invariance.

Discussion

Polytomous scoring using six response categories (scores 0 to 5) for the four THE instrument items reveals that none of the items fit into the PCM and GPCM. When estimated using the

GRM, however, only one item (i.e., item 1) fits the model. This indicates that polytomous scoring using six response categories produces unsatisfactory estimates for all estimation models. Several studies suggest that four response categories produce a good model fit (Abal et al., 2017; Lozano et al., 2008). Based on that, it makes sense that using six response categories results in an unsatisfactory model fit. As a result, scoring improvements or corrections are required to improve the model's fit.

The estimation results using six response categories were unsatisfactory, so the polytomous score of THE instrument had to be rescaled (rescaling needs to be made). This refers to the CRF curves for all estimation models, demonstrating that some response categories do not work well. Several categories of responses that did not work were combined with other response categories to provide better information for estimating the minimum ability of students to score in these response categories. Finally, the initial scoring using six response categories was rescaled to three response categories (i.e., scores 0 to 2). Scoring with the three response categories yields better estimates for all polytomous IRT models, and each response category on the CRF curve functions correctly (see Figure 2).

Estimation results using three response categories for THE instrument show that all items fit the PCM model, whereas estimation using the GRM and GPCM only yields three out of four items that fit the model. Thus, the estimation of polytomous scoring of THE instrument using the PCM model is far more satisfying than the GRM and GPCM. This finding differs from the findings of Yılmaz (2019), who found that GRM produces smaller RMSEA values than GPCM. However, Yılmaz's study (2019) did not directly compare GRM and GPCM. The findings of this study are also different from those of Suciati et al. (2022), who reported that the PCM did not meet the model fit for the constructed-response test instrument they developed. The results of other studies are also inconsistent with ours, such as the study of Auné et al. (2020), who reported that the GRM produced the best model fit parameters than the GRM, GPCM, and PCM. As a result, there has been no consistency in research results to determine the best polytomous scoring model up to this point. However, our study strengthens the future discourse so that other researchers can continue investigating this issue.

The results of this study contribute to improving the quality of assessment using the polytomous scale instrument. The results are significant because they can be used as a consideration for test developers to choose the appropriate response category when creating a polytomous scale instrument scoring rubric, such as a constructed-response test. The selection of the correct number of response categories and impacting model fit also guarantees that the test meets the principle of fairness for participants. In other words, choosing the correct number of response categories can reduce bias regarding the actual test taker's ability. No less essential, these results also guide researchers, psychometricians, and even test developers to choose the estimation model that best fits the characteristics of the data to be analyzed.

Study Limitations and Implications

Although this study successfully provided empirical evidence regarding scoring correction procedures and their impact on model fit, it still has several limitations. *First*, because the data used is limited to a single case, researchers cannot guarantee the consistency of the research findings. An important question that must be addressed next is whether the same procedures and estimates will yield the same results when applied to other data sets. Thus, future research should broaden this study by involving more data sets with the same characteristics to strengthen or correct the findings from this study. *Second*, there were only four items used in this study. The intriguing question is whether the same procedures and estimates applied to larger instruments will yield the same results. This question is a guideline and an opportunity for

future research, and it is expected to complement the findings of this study. *Third*, this study only looks at the effect of scoring corrections on the fit of the estimation model and not on the ability parameter estimates. As a result, it is critical to conduct research on this topic in the future.

CONCLUSION

This study reveals that the number of response categories on polytomous scoring impacts model fit. Scoring with fewer response categories results in a better model fit. Scoring with more response categories will likely cause some response categories not to work. This has an impact on the difficulty of obtaining optimal model fit. The study also revealed that the PCM produced the best fit compared to the GRM and GPCM. This indicates that the test used (i.e., THE developed by the Open University) is best analyzed using the PCM. The PCM analysis would produce the most accurate estimation of item and person parameters. This accuracy is important to ensure the quality of the test and fulfill the principle of fairness in estimating participants' abilities. Finally, we recommend that test developers be careful in creating essay-scoring rubrics. In the rubric developed, the test developer should pay attention to the functioning of the given score category. There must be a clear definition of when participants score 1, 2, and so on. In addition, the model fit parameters must be considered when choosing an analysis model on polytomous scoring.

REFERENCES

Abal, F. J. P., Auné, S. E., Lozzia, G. S., & Attorresi, H. F. (2017). Funcionamiento de la categoría central en ítems de confianza para la matemática. *Revista Evaluar*, *17*(2). https://doi.org/10.35670/1667-4545.v17.n2.18717

Abdelhamid, G. S. M., Bassiouni, M. G. A., & Gómez-Benito, J. (2021). Assessing cognitive abilities using the WAIS-IV: An item response theory approach. *International Journal of Environmental Research and Public Health*, *18*(13), 6835. https://doi.org/10.3390/ijerph18136835

Auné, S. E., Abal, F. J. P., & Attorresi, H. F. (2020). Análisis psicométrico mediante la Teoría de la Respuesta al Ítem: modelización paso a paso de una Escala de Soledad. *Ciencias Psicológicas*, *14*(1). https://doi.org/10.22235/cp.v14i1.2179

Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). Pearson.

Chalmers, R. P. (2012). mirt : A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Hambleton, R. K., & Jones, R. W. (2005). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

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

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education*, *1*(4), 279–297. https://doi.org/10.1207/s15324818ame0104_2

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

Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I., & Rosyada, M. N. (2023). What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination. *Pedagogical Research*, *8*(1), em0145. https://doi.org/10.29333/pr/12657

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Parama Publishing.

Retnawati, H., Hadi, S., Nugraha, A. C., Ramadhan, M. T., Apino, E., Djidu, H., Wulandari, N. F., & Sulistyaningsih, E. (2017). *Menyusun laporan hasil asesmen pendidikan di sekolah: referensi untuk pendidik, mahasiswa, & praktisi pendidikan*. UNY Press.

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction*, *10*(103), 257–276. https://doi.org/10.12973/iji.2017.10317a

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment*, *22*(1), 1–4. https://doi.org/10.1037/a0018811

Safitri, A., & Retnawati, H. (2020). The estimation of mathematics literacy ability of junior high school students with partial credit model (pcm) scoring on quantity. *Journal of Physics: Conference Series*, *1581*, 012030. https://doi.org/10.1088/1742-6596/1581/1/012030

Samejima, F. (1970). Erratum Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *35*(1), 139–139. https://doi.org/10.1007/BF02290599

Suciati, Munadi, S., & Sugiman. (2022). Estimation of test item parameters with polytomous item response using Partial Credit Model (PCM). *Proceedings of the 2nd International Conference on Innovation in Education and Pedagogy (ICIEP 2020)*. https://doi.org/10.2991/assehr.k.211219.042

Sun, X., Zhong, F., Xin, T., & Kang, C. (2021). Item response theory analysis of general self-efficacy scale for senior elementary school students in China. *Current Psychology*, *40*(2), 601–610. https://doi.org/10.1007/s12144-018-9982-8

The R Development Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.yumpu.com/en/document/view/6853895/r-a-language-and-environment-for-statistical-computing

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

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x