Manuscript

by Agus Santoso

REiD (Research and Evaluation in Education), ...(…), 20.., …-…
Available online at: http://journal.uny.ac.id/index.php/reid

The effect of scoring correction and model fit on the estimation of ability parameter and person fit on polytomous item response theory

Anonymous

ARTICLE INFO

Article History
Submitted:
--
Revised:
--
Accepted:
--

Keywords
ability estimate; model fit; person fit; polytomous IRT; scoring correction

ABSTRACT

Scoring quality has been recognized as one of the important aspects that should be of concern to both test developers and users. This study aimed to investigate the effect of scoring correction and model fit on the estimation of ability parameter and person fit in the polytomous item response theory. The results of 165 students in the Statistics course (SATS4410) test at one of the universities in Indonesia were used to answer the problems in this study. The polytomous data obtained from scoring the test results were analyzed using the Item Response Theory (IRT) approach with the Partial Credit Model (PCM), Graded Response Model (GRM), and Generalized Partial Credit Model (GPCM). The effect of scoring correction and model fit on the estimation of ability and person fit was tested using multivariate analysis. Among the three models used, GRM showed the best fit based on *p*-value and RSMEA. The results of the analysis also showed that there was no significant effect of scoring correction and model fit on the estimation of the test taker's ability and person fit. From the results of this study, we recommend the importance of evaluating the levels or categories used in scoring on student work on a test.

How to cite:

INTRODUCTION

The quality of the instrument items is recognized as one of the most influential factors on the measurement results. Inaccuracies in the development and arrangement of instrument items lead to biased measurement results. The development of test theory so far cannot be separated from the analysis of the quality of the items and the instrument itself. In modern test theory such as Item Response Theory (IRT), the fulfillment of the underlying assumptions in the models used is also very dependent on the quality of the items. The quality of these items could be assessed from two perspectives, namely content and statistics (Paek et al., 2021). A test item developer should have good knowledge of item development to produce good items in terms of content quality. Meanwhile, the item parameters yielded from the analysis by applying IRT such as difficulty level (*b*), discriminating power (*a*), and pseudo-guessing (*c*) can be examined statistically if the model used satisfies the assumptions (Hambleton & Swaminathan, 1985; Retnawati, 2014).

Methods for assessing the extent to which test takers' responses to each item fit a test measurement model, often referred to as person fit statistics, have been investigated by several researchers in recent decades (e.g., Cui & Mousavi, 2015; Dodeen & Darabi, 2009; Meijer & Sijtsma, 2001; Mousavi et al., 2019; Pan & Yin, 2017). Person fit evaluates the suitability of the test taker's response pattern with the expected response pattern of an IRT model (Djidu & Retnawati, 2022; Hambleton & Swaminathan, 1985). In other words, person fit is used to detect anomalies or deviations in the test taker's response pattern (Pan & Yin, 2017). As an illustration, if a test taker answers correctly on a difficult item but fails to answer correctly on an easier item, then this test taker's response pattern would be considered a misfitting response (Cui & Mousavi, 2015). Thus, test results of items with poor person fit would fail to provide accurate information about the test taker's measured ability.

Meijer and Sijtsma (2001) conducted a comparative study of more than 40 statistics to evaluate person fit methods. They have concluded that the methods available at that time had weaknesses. Among the weaknesses they point out are that the method used to determine person fit is influenced by three things, namely the type of misfitting behavior, the value of the person parameter, and the length of the test. In addition, one of the factors that influence the threshold value to determine person fit is item scoring (Meijer & Sijtsma, 2001). Evaluation of scoring quality is one of the aspects recommended by the Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) as cited by Wind and Walker (2019). Test developers and users should monitor and properly document the scoring carried out so that the results of the assessment resulting from the implementation of the test could be accounted for its quality (Wind & Walker, 2019).

The current study aims to evaluate the effect of scoring methods and the suitability of the polytomous IRT model on the estimation of ability and person fit by utilizing examination results data at the *Universitas Terbuka* (UT) or Indonesia Open University. The examination is known as the Take Home Exam (THE), which is an online examination held at the end of the semester. Students could start taking tests on the examination six hours after the test items were uploaded or available on the online learning platform used by UT. Student work on the examination was scored by the respective subject lecturer through the online learning platform. The final score of students so far was obtained by adding up the scores of each item using the Classical Test Theory (CTT) approach. Thus, in this study, three IRT models, namely the Partial Credit Model (PCM), Graded Response Model (GRM), and the Generalized Partial Credit Model (GPCM) were used to estimate students' abilities from the results of the Statistics course test.

METHOD

Design and Data Source

This is an exploratory descriptive study with a quantitative approach. The data source came from the results of the Statistics course test (denoted by SATS4410) as part of the Take Home Exam (THE) at the *Universitas Terbuka* (UT) or Indonesia Open University in 2022. The test carried out using a take home model with four essay (constructed response) questions. A total of 165 students took this test. Students' work on test was collected by them through the learning management system owned by the UT. The student's work on the test was then scored by the lecturer or rater with a maximum score of 25 for each item.

Data Analysis

In accordance with the purpose of this study, this study would explore the effect of scoring correction and model fit on the results of the estimation of ability parameter and person fit. Therefore, data analysis was generally carried out in several stages. First, we converted the rater's assessment results to the exam results using a scale of 0 – 5. Conversion of scores from the rating given by the rater was done using the following equation: $X_n = [\![X_r/5]\!]$, where X_n is the score of the conversion result and X_r is the score of the student's answer from the rater. The student scores obtained from this conversion result into initial data (stored under the name Data 4410) in the analysis process in this study. Second, the frequency of data on each item was calculated to obtain the distribution of student scores. Three models in the Item Response Theory (IRT) approach, namely Partial Credit Model (PCM), Graded Response Model (GRM), and Generalized Partial Credit Model (GPCM) were used to obtain information about model fit, ability estimation, and person fit. Afterwards, the scoring correction was carried out by considering the frequency of

ISSN: 2460-6995 (Online)

scores and the curve of Category Response Function (CRF) generated in the results of the first stage of data analysis, namely score conversion. The results of the scoring correction (the data were named as Data 4410BB) were analyzed with three IRT models (i.e., PCM, GRM, and GPCM) to obtain information about model fit, ability estimation, and person fit. The fit of the model was based on the *p*-value and the Root Mean Square of Error (RMSE).

The effect of scoring corrections and the IRT model used on the results of the estimation of the student's ability and person fit was tested using multivariate analysis. The analysis involved two factors (i.e., scoring correction and the IRT model used) and two dependent variables (i.e., ability (θ) and person fit). This multivariate analysis was used as a basis for inferring whether or not there were differences in the results of the ability estimation and the suitability of the response pattern or student scores after scoring correction from the estimation results using three different polytomous IRT models (i.e., PCM, GRM, and GPCM). Analysis of model fit, ability estimation, and person fit of the data was carried out using the 'mirt' package available in RStudio (Chalmers, 2021; R Core Team, 2022), while the multivariate analysis was carried out using the MANOVA function available in the 'stats' package (Revelle, 2022; Sarkar, 2008).

FINDINGS AND DISCUSSION

Findings

Findings of the Analysis of Test Items Using the Polytomous IRT Approach

The results of the analysis presented in Table 1 show that the proportion of the score 0 in items 1, 2, and 3 is very small and does not even reach 1%. In addition, in items 1 and 4, it can be seen that the proportions for the minimum score (0) and maximum score (5) do not reach 1%. Even in item 4, most of the scores are distributed on a score of 1 to a score of 3.

Table 1. Frequency of Student Scores on Statistics Course Test on Take Home Exam (Data 4410)

Score	Item 1		Item 2		Item 3		Item 4	
	Frequency	Prop.	Frequency	Prop.	Frequency	Prop.	Frequency	Prop.
0	1	0.01	1	0.01	6	0.04	13	0.08
1	2	0.01	21	0.13	32	0.19	82	0.50
2	47	0.28	59	0.36	52	0.32	28	0.17
3	61	0.37	19	0.12	17	0.10	23	0.14
4	51	0.31	21	0.13	34	0.21	8	0.05
5	3	0.02	44	0.27	24	0.15	11	0.07

Based on the Classical Test Theory (CTT) approach, the results of the analysis shown in Table 1 also represent the level of difficulty of the items. For instance, in items 1 and 4, the frequency (and proportion) of the score 5 is very small (less than 1%) of the number of students taking the test. This means that most students were not able to obtain the maximum score on the two items. Even in item 4, it can be seen that the score 4 has a small proportion as well. It is different with items 2 and 3, which show the proportion for a score of 5 which is relatively larger when compared to the other two items, namely 27% for item 2 and 15% for item 3. Based on these results, items 1 and 4 are more difficult than items 2 and 3. Meanwhile, based on the Item Response Theory (IRT) approach, the distribution of the frequency of these scores does not necessarily represent the level of difficulty of the items. Estimation of the probability of students to answer correctly or achieving a certain maximum score would be the basis for determining the level of difficulty of the items. Furthermore, information about the characteristics of these items would affect the results of the estimation of the ability of Statistics course test takers at the UT.

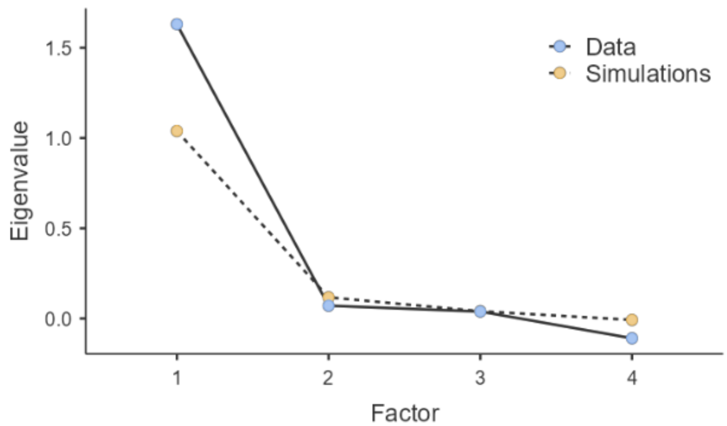


Figure 1. Results of Factor Analysis on Data 4410

Estimation of instrument reliability has also been carried out by calculating Cronbach's alpha and obtained a reliability estimate of 0.685. The result of the estimated reliability obtained according to Miller et al. (2009) and Nitko and Brookhart (2014) has been accepted for the instrument used in small-scale assessment (classroom assessment). In addition, factor analysis was also carried out before carrying out the analysis using the IRT approach to provide evidence that the instrument only measures one dominant dimension so that the unidimensional assumption is satisfied. The result of factor analysis showed that there is only one dominant factor measured by the instrument used in the Statistics course test (see Figure 1).

Table 2. Results of Analysis on Initial Data (Data 4410) Using PCM, GRM, and GPCM

Item	PCM			GRM			GPCM		
	RMSEA	***p***	**Model Fit**	**RMSEA**	***p***	**Model Fit**	**RMSEA**	***p***	**Model Fit**
1	0.048	0.164	Fit	0.00	0.455	Fit	0.00	0.475	Fit
2	0.047	0.142	Fit	0.00	0.609	Fit	0.00	0.514	Fit
3	0.061	0.040	Not Fit	0.05	0.153	Fit	0.05	0.154	Fit
4	0.058	0.069	Fit	0.00	0.515	Fit	0.00	0.633	Fit

The results of the analysis on Data 4410 using three polytomous IRT models indicate that the four test items fit with the GRM and GPCM models. Meanwhile, item 3 does not fit with the PCM model. Based on the RMSEA and *p*-value, all items showed a better fit with the GPCM model than with the GRM model. Three of the four items (i.e., items 1,3, and 4) showed a higher *p*-value although with a small difference. Furthermore, the analysis on the Data 4410 using PCM, GRM, and GPCM also resulted in the characteristics of the items in the form of difficulty level (*a*) and discriminating power (*b*) which are shown in Table 3. The results of the initial data analysis are in line with the model fit analysis as shown in Table 2. As we mentioned earlier, the level of difficulty indicates the probability of a student with a certain ability (θ) to obtain a score of 1, 2, 3, 4, or 5. Table 3 shows the level of the difficulty of the four test items obtained after being analyzed with the PCM, GRM, and GPCM models.

Table 3 shows that with the GRM model, the difficulty level for attaining a higher score is always greater than the difficulty level for attaining a lower score. However, in PCM and GPCM models, this condition does not always apply. For instance, the $b2$ parameter as a result of analysis with PCM model in the first item shows a value of –4.702 which is much smaller than $b1$ which is –2.35. Likewise, the $b2$ parameter generated by the GPCM model in the first item shows

ISSN: 2460-6995 (Online)

a lower value than b1, namely –6.939 for *b*2 and –2.483 for *b*1. That result is different from the GRM model which produces *b*2 = –6.19, which is larger than *b*1 = –7.884. A situation that is not much different is also shown in the *b* parameter on other items as has been given an asterisk (*) in Table 3. After scoring correction, the *b* parameter for all items shows an order from the smallest to the largest. The higher the score, the higher the difficulty level for all models (i.e., PCM, GRM, and GPCM). This condition is different from the characteristics of the *b* parameter based on the analysis of the initial data, where a high score does not always have a greater difficulty level than a low score. This condition demonstrates the most obvious impact of the score changes that occur after the scoring correction is made.

Table 3. Item Difficulty and Discriminating Power

Model	Parameter	Item 1		Item 2		Item 3		Item 4	
		Before Scoring Correction	After Scoring Correction	Before Scoring Correction	After Scoring Correction	Before Scoring Correction	After Scoring Correction	Before Scoring Correction	After Scoring Correction
PCM	*a*	1	1	1	1	1	1	1	1
	*b*1	–2.35	–5.21	–4.317	–4.674	–2.699	–2.971	–2.496	–2.651
	*b*2	–4.072*	–0.663	–1.715	–1.793	–1.005	–1.088	0.983	0.41
	*b*3	–0.572	0.561	0.916	–0.258	1.052	0.126	0.591*	2.643
	*b*4	0.47	4.049	0.12*	-	–0.321*	1.603	1.868	-
	*b*5	3.732	-	–0.09*	-	1.152	-	0.878*	-
GRM	*a*	0.673	0.724	2.835	3.069	6.036	3.324	0.941	0.983
	*b*1	–7.884	–7.386	–2.918	–2.922	–1.764	–1.989	–2.96	–2.862
	*b*2	–6.19	–1.225	–1.282	–1.27	–0.805	–0.875	0.332	0.336
	*b*3	–1.288	1.147	–0.073	–0.047	0.095	0.128	1.24	3.038
	*b*4	1.221	5.824	0.297	-	0.423	1.226	2.392	-
	*b*5	6.214	-	0.744	-	1.134	-	3.123	-
GPCM	*a*	0.502	0.492	2.062	2.803	2.772	2.692	0.468	0.775
	*b*1	–2.483	–8.556	–3.225	–2.914	–1.995	–1.983	–4.402	–2.937
	*b*2	–6.939*	–0.677	–1.363	–1.259	–0.846	–0.849	2.2	0.477
	*b*3	–0.743	0.562	0.301	–0.072	0.347	0.107	0.69*	2.844
	*b*4	0.563	6.386	0.22*	-	0.22*	1.212	2.879	-
	*b*5	6.286	-	0.427	-	1.15	-	0.28*	-

Note. * : The value of *b*n (n = 1, 2, 3, 4, 5) is smaller than the value of *b*(n – 1)

The situation that occurs with the *b* parameter in the test items used can be explained by referring to the Table 1 which demonstrates student score analysis results for each item. In the GRM model, the level of difficulty to obtain a higher score is always higher than that of a lower score. This situation is caused by the different assumptions used by PCM and GPCM with GRM in terms of the nature of the scores, namely whether they are ordered/leveled or not (Djidu et al., 2022; Retnawati, 2014). The difference in the proportion of obtaining a score at a certain step is in line with the *b*n parameter in the PCM and GPCM models. In the fourth item, it can be seen that *b*5 is always smaller than *b*4 for the PCM and GPCM models which is also in line with the proportion of obtaining the score of 5 on the item is higher than the proportion of obtaining the score of 4.

ISSN: 2460-6995 (Online)

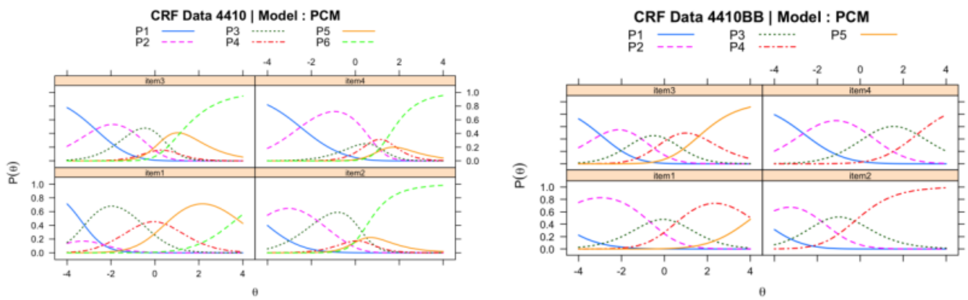


Figure 2. CRF for the Data 4410 and 4410BB Estimated Using PCM

The item parameters which include difficulty level (b) and discriminating power (a) are also shown by CRF (see Figure 2, Figure 3, and Figure 4). Figure 2 shows that in the first item (Data 4410), the P1 curve is above the P2 curve. In other words, it can be said that on this item, a student with a certain level of ability has a greater chance of obtaining a score of P1 than obtaining a score of P2. The same situation also occurs in the first item in the GPCM model as shown in Figure 3, where the position of the P1 curve tends to be above the P2 curve, which means that the probability of a student with a certain level of ability to obtain a score of P1 would be greater than his probability of obtaining a score of P2.

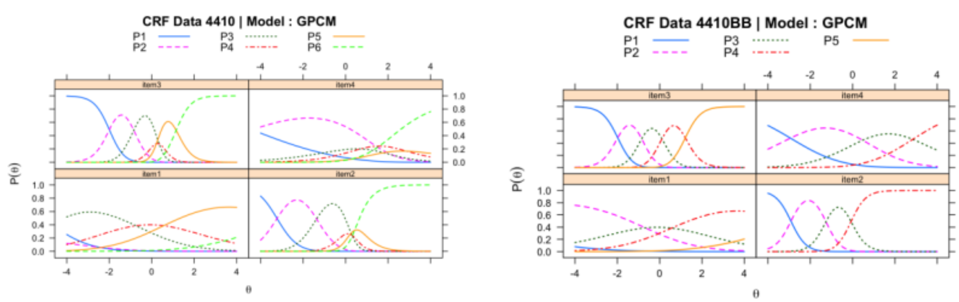


Figure 3. CRF for the Data 4410 and 4410BB Estimated Using GPCM

The results of the analysis on Data 4410 also have shown that certain scores have a lower probability of being obtained than the probability of obtaining other scores. In the PCM and GPCM models, the lowest probability of obtaining a certain score occurs at a score of 1 (item 1), at a score of 3 (item 2), at a score of 3 (item 3), and at a score of 4 (item 4) . The same thing is also shown in the curve formed from the analysis with the GRM model.

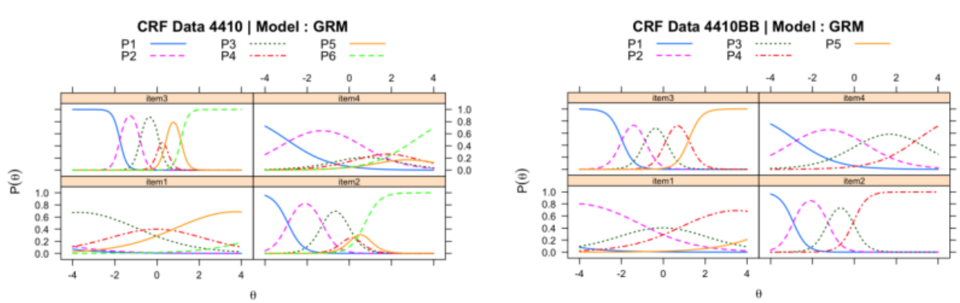
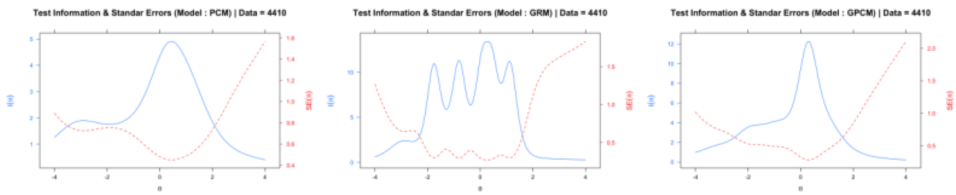
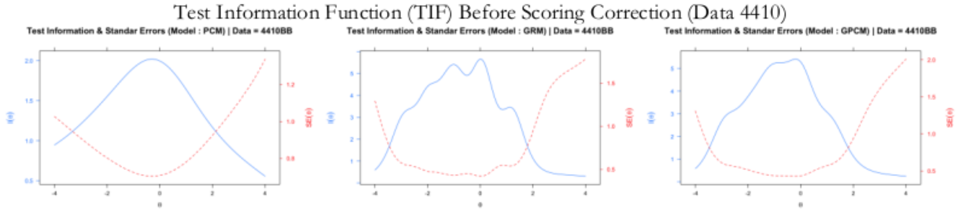


Figure 4. CRF for the Data 4410 and 4410BB Estimated Using GRM

Furthermore, the test information function and standard error yielded from the analysis on Data 4410 using the PCM, GPCM, and GRM models (see Figure 5) demonstrate that the maximum value of the information function in the PCM model is in the theta range from 0 to 1. In addition, the graph of the information function generated by the GRM model (see Figure 5) oscillates in the theta range from –2 to 1. Furthermore, the analysis results have shown that the highest total information is yielded in the modeling with GRM.



Test Information Function (TIF) Before Scoring Correction (Data 4410)



Test Information Function (TIF) After Scoring Correction (Data 4410BB)

Figure 5. Test Information Function (TIF) Before and After Scoring Correction

Table 4 presents the results of scoring corrections made to student test scores. The scoring correction on the first item was carried out by reducing the scores 2 and 1 to a score of 1, then, the scores 3, 4, and 5 are each made into a score of 2, 3, and 4, respectively. The scoring correction on the second item was done by reducing the scores from 3 to 5 to a score of 3 so that the maximum score on item 2 is 3. The maximum score on item 3 is also 3 which was obtained after the scoring correction was carried out by reducing the scores 3, 4, and 5 to a score of 3. Meanwhile, the scoring correction on item 4 was done by reducing the scores 2, 3, and 4 to a score of 2 so that the maximum score for that item after scoring correction is 3.

Table 4. Frequency of Student Scores on Statistics Course Test on Take Home Exam (Data 4410BB)

Score	**Item 1**		**Item 2**		**Item 3**		**Item 4**	
	Frequency	**Prop.**	**Frequency**	**Prop.**	**Frequency**	**Prop.**	**Frequency**	**Prop.**
0	1	0.01	1	0.01	6	0.04	13	0.08
1	49*	0.29*	21	0.13	32	0.19	82	0.50
2	61	0.37	59	0.36	52	0.32	59*	0.36*
3	51	0.31	84*	0.52*	75*	0.46*	11	0.07
4	3	0.02	-	-	-	-	-	-

Note. * : The results of scoring correction

The results of the model fit test on all items showed no difference with the results of the analysis of Data 4410, especially on the GRM and GPCM models. The *p*-value and RMSEA for these two models did not appear to experience significant changes in the GRM and GPCM models. On the other hand, the PCM model depicts a difference in the number of fit items. The first and second items experienced a significant decrease in *p*-value from 0.164 to 0.039 and 0.142 to 0.002, respectively. However, the scoring correction on the item 4 shows an increase in the *p*-value to 0.46, from the initial condition of 0.069 (Data 4410). Changes in RMSEA and *p*-values

ISSN: 2460-6995 (Online)

resulting from the model fit test on Data 4410BB are shown in Table 5. The values in brackets with the signs "▲" and "▼" indicate changes in both RMSEA and *p*-value compared to the results of the model fit test on Data 4410.

Table 5. Model Fit After Scoring Correction (Data 4410BB)

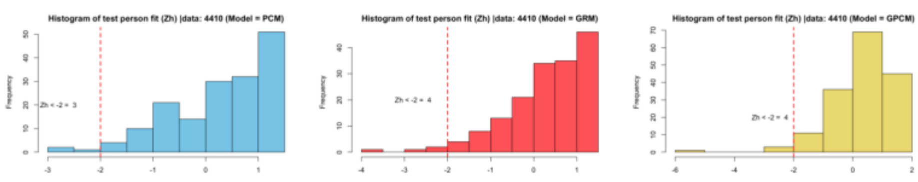
Item	PCM			GRM			GPCM		
	RMSEA	***p***	**Model Fit**	**RMSEA**	***p***	**Model Fit**	**RMSEA**	***p***	**Model Fit**
1	0.079 (▲0.03)	0.039 (▼0.13)	Not Fit	0.00 (0)	0.598 (▲0.14)	Fit	0.00 (0)	0.550 (▲0.08)	Fit
2	0.122 (▲0.08)	0.002 (▼0.14)	Not Fit	0.00 (0)	0.518 (▼0.09)	Fit	0.00 (0)	0.513 (0)	Fit
3	0.080 (▲0.02)	0.038 (0)	Not Fit	0.067 (▲0.02)	0.124 (▼0.03)	Fit	0.07 (▲0.02)	0.109 (▼0.05)	Fit
4	0.000 (▼0.06)	0.463 (▲0.39)	Fit	0.00 (0)	0.585 (▲0.07)	Fit	0.00 (0)	0.717 (▲0.08)	Fit

Note. : Increase in the value after scoring correction, : Decrease in the value after scoring correction

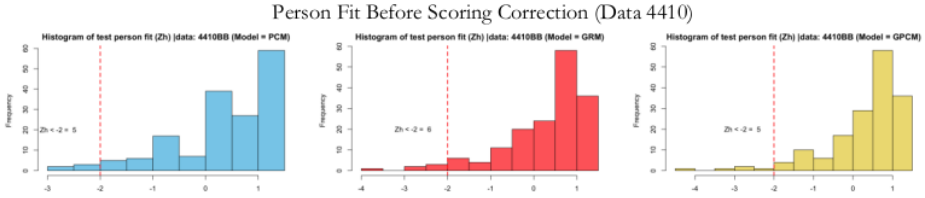
The scoring correction that has been made can also be seen from the CRF formed from the Data 4410BB (see Figure 2, Figure 3, and Figure 4) which also experienced changes. Changes are also found in the information function (see Figure 5), where the amount of information generated in a certain range of abilities (θ) is more stable (monotone increases at wider intervals). Significant changes can also be seen in the graph of the information function generated from the GRM model (see Figure 5). The total information generated by each model is 13.982 (PCM), 25.063 (GRM), and 23.226 (GPCM).

Findings of the Multivariate Analysis on the Effect of Scoring Correction on the Estimation of Ability Parameter and Person fit

This section provides answers to questions related to the effect of the scoring correction made on student test results on the results of theta estimate and person fit using the Polytomous IRT approach. The results of the model fit analysis in Table 5 have shown how the item fit to the model has changed after scoring correction is made (i.e., Data 4410 becomes Data 4410BB). The multivariate analysis in this section was intended to examine whether changes in the scoring made on the initial data (i.e., Data 4410) have an effect on the estimation of ability parameter (θ) and person fit.



Person Fit Before Scoring Correction (Data 4410)



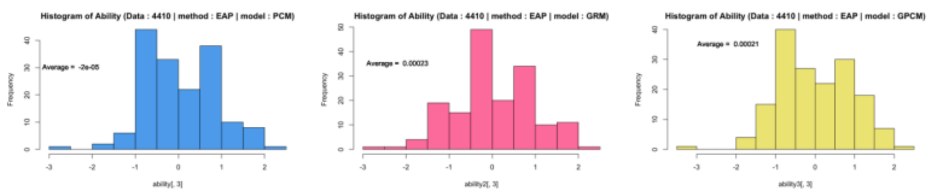
Person Fit After Scoring Correction (Data 4410BB)

Figure 7. Person Fit Before and After Scoring Correction

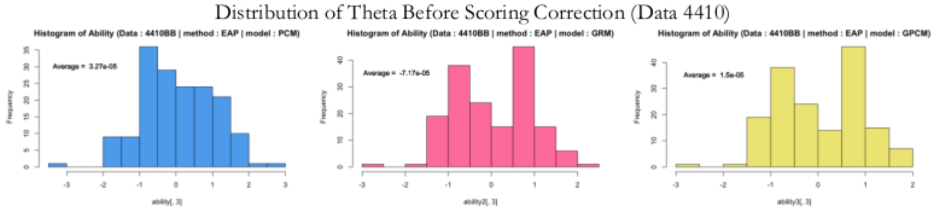
ISSN: 2460-6995 (Online)

Figure 7 shows in detail the change in the distribution of person fit before and after the scoring correction. The data has shown a decrease in the number of patterns of responses/scores that fit the three models after the scoring correction. However, the change in the number of person fit is not too significant, namely at most two response patterns from 165 students. The number of responses that do not fit the most is shown in the results of the analysis using the GRM model after the scoring correction which shows six patterns of responses/scores that do not fit the model.

The results of the multivariate analysis have demonstrated that the scoring correction does not affect the results of ability (θ) estimate and person fit in all models tested. In detail, the multivariate analysis showed that there was no significant difference between before and after the scoring correction in terms of estimation of the ability parameter and person fit based on the PCM model (Wilks' Lambda = 0.99, $F(1, 328) = 1.578$, $p = 0.208$), GRM model (Wilks' Lambda = 0.995, $F(1, 328) = 0.799$, $p = 0.45$), and GPCM model (Wilks' Lambda = 0.995, $F(1, 328) = 0.799$, $p = 0.45$). In other words, these results indicate that the scoring correction does not cause differences in the results of the estimation of the student's ability level and the model fit of the student's response/score in the Statistics course test.



Distribution of Theta Before Scoring Correction (Data 4410)



Distribution of Theta After Scoring Correction (Data 4410B)

Figure 8. Distribution of Student's Ability (θ) Before and After Scoring Correction

The ability estimation from the initial data (Data 4410) and after the scoring correction (Data 4410BB) showed results that were not much different, as the multivariate test which had obtained the result that the score correction did not affect the ability (θ) estimation. Figure 8 presents the distribution of scores on Data 4410 and Data 4410BB for the three models tested (i.e., PCM, GRM, and GPCM). The three models produce theta estimates that do not differ, both in distribution and in the average theta estimates. In addition, the histogram presented in Figure 8 also shows that PCM is a model that produces a different score distribution from the other two models. Changes in the distribution pattern of scores before and after scoring correction in the PCM model are also more contrasting than the GRM and GPCM models. These results are in line with the results of multivariate analysis, where the p value yielded based on the PCM model is the smallest (i.e., 0.21) compared to the p value yielded based on the GPCM and GRM models (i.e., 0.45 for each). Therefore, although the scoring correction did not produce a significant difference in the estimation of ability and person fit based on the three existing models, it can be said that the PCM model is the most affected by the scoring correction than the other two models.

Lastly, we examined the effect of scoring corrections (i.e., Data 4410 and Data 4410BB) and the polytomous IRT models used (i.e., PCM, GRM, and GPCM) on ability (θ) estimates and person fit. This analysis is a simultaneous form of the previous test that examines the effect of the scoring correction on the ability (θ) estimates and person fit partially in each of the tested models. The results of the analysis have shown that the scoring correction (Wilks' Lambda = 0.9992, $F(1, 984) = 0.381$, $p = 0.683$) and the polytomous IRT models used (Wilks' Lambda = 0.9998, $F(2, 984) = 0.036$, $p = 0.998$) do not significantly affect the estimation of student ability (θ) and person fit. The test of the effect of the interaction between the scoring correction and the polytomous IRT models used indicated that the interaction did not cause a significant difference in the ability (θ) estimate and person fit (Wilks' Lambda = 0.9999, $F(2, 984) = 0.001$, $p = 0.9999$).

Based on the results of the analysis described in the first and second sections, we found that the scoring correction does not affect the results of the ability estimate and the fit of the pattern of the responses/scores of students to the polytomous IRT models based on the results of student work on the Statistics course (Code: 4410) test which was part of Take Home Exam (THE) conducted by *Universitas Terbuka* (UT). However, the scoring corrections made have an effect on the item fit to the model (see Table 5). PCM is the model that is most affected by the scoring correction, which can be seen from the increase in the number of items that do not fit the model. In other words, when compared to the other two polytomous IRT models, i.e., GRM and GPCM, PCM is the model most affected by the scoring correction.

Discussion

The results showed that the scoring correction and model fit did not affect the estimation of ability parameter and person fit. The results of the ability estimation produced by the three examined polytomous IRT models were not significantly different. This result indicated that the estimation of a test taker's ability does not differ even though it is analyzed with different models. These results are consistent with the results of a study conducted by Si and Schumacker (2004) who found the similarity of the ability estimation results from several polytomous IRT models used. In their study, they only tested the differences in the results of the ability parameter estimates separately, while the results of this study showed that the model and scoring corrections also had no effect on the ability estimates.

The difference found was precisely the fit of the model after the scoring correction. The results of the model fit test showed that there was a difference in the PCM model, where the scoring correction actually resulted in two items failed to fit adequately with the model. Unlike the PCM, the GRM and GPCM models did not change the number of fit items after the scoring correction. The results of the multivariate test on the results of the ability estimate and person fit that have been carried out also demonstrated that the GRM and GPCM models have a higher level of similarity than those generated from the PCM model.

The GRM model showed the best model fit index among the other two models based on RMSEA and p-value. This means that the discriminating power parameter (a) of the items has a large contribution to the ability estimation. If it is reviewed from the scoring model given to student test results, the GRM model was more suitable because each score was graded so that the level of difficulty in achieving a higher score on an item would increase. This can be seen from the difficulty level parameter (b) which increases with higher scores in the GRM model. However, this did not apply to PCM and GPCM models. A higher score may have a lower level of difficulty than a lower score because PCM and GPCM use the assumption that the score is partial and not tiered (Djidu et al., 2022; Hambleton et al., 1991; Linden & Hambleton, 1997).

The evaluation of the number of response patterns that fit the three models tested in this study demonstrated that there were additional unfit response patterns. However, the resulting capability estimates were not significantly different. Based on this perspective, the results of this study are relevant to the results of the study of Dodeen and Darabi (2009) which found that person fit and cognitive performance of test takers in mathematics did not have a significant relationship. They discovered non-cognitive aspects such as motivation and attitudes that greatly affect person fit. Equivalent results were also found in the study conducted by Spoden et al. (2020) that found a significant influence on non-cognitive aspects such as anxiety and motivation to do mathematics tests on person fit. In other words, the mismatch of response patterns to the IRT model used indicates that the respondent is not serious in responding and does not always reflect the respondent's level of ability.

Other studies, one of which is a study conducted by Wind and Walker (2019) found a very close relationship between scoring and person fit. Their study highlights the scores given by some raters to essays written by students and the effect of such scoring on person fit. The results of the study by Wind and Walker (2019) differ from the results of this study for three reasons. First, the aspect measured in the study by Wind and Walker (2019) was writing ability, while this study was focused on the result of student examinations in the Statistics course which is very closely related to mathematics. Second, the scoring correction in the current study was carried out on all student responses based on the CRF and the results of the analysis of the frequency of student scores, while the scoring correction in their study was done on certain scores only based on the misfit of the response patterns of the students. Third, the context of measurement in this study focused on the cognitive aspect in the form of students' ability to solve questions in the form of a test, while the focus of measurement in the study conducted by Wind and Walker (2019) was students' performance in writing essays.

Although the scoring correction and the polytomous IRT model used in analyzing student examination results in this study showed no significant differences, the results of this study still need to be strengthened using a larger number of respondents. In addition, the number of questions in this study is only four so that the evaluation of the suitability of the items may be different if applied to a test with a larger number of items. Furthermore, the number of responses or answers of the test takers used can be analyzed by using more responses so that the evaluation of person fit and ability estimation is more comprehensive.

3

CONCLUSION

Based on the description of the results of this study, it was concluded that the correction of scoring and model fit did not affect the estimation of ability parameter and person fit in the polytomous IRT model. The GRM and GPCM models produce more stable estimates than PCM, although all three produced estimates that are not significantly different. The results of this study emphasize the importance of the quality of scoring in the administration of the test. Score correction in this study was carried out systematically based on the CRF obtained from the analysis using the IRT approach. In addition, score correction was also applied to all test takers because the scoring correction did not intend to improve or re-score the scores of certain test takers. As a result, there is no difference in the estimation of ability parameter and person fit. This result could be a reference for simplification in providing scores to make it easier for lecturers or teachers who would administer tests in the future. A test that has been piloted with certain scoring categories or levels could be re-evaluated regarding its scoring model used by simplifying the number of scoring categories or levels which could accelerate the next scoring process.

REFERENCES

Chalmers, R. P. (2021). *mirt: Multidimensional item response theory*. https://cran.r-project.org/package=mirt

Cui, Y., & Mousavi, A. (2015). Explore the usefulness of person-fit analysis on large-scale assessment. *International Journal of Testing*, *15*(1), 23–49. https://doi.org/10.1080/15305058.2014.977444

Djidu, H., Ismail, R., Sumin, Rachmaningtyas, N. A., Imawan, O. R., Suharyono, Aviory, K., Prihono, E. W., Kurniawan, D. D., Syahbrudin, J., Nurdin, Marinding, Y., Firmansyah, Hadi, S., & Retnawati, H. (2022). *Analisis instrumen penelitian dengan pendekatan teori tes klasik dan modern menggunakan program* R [*Analysis of research instruments with classical and modern test theory approaches using the* R *program*]. UNY Press.

Djidu, H., & Retnawati, H. (2022). IRT unidimensi penskoran dikotomi [Unidimensional IRT for dichotomous scoring]. In H. Retnawati & S. Hadi (Eds.), *Analisis instrumen penelitian dengan pendekatan teori tes klasik dan modern menggunakan program* R [*Analysis of research instruments with classical and modern test theory approaches using the* R *program*] (pp. 89–141). UNY Press.

Dodeen, H., & Darabi, M. (2009). Person-fit: Relationship with four personality tests in mathematics. *Research Papers in Education*, *24*(1), 115–126. https://doi.org/10.1080/02671520801945883

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Nijhoff Publishing.

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

Linden, W. J. van der, & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer Science + Business Media. https://doi.org/10.1007/978-1-4757-2691-6

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching*. Pearson Education.

Mousavi, A., Cui, Y., & Rogers, T. (2019). An examination of different methods of setting cutoff values in person fit research. *International Journal of Testing*, *19*(1), 1–22. https://doi.org/10.1080/15305058.2018.1464010

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (6th ed.). Pearson.

Paek, I., Liang, X., & Lin, Z. (2021). Regarding item parameter invariance for the Rasch and the 2-parameter logistic models: An investigation under finite non-representative sample calibrations. *Measurement: Interdisciplinary Research and Perspectives*, *19*(1), 39–54. https://doi.org/10.1080/15366367.2020.1754703

Pan, T., & Yin, Y. (2017). Using the bayes factors to evaluate person fit in the item response theory. *Applied Measurement in Education*, *30*(3), 213–227. https://doi.org/10.1080/08957347.2017.1316275

R Core Team. (2022). *R: A language and environment for statistical computing*. https://www.r-project.org/

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana* [*Advanced item response theory and its applications: For researchers, measurement and testing practitioners, graduate students*]. Parama Publishing.

Revelle, W. (2022). *psych: Procedures for psychological, psychometric, and personality research*. https://personality-project.org/r/psych/

Sarkar, D. (2008). *Lattice: Multivariate data visualization with* R. Springer. http://lmdvr.r-forge.r-project.org

Si, C.-F., & Schumacker, R. E. (2004). Ability estimation under different item parameterization and scoring models. *International Journal of Testing*, *4*(2), 137–181. https://doi.org/10.1207/s15327574ijt0402_3

Spoden, C., Fleischer, J., & Frey, A. (2020). Person misfit, test anxiety, and test-taking motivation in a large-scale mathematics proficiency test for self-evaluation. *Studies in Educational Evaluation*, *67*, 1–7. https://doi.org/10.1016/j.stueduc.2020.100910

Wind, S. A., & Walker, A. A. (2019). Exploring the correspondence between traditional score resolution methods and person fit indices in rater-mediated writing assessments. *Assessing Writing*, *39*, 25–38. https://doi.org/10.1016/j.asw.2018.12.002

Manuscript

ORIGINALITY REPORT

10%	8%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.uny.ac.id Internet Source	2%
2	docslib.org Internet Source	1%
3	ismei.qitepinmath.org Internet Source	1%
4	eprints.sinus.ac.id Internet Source	1%
5	worldwidescience.org Internet Source	1%
6	www.uniting4kids.com Internet Source	<1%
7	jurnal.fkip.unila.ac.id Internet Source	<1%
8	Ernest H. O'Boyle, Larry J. Williams. "Decomposing model fit: Measurement vs. theory in organizational research using latent variables.", Journal of Applied Psychology, 2011 Publication	<1%

9	anketimvar.net Internet Source	<1%
10	Eisa Abdul-Wahhab Al-Tarawnah, Mariam Al-Qahtani. "The Effect of Test Length on the Accuracy of Estimating Ability Parameter in the Two- and Three-Parameter Logistic Models: Comparison by Using the Bayesian Method of Expected Prior Mode and Maximum Likelihood Estimation", Journal of Educational and Social Research, 2022 Publication	<1%
11	Menghua She, Yaling Li, Dongbo Tu, Yan Cai. "Computerized Adaptive Testing for Sleep Disorders: Development of An Item Bank and Validation in A Simulated Study", Research Square Platform LLC, 2020 Publication	<1%
12	Submitted to Program Pascasarjana Universitas Negeri Yogyakarta Student Paper	<1%
13	sumc.lt Internet Source	<1%
14	openaccess.leidenuniv.nl Internet Source	<1%
15	Submitted to Rutgers University, New Brunswick Student Paper	<1%

16	Submitted to Ohio University Student Paper	<1 %
17	livrepository.liverpool.ac.uk Internet Source	<1 %
18	pdfs.semanticscholar.org Internet Source	<1 %
19	Rob Meijer. "Detection and Validation of Unscalable Item Score Patterns Using Item Response Theory: An Illustration with Harter's Self-Perception Profile for Children", Journal of Personality Assessment, 5/2008 Publication	<1 %
20	www.estosite.org Internet Source	<1 %
21	www.mdpi.com Internet Source	<1 %
22	www.pegegog.net Internet Source	<1 %
23	doaj.org Internet Source	<1 %
24	repository.kaust.edu.sa Internet Source	<1 %
25	www.ijstr.org Internet Source	<1 %

26	C. A. W. Glas, Anna Villa T. Dagohoy. "A Person Fit Test For Irt Models For Polytomous Items", Psychometrika, 2006 Publication	<1 %
27	Stefanie A. Wind. "Detecting Rating Scale Malfunctioning With the Partial Credit Model and Generalized Partial Credit Model", Educational and Psychological Measurement, 2022 Publication	<1 %
28	ioinformatics.org Internet Source	<1 %
29	journal.unnes.ac.id Internet Source	<1 %
30	jurnal.utu.ac.id Internet Source	<1 %
31	web.archive.org Internet Source	<1 %
32	Christian Spoden, Jens Fleischer, Andreas Frey. "Person misfit, test anxiety, and test-taking motivation in a large-scale mathematics proficiency test for self-evaluation", Studies in Educational Evaluation, 2020 Publication	<1 %

33	Ching-Fung Si, Randall E. Schumacker. "Ability Estimation Under Different Item Parameterization and Scoring Models", International Journal of Testing, 2004 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Manuscript

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12