JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v13i2.40344
http://journal.uinjkt.ac.id/index.php/jp3i

Mplus and the R mirt Package: A Comparison of Model Parameter Estimation for Generalized Partial Credit Model (GPCM)

Arif Budiman Al Fariz
Faculty of Psychology, Universitas Gadjah Mada
arifbudimanalfariz@mail.ugm.ac.id

Abstract

This article aims to carry out an empirical demonstration to calibrate data using the generalized partial credit model (GPCM) and compare the results of GPCM analysis on paid software, namely Mplus, and open-source software, namely R Package Mirt. The data used in this study used secondary data in the form of item scores from the future orientation scale or *Skala Orientasi Masa Depan* (S-OMD) with a total of 326 participants using a Likert scale with 4 response options. The results of this study show that GPCM is fit for OMD scale data. A comparison of analysis results using Mplus and R Package mirt shows the same output, such as discrimination parameters and item difficulty levels. The resulting factor score correlation also has a perfect correlation, or the coefficient of correlation is one. In conclusion, open-source software can have the same computing performance as paid software and even has several additional features not found in paid software, especially in the context of GPCM calibration.

Keywords: Generalized partial credit model, irt, mirt, mplus

Abstrak

Tujuan dari penelitian ini adalah melakukan demonstrasi empiris untuk melakukan kalibrasi data menggunakan generalized partial credit model (GPCM) dan perbandingan antara hasil analisis GPCM pada software berbayar yaitu Mplus dengan software open-source yaitu R Package mirt. Data yang digunakan dalam penelitian ini adalah data sekunder berupa skor butir dari skala orientasi masa depan (S-OMD) dengan jumlah partisipan sebanyak 326 orang menggunakan skala Likert dengan 4-opsi respons. Hasil penelitian menunjukkan GPCM fit terhadap data skala OMD. Perbandingan hasil analisis menggunakan Mplus dan R Package mirt menunjukkan output yang sama, seperti parameter diskriminasi dan tingkat kesukaran butir. Korelasi skor faktor yang dihasilkan memiliki korelasi sempurna atau korelasi sebesar satu. Kesimpulannya software open-source mampu memiliki performa komputasi yang sama dengan software berbayar, bahkan memiliki beberapa fitur tambahan yang tidak terdapat pada software berbayar terutama dalam konteks kalibrasi GPCM.

***Kata Kunci**: Generalized partial credit model, irt, mirt, mplus*

Introduction

Item response theory (IRT), also known as modern test theory, has been developed to overcome the limitations of classical test theory (CTT) (Mair, 2018). The CTT model generally conceptualizes observed test scores as a linear combination of correct and error scores, whereas CTT also assumes that measurement errors are equal for all persons (Wainer & Thissen, 2009; Zanon et al., 2016). Unlike CTT, IRT provides standard error measurement for each respondent (Bock & Gibbons, 2021). This feature shows that information produced by IRT is better than CTT.

Furthermore, IRT can estimate item parameters (i.e., item discrimination and item difficulty) as well as person parameters (i.e., ability level) on the same scale (Baker & Kim, 2004). In addition, the fit of the model to the data (goodness-of-fit), the fit of the items to the model (item fit), and the fit of the person to the model (person fit) can be tested with various statistic properties (Debelak, 2019; Maydeu-Olivares, 2015; Tay et al., 2011). Thus, IRT provides more in-depth psychometric evaluation results for an instrument and more detailed diagnostic information on improving the scale (Petrillo et al., 2015).

With various advantages, IRT-based methods have been used widely across disciplines in the social, behavioral, and health sciences (see Thissen & Steinberg, 1986). Furthermore, on a larger scale, IRT has become a popular methodological framework for modeling response data from large-scale assessments in education, such as the National Assessment of Educational Progress (NAEP) and the Maryland School Performance Assessment Program (MSPAP) (Thissen et al., 2009), Trends in International Mathematics and Science Study (TIMSS) (Yamamoto & Kulick, 2000) and The Programme for International Student Assessment (PISA) (OECD, 2024), where the use of modern test theories such as IRT or Rasch measurement theory, making the use of more CTT.

In fact, many models of IRT exist, which can be classified according to their dimensionality, number of categories per item, and number of parameters (Buchbinder et al., 2012); in various large-scale assessments, one of the most widely used IRT models is GPCM (Muraki, 1992; von Davier & Yamamoto, 2004). GPCM is widely used in analyzing polytomous data both in large-scale assessments and studies testing the psychometric properties of instruments with a polytomous score format (e.g., OECD, 2024; Schauberger & Mair, 2020; Wallmark et al., 2023; Wind, 2023; Yamamoto & Kulick, 2000), where the polytomous score in question includes various formats of attitude scale (e.g., Likert scale, Guttman scale) and various item types such as essay questions, forced-choice questions and other formats that produce polytomous scores (Muraki, 1992; Wu et al., 2016). One large-scale assessment that uses GPCM consistently from cycle to cycle is TIMSS. However, many researchers, especially in Indonesia, often need help understanding the optimal application of GPCM or reporting it properly (e.g., Kurnia, 2019; Samritin, 2018), although it cannot be denied that the use of GPCM is popular in Indonesia.

Because of implementation in large-scale assessments, the popularity of using IRT cannot be separated from the development of software used in the calibration process or data analysis (see Wang, 2018). Among the various available software, there is software that is commercial and requires a fee to use (for example, IRTPRO, BILOG-MG, Mplus, etc.), as well as software that is open source and can be used for free (for example, various packages in R programs). These programs can be used to perform GPCM analysis. Many studies have compared multiple software in performing GPCM analysis (Fu, 2020; Huggins-Manley & Algina, 2015).

A previous study compared three programs for analyzing GPCM, namely Mplus, SAS, and IRTPRO, where this study found that Mplus, SAS, and IRTPRO produced the same solution but with different completeness (Huggins-Manley & Algina, 2015). Apart from that, another study compared the analysis results of two models, namely 3-PL and GPCM, using several software such as Mplus, Xcalibre, IRTPRO, and ltm found that the quality of the software varied in resulted output calibration, for example, differences in estimates of the level of difficulty and the *d* parameter in GPCM (Wang, 2018). Other research also compared the results of multidimensional IRT model analysis using the IRTPRO program with Mplus and found that the two provided equivalent estimated parameter results (Sims, 2017). In contrast, a similar study was conducted in Indonesia comparing the estimation results of three software, namely Mplus, IRTPRO, and WINSTEPS but in the context of Rasch models and not IRT such as GPCM (Hayat et al., 2020). The latest Educational Testing Services (ETS) study compared five software for analyzing GPCM and found that flexMIRT was the most optimal software for this kind of study (Fu, 2020). Based on previous research, each software presents varying output quality, so reporting on the software used to perform data analysis is very important (Wang, 2018).

This study aims to provide empirical illustrations as part of introducing two types of calibration software, Mplus and Mirt, for the IRT model, GPCM. The software was chosen based on its popularity in data calibration, and many comparisons were conducted in previous studies. They represent paid and free application programs, and both software have their advantages. Mplus, although paid, has several advantages such as Mplus has advanced features in handling missing data such as multiple imputation (MI), full information maximum likelihood (FIML) and other traditional methods (Wang & Wang, 2020), able to handle various data such as normal, non-normal and censored data, there are various choices of integration types, estimators and algorithms types including Markov chain monte carlo (MCMC) for bayesian estimators and a user-friendly interface (Muthén & Muthén, 1997/2017). On the other hand, Mirt is a package of the R program with the advantage of being free to use. In addition, other advantages of Mirt are flexibility in modeling complex unidimensional and multidimensional IRT analysis, flexibility in IRT and traditional parameterization in GPCM, the capability of performing latent class analysis, and Exploratory and confirmatory models can be estimated with quadrature (EM) or stochastic (MHRM) methods (Chalmers, 2012).

Understanding different software features will provide vital information for researchers before applying GPCM, such as differences in IRT "tradition" (e.g., Hayat et al., 2020), differences in estimation methods (e.g., Finch & French, 2019), and differences in software features in general, such as plots, graphs, or user interface (e.g., Paek & Cole, 2020). Therefore, this study aims to compare the results of the GPCM analysis as in the study conducted by Huggins-Manley and Algina (2015) but using the comparison criteria used in the study conducted by Hayat et al. (2020) but with a different model, namely GPCM. In this research, the two programs compared are Mplus as commercial software and the R package 'mirt' as an open-source program that is free to use. The reason for choosing 'mirt' is because the estimation method is in line with Mplus (i.e., Bock & Aitkin, 1981), so comparisons can be made apples-to-apples without worrying about differences in estimation methods between the two programs. Another objective of this study is to provide empirical illustrations and to be learning materials for researchers and readers when calibrating data using GPCM using paid software represented by Mplus and free software represented by R Package Mirt.

Theoretical framework: GPCM

GPCM is an extension of PCM by estimating item discrimination parameters (Muraki, 1992). PCM (Masters, 1982) is a model part of the Rasch family, including dichotomous Rasch (Rasch, 1960) and RSM (Andrich, 1978), which assumes parallel item characteristic curves (ICC) or item discrimination parameter is constrained to 1 for all items (Andersen, 1973). Although GPCM is an extension of PCM by relaxing the item discrimination of items to vary, GPCM is not part of the Rasch analysis family. The following is the GPCM equation developed by Muraki (1992):

$$\Pr(X_{ni} = x) = \frac{exp \sum_{k=1}^{x} \alpha_i(\theta_n - \delta_{ik})}{\sum_{h=0}^{mi} exp \sum_{k=0}^{x} \alpha_i(\theta_n - \delta_{ik})} \quad (1)$$

The meaning of this formula is that the probability of a person (*n*) on item *i* to get score *x* (x = 0, 1, . . . k), which is the exponent of the summation α (item discrimination) item *i* multiplied by the result of subtracting θ (ability) of the person to *n* minus δ (item difficulty) item *i* in category *k* with *k*-1, with the denominator being the sum of all the numerators. Formula (1) is also said (e.g., Yen, 1993) as the two-parameter partial credit (2PPC) model (de Ayala, 2022). Formula (1) is implemented directly in software such as the R Package mirt with the gpcmIRT model or GPCM with classical IRT parameterization (Chalmers, 2012).

The GPCM formula for polytomous data or scores of more than two can be broken down into several score categories. The following is the equation for the chance of getting a score when using four scores or categories (x = 0, 1, 2, 3) with the principle that δ_0 is 0, so exp (0) is 1:

$$P0 = \Pr(x = 0) = \frac{1}{1 + exp\ \alpha(\theta - \delta 1) + exp\ \alpha\big(2\theta - (\delta 1 + \delta 2)\big) + exp\ \alpha(3\theta - (\delta 1 + \delta 2 + \delta 3))} \quad (2)$$

$$P1 = \Pr(x = 1) = \frac{exp\ \alpha(\theta - \delta 1)}{1 + exp\ \alpha(\theta - \delta 1) + exp\ \alpha\big(2\theta - (\delta 1 + \delta 2)\big) + exp\ \alpha(3\theta - (\delta 1 + \delta 2 + \delta 3))} \quad (3)$$

$$P2 = \Pr(x = 2) = \frac{exp\ \alpha\big(2\theta - (\delta 1 + \delta 2)\big)}{1 + exp\ \alpha(\theta - \delta 1) + exp\ \alpha\big(2\theta - (\delta 1 + \delta 2)\big) + exp\ \alpha(3\theta - (\delta 1 + \delta 2 + \delta 3))} \quad (4)$$

$$P3 = \Pr(x = 3) = \frac{exp\ \alpha(3\theta - (\delta 1 + \delta 2 + \delta 3))}{1 + exp\ \alpha(\theta - \delta 1) + exp\ \alpha\big(2\theta - (\delta 1 + \delta 2)\big) + exp\ \alpha(3\theta - (\delta 1 + \delta 2 + \delta 3))} \quad (5)$$

The interpretation of formula (2) is that it calculates the probability of a person receiving a score of 0 or being placed in category 0, given known values for theta, alpha, and delta. Similarly, formulas (3), (4), and (5) calculate the probability of a person obtaining a higher score. These formulas are derived from formula (1) when there are four score categories. The

results section of this study provides an illustration of how to read and interpret these formulas. The GPCM, as expressed in formula (1), can also be represented using the following formula (e.g., Asparouhov & Muthén, 2020; de Ayala, 2022; Reckase, 2009):

$$\Pr(X_{ni} = x) = \frac{exp \sum_{k=1}^{x} \alpha_i((\theta_n - \delta_i) + \tau_k)}{\sum_{h=0}^{mi} exp \sum_{k=0}^{x} \alpha_i((\theta_n - \delta_i) + \tau_i)} \quad (6)$$

The difference between formulas (1) and (6) lies in the parameterization, classical parameterization, and IRT parameterization, which can be seen in the thresholds. In formula (1), between steps have the same thresholds. Formula (6) describes the components of item location *i* and thresholds in categories between *k* and *k-1* or performs substitution, meaning replacing the level of difficulty of items between categories k and k-1 into the level of difficulty or location of items as a whole minus the thresholds in a particular category, with that 0 or the sum of all results. Formula (6) is also called a generalized rating scale (GRS) because of the constraints (de Ayala, 2022; Furr, 2017). Instead of discussing these differences, this study will try to demonstrate the results of the GPCM analysis using formula (1) or classical parameterization, which is implemented in this study. The difference between the two formulas is provided to make it clear as study material that GPCM sometimes uses IRT parameterization in some software because they have different interpretations.

Methods

Data Resource

This study utilizes secondary data from prior research by Putra and Tresniasari (2015), which introduced the *Skala Orientasi Masa Depan* (S-OMD). The S-OMD comprises eight items rated on a 4-point Likert scale. Before conducting the research, the researcher corresponded with the corresponding author and received permission to use the data for empirical illustration in this study. The dataset analyzed in this study is openly accessible at: https://osf.io/2uz7h/?view_only=cd6cf0ff41244fcd89f2a45d1449b94c

Analysis Program

Mplus

The Mplus software used in this research is version 8.3. In carrying out GPCM analysis using Mplus, it is based on the Syntax in Appendix 1. The Syntax was created as is generally the case using unidimensional CFA (see Muthén & Muthén, 1998/2017). In the VARIABLE command, the statement CATEGORICAL ARE u1-u8 (gpcm) is added, which indicates that the item is treated as category data, (gpcm) is needed to estimate the GPCM model in addition to IRT 2-PL. In the MODEL command, (*) is added to free the default constraint on the factor loading of the first item, and O@1 means that the variance of the factor is constrained to 1, as in the IRT tradition. If O*1 is used, calibration will be carried out using the Partial Credit Model (PCM). PLOT1 and PLOT3 are instructed to report useful plots in IRT. The computational method in Mplus in this analysis applies the robust maximum likelihood estimation method (ESTIMATOR = MLR), which is equivalent to the marginal maximum likelihood (MML) with expectation–maximization (EM) algorithm (Bock & Aitkin, 1981).

MIRT

In performing GPCM analysis using the 'mirt' package (Chalmers, 2012), several codes were needed (see Appendix 2). The code used in the first line is a command specification for calibrating the GPCM model with the statement "gpcmIRT." This statement functions to order GPCM calibration with conventional parameterization, and it will be different if the "GPCM" is used. The second line is a command to display the parameter output from GPCM. The third line is a command to estimate the person's ability parameters (factor scores) for each participant. The fourth line is a command to display the item characteristic curve (ICC) plot for all items. Meanwhile, the last line is a command to calculate the goodness-of-fit statistics, which is the M2 statistics (Maydeu-Olivares & Joe, 2005). The estimation method for all model parameters is the marginal maximum likelihood with an expectation–maximization (EM) algorithm (Bock & Aitkin, 1981).

Analysis Strategy

GPCM is used to compare the performance of two applications, Mplus and Mirt, using the formula (1). Some of the psychometric properties compared are following the testing of the required assumptions and parameters. First, dimensionality testing (Chou & Wang, 2010; Christensen et al., 2013) uses statistical fit indices produced by both software. Second, item discrimination and step parameters use IRT or traditional parameterization (Chalmers, 2012). The third is the item characteristic curve (ICC). The fourth is the test information function (TIF). Lastly is the correlation of factor scores or theta OMD scales produced by each software.

Results and Discussion

Results

Model fit index

After calibrating using the GPCM model on 326 data, several important pieces of information were obtained by comparing the analysis output from the Mplus and mirt programs. In Table 1, the fit model contains several pieces of information, namely the Akaike information criterion (AIC) and Bayesian information criterion (BIC). AIC and BIC function to compare model fit or nested models from several models being tested (de Ayala, 2022). The AIC and BIC coefficients estimated from mirt and Mplus have similar coefficients, namely 6559 and 6680 for AIC and BIC, respectively. This means that when AIC and BIC are the same, the calculation method used by both software is the same when producing the output.

Table 1. Model fit Mplus and mirt

	mirt	Mplus
AIC	6559.5	6559.5
BIC	6680.7	6680.7
M2	5.5218	-
P-value	0.238	-
RMSEA	0.034	-
TLI	0.968	-
CFI	0.989	-

Sources: Personal data (2024).

The difference in output from the results of GPCM analysis using Mplus and mirt lies in the statistical fit index. It can be seen in Table 1 that the GPCM output from mirt produces M2 statistics for polytomous data (Chalmers, 2012; Maydeu-Olivares & Joe, 2005). Several statistical fit indices for the GPCM model show good model fit: p-value = 0.238 or >0.05, RMSEA = 0.034 or <0.05, CFI and TLI respectively 0.989 and 0968 with criteria >0.95. Meanwhile, Mplus analysis results when using the ML/MLR estimator for categorical data; Mplus does not produce X2 or T2 statistics because the FIML algorithm integration requires very complex calculations to model categorical data analysis (Wang & Wang, 2020). Besides that, Mplus is software that was not developed specifically for IRT. Therefore, one of the findings and advantages of calibrating GPCM using mirt is that it can assess model fit using available statistical fit indices, M2, so that the diagnosis of the unidimensional assumption in IRT can be fulfilled statistically. From Table 1, the GPCM is fit for data from the OMD scale with eight items.

Item Parameters

After assessing the statistical model fit of the GPCM model against the OMD scale data, interpretation of the item parameter results can be carried out. Table 2 shows information on the parameters or psychometric properties of the eight items, namely the item discrimination parameter or α and the difficulty level parameter for each response category or step parameter (***δ***). The item discrimination of Item1 to Item8 ranges from 0.660 to 1.030, with the smallest being Item6 and the largest being Item1. The item discrimination parameter output from calibration using Mplus with mirt produces a similar output between 0.660 and 1.030.

Table 2. Output item parameters Mplus and mirt

	Mplus				**MIRT**			
Item	**α (Mplus)**	***δ*1**	***δ*2**	***δ*3**	**α (mirt)**	***δ*1**	***δ*2**	***δ*3**
Item1	1.030	1.287	-1.211	-0.111	1.029	-1.287	1.215	0.113
Item2	0.811	1.417	-0.482	-0.949	0.813	-1.413	0.486	0.946
Item3	0.887	0.927	-0.741	-0.585	0.887	-0.926	0.743	0.588
Item4	1.009	0.156	0.011	-0.494	1.008	-0.155	-0.011	0.497
Item5	0.685	1.309	0.080	-1.168	0.685	-1.304	-0.082	1.168
Item6	0.668	1.307	-0.131	-0.869	0.667	-1.304	0.131	0.876
Item7	0.808	0.571	-0.584	-0.278	0.812	-0.569	0.583	0.282
Item8	0.911	0.019	-0.418	-0.255	0.900	-0.017	0.423	0.257

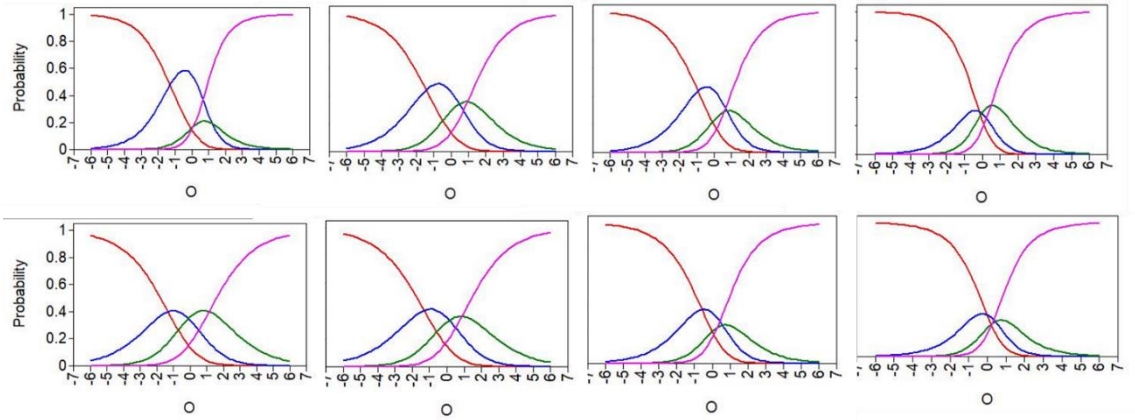
Sources: Personal data (2024).

Then, for the δ parameter, because the OMD scale data has four response options, there are three step or δ parameters. The step parameter in Mplus and mirt produces the same output. Just note that the equalization or standardized parameters in Mplus with IRT are obtained from the following transformation: $\beta_i = -\tau_i$ (Asparouhov & Muthén, 2020). This means that in Mplus can be balanced with the equation above to be equivalent to the results from mirt. However, to obtain the δ coefficient in Mplus, some manual calculations must be done first, especially for the classical GPCM parameterization, namely by subtracting the Item Categories of each response from the Item location of each item or by dividing the Step parameter in the unstandardized output by the item discrimination power. The difference in numbers in the δ parameter is not too big or only differs in rounding to two decimal places, so it can be concluded that the resulting output is similar for all items, Item1 to Item8. In

GPCM and PCM, parameter steps or thresholds are not required to be graded or sequential as in GRM (Wu et al., 2016).

Item characteristic curve (ICC)

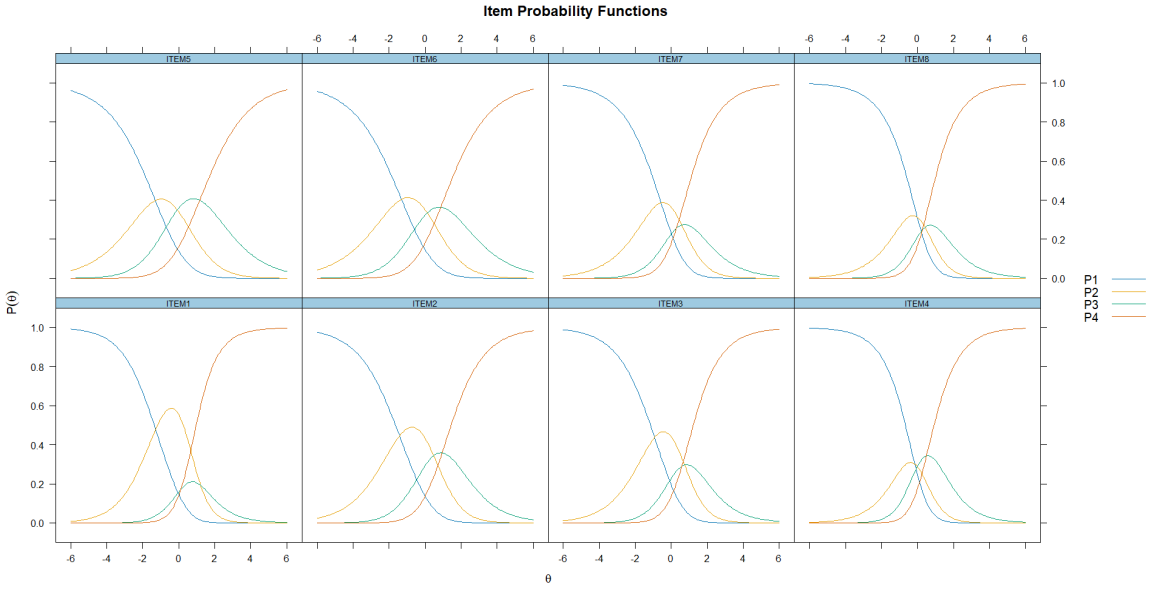
Another property of conducting IRT analysis is the presence of ICC. ICC on polytomous data functions to calculate and determine the probability of participants at a certain trait level choosing an answer at a certain score, in this case, a score of 0 to 4. The respondents' likelihood of answering a certain score category can be calculated using formulas (2), (3), (4), and (5) for four score categories. From these four formulas, the greatest probability at a certain trait level is in what score so that a person can predict the probability of answering an item.



Sources: Personal data (2024).

Figure 1. Item characteristic curve (ICC) Mplus

Figure 1 and Figure 2 show the ICC output from the Mplus and mirt programs, respectively. In Figure 1, starting from the first row from left to right is Item1 to Item4 and the second row from left to right is Item5 to Item8. Meanwhile, in Figure 2 Item1 to Item4 are in the second row or below, and the first row or above are Item5 to Item8. The ICC graph represents the step parameters, namely $\delta 1$, $\delta 2$, and $\delta 3$.



Sources: Personal data (2024).

Figure 2. Item characteristic curve (ICC) mirt

To provide a clearer picture, the following is a demonstration of calculating the magnitude of a person's likelihood at a certain latent trait level (theta), for example, theta -2, 0 and +2 for Item3 using formulas (2), (3), (4) and (5). First, calculate the divider with the formula $1 + exp\ \alpha(\theta - \delta 1) + exp\ \alpha\big(2\theta - (\delta 1 + \delta 2)\big) + exp\ \alpha(3\theta - (\delta 1 + \delta 2 + \delta 3))$, or we denote it as G. For Item3 it has an alpha of 0.887 and a delta of 1 or d or δ1= -0.926, δ2=0.743, and δ3=0.588. G can be calculated for each theta value being explored. The G values for theta -2, 0, and 2 are 1.423, 5.148, and 198.26, respectively. In terms of likelihood, for theta -2, the highest probability is for obtaining a score of 0 or selecting response option 1. For theta 0, the likelihood is higher for getting a score of 1 or choosing response option 2, while for theta 2, it is most likely to get a score of 3 or select response option 4. This pattern can also be confirmed on the ICC graph. For instance, on the ICC Mplus Item3 graph for theta -2, drawing a straight line upward from the x-axis at -2 shows that the highest probability aligns with the red line. Similarly, for theta 0 and theta 2, the highest probabilities correspond to the blue and pink lines, respectively. A detailed summary of the likelihood calculations can be seen in Table 3 below.

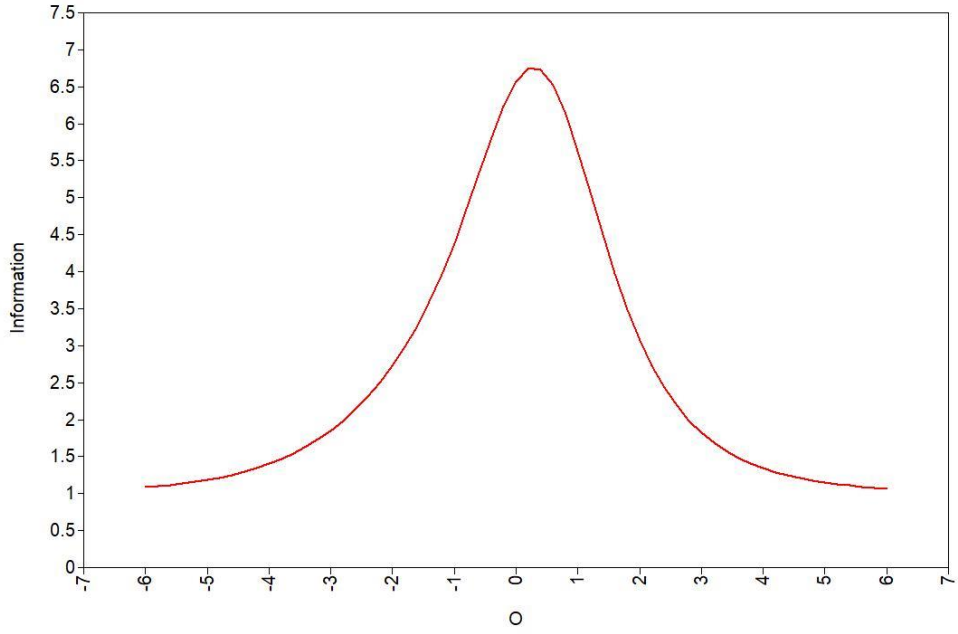
Table 3. *Likelihood ICC theta -2, 0 and 2 on Item3*

Theta	Alpha	G	P0	P1	P2	P3
-2	0.887	1.423	**0.703**	0.271	0.024	0.002
0	0.887	5.148	0.194	**0.442**	0.228	0.136
2	0.887	198.26	0.005	0.067	0.206	**0.721**

Sources: Personal data (2024).

Test Information Function (TIF)

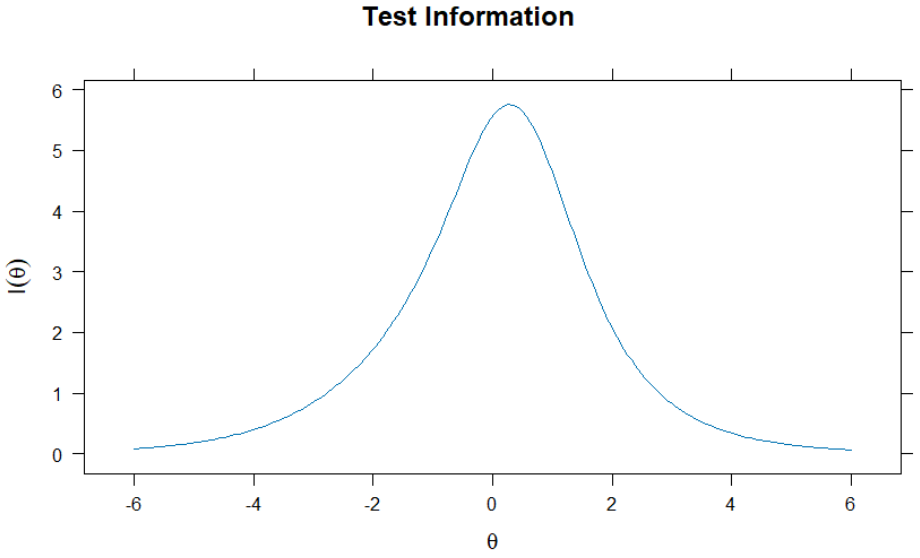
Another property that can be compared is the test information function (TIF). TIF is generated from the sum of all information for each item or the sum of all item information (Rahayu et al., 2023). TIF functions as information about the accuracy of people's estimated score results or reliability in the population because it accommodates all trait levels in the test; this is a property that CTT does not have (Samejima, 1990, 1994).



Sources: Personal data (2024).

Figure 3. TIF Mplus

The results in this study obtained TIF estimates from Mplus and mirt. In Figure 3, the TIF is resulted from Mplus. TIF from Mplus has a peak at a latent trait level of 0.200 with information of 6.748. Meanwhile, Figure 4 shows the GPCM TIF estimate from mirt. The TIF from mirt software has a peak slightly different from the TIF from Mplus, namely at the latent trait level or OMD of 0.271, and produces information of 5,757. The information from TIF produced by mirt software is less than Mplus, but the difference is not too big because it is at the OMD level or at theta 0.20.



Sources: Personal data (2024).

Figure 4. TIF mirt

Correlation of factor scores

The final property that can be compared is the factor score, latent trait score, or, in this context, the OMD scale score from the GPCM analysis. The results of GPCM analysis using Mplus and mirt software are in Table 4 and Figure 3. The Pearson product-moment correlation of the resulting factor scores between mirt and Mplus shows a perfect correlation or 1. This can be interpreted as the factor score result or theta (θ) from Mplus and mirt equivalent.

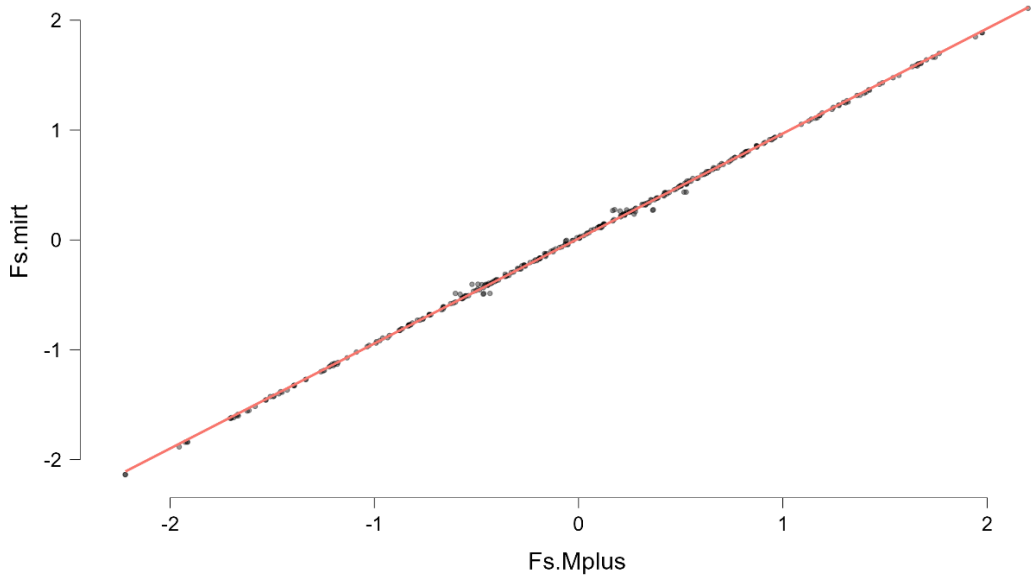
Table 4. Factor score Mplus and mirt

	Fs.mirt	SE Mirt	Fs.Mplus	SE Mplus
Fs.mirt	1			
SE Mirt	-0.39	1		
Fs.Mplus	1.00**	-0.38	1	
SE Mplus	-0.35	1.00**	-0.35	1

** significance <0.01

Sources: Personal data (2024).

This can also be confirmed by Figure 3, which shows that the GPCM score factors of mirt and Mplus are on the line, meaning they show a high correlation, even reaching 1. This factor score can be followed up using other advanced analyses; for example, research needs to analyze regression, moderation, and other use factor scores from the GPCM. GPCM factor scores have a more minor standard error, are more precise, and have an interval scale. Hence, they are more reliable than using total scores or raw scores, which still contain errors in the summations of classical test theory.



Sources: Personal data (2024).

Figure 5. Mplus and mirt score factor correlation graph

Discussion

The development of the psychometric paradigm cannot be separated from industrial and technological developments, especially software developments that operate on open-source software (Wang, 2018). Free software development is an advantage for researchers because it reduces research costs. Even so, the reliability of free software still requires performance testing against paid software. Therefore, this research aims to see to what extent the performance of free or open-source software, namely mirt, with paid software, namely Mplus. This comparison was carried out in the IRT paradigm because IRT requires computing using adequate software. This research uses the IRT GPCM model because it has been widely used in educational and psychological measurements (e.g., OECD, 2024; Schauberger & Mair, 2020; Wallmark et al., 2023; Wind, 2023; Yamamoto & Kulick, 2000).

The findings in this study are in line with previous research that the results of each software are no different (Hayat et al., 2020; Huggins-Manley & Algina, 2015; Sims, 2017; Wang, 2018), although there is a slight difference but it is not very significant. Some of the compared properties are model fit testing, item parameters, ICC, TIF, and factor scores. Overall, the computational procedure or estimation calculation of the GPCM model between Mplus and Mirt produces the same output in model fit and item parameters, including item discrimination power and step parameters, ICC, TIF, and factor scores. This equation shows that the performance of open-source software is reliable. For example, in estimating fit

indices, namely AIC and BIC, mirt and Mplus produce the same coefficients. Furthermore, the resulting factor score correlation correlates 1; this means there is no difference in the estimation results of theta or factor scores from the two applications, likewise with other psychometric properties.

The difference only lies in a few things: Mirt is superior in producing more statistical fit indices. At the same time, Mplus has no statistical fit indices. One reason is that the Mplus program was not explicitly developed for IRT modeling. However, Mplus is flexible statistical modeling that provides a variety of models, estimators, and algorithms for analyzing continuous, categorical, binary, and censored data (Muthén & Muthén, 1998/2017). Meanwhile, mirt or multidimensional item response theory is a program specialized in IRT modeling (Chalmers, 2012).

The statistical fit index produced by MIRT uses M2 statistics (Maydeu-Olivares & Joe, 2005) so that it can produce several fit indices such as RMSEA, CFI, TLI, and SRMR to evaluate the assumption of a unidimensional fit model. Another difference lies only in the information generated from TIF. The information from Mplus is slightly greater than that of mirt. However, the difference is not too big, so mirt and Mplus have the same performance in performing GPCM calibration. Therefore, the findings from this study state that open-source-based software is proven to have the same performance and even has additional features compared to paid software in IRT modeling calibration, in this case, GPCM. However, Mplus provides more detailed output and more information than Mirt. In one analysis, Mplus's advantage is that it can produce two GPCM parameters in one output file, IRT parameterization, and classical parameterization, although it requires some manual calculations. In addition, Mplus provides a single location item for all response options so that comparative interpretations can be made between items directly. By default, Mplus calculated item parameters in GPCM using the formula (6).

The results of this research provide several recommendations for other researchers who will use GPCM. First, researchers can use open-source software without doubt about the output results. It is proven from this finding that open-source software that is free to use can perform similarly to paid software. Second, researchers are expected to be careful when interpreting GPCM because there are two parameterizations commonly used in GPCM. Each software has its different parameterization, such as mirt and ltm. The parameterization referred to in GPCM is divided into IRT or classical parameterization and NRM parameterization. The two parameterizations have different interpretations; namely, the interpretation of the step parameter is some as *b* and some as *d*. Several parameterization formulas for *d* can be found in several references (e.g., Asparouhov & Muthén, 2020; de Ayala, 2022; Reckase, 2009), while parameter *b* is used in this research.

Conclusion

A comparison of the GPCM analysis results between the Mplus software and the R package mirt produces precisely the same results. The calculation methods used by both software apply the EM algorithm as Muraki did. The intended comparison of psychometric properties is the model fit index, alpha parameters or discrimination power, difficulty level parameters or step parameters (delta), ICC, TIF, and score factors. The results show the same or similar results. However, the difference lies in the output of mirt, which can produce statistical fit indices using M2 statistics such as p-value, RMSEA, CFI, and TLI. In Mplus,

there are no model fit statistics because categorical data analysis using a maximum likelihood estimator is very computationally demanding, so it cannot produce a statistical fit index to assess model fit. Besides the lower output difference between both software, Mplus has many advantages, especially for the features included as it should be a paid software. The last, this study and Mirt Package as open-source software can be used for learning material for students when calibrating GPCM.

Acknowledgment

The author would like to thank to all parties for their support especially for the parties who allowing the author to re-use the OMD data.

Conflict of Interest

The author declares no conflicts of interest of this study.

Authors Contribution

ABALF contributed to conception, methodology, data analysis, writing original draft preparation and editing. The author have read and agreed to publish this version of the manuscript.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. . https://doi.org/10.1007/BF02291180

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Asparouhov, T., & Muthén, B. O. (2020). IRT in Mplus Version 4. *Mplus Technical Appendix*, 1–16. www.statmodel.com

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: parameter estimation techniques* (2nd ed.). Taylor & Francis. https://doi.org/10.4324/9780203181287-36

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Gibbons, R. D. (2021). *Item response theory*. Wiley.

Chalmers, R. P. (2012). Mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chou, Y., & Wang, W. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, *70*(5), 717–731. https://doi.org/10.1177/0013164410379322

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). Rasch model in health. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *John Wiley & Sons* (1st ed.). John Wiley & Sons.

de Ayala, R. J. (2022). *The theory and practice of item response theory* (T. D. Little (ed.); 2nd

ed.). The GuilfordPress.

Debelak, R. (2019). An evaluation of overall goodness-of-fit tests for the rasch model. *Frontiers in Psychology*, *9*(JAN). https://doi.org/10.3389/fpsyg.2018.02710

Finch, W. H., & French, B. F. (2019). Educational and Psychological Measurement. In *Educational and Psychological Measurement*. Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315650951

Fu, J. (2020). *A preliminary comparison of five software applications to estimate unidimensional item response theory models (Research Memorandum No. RM-20-02)*. https://www.ets.org/Media/Research/pdf/RM-20-02.pdf

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), 39–50. https://doi.org/10.21831/pep.v24i1.29871

Huggins-Manley, A. C., & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in Mplus. *Structural Equation Modeling*, *22*(2), 308–318. https://doi.org/10.1080/10705511.2014.937374

Kurnia, A. (2019). Analisis Tes Kemampuan Berpikir Kritis Matematika Siswa dengan Menggunakan Generalized Partial Credit Model (GPCM). *PEDIAMATIKA: Journal of Mathematical Science and Mathematics Education*, *01*(02), 105–114. http://digilib.uinsgd.ac.id/22038/

Mair, P. (2018). *Modern psychometrics with R*. Springer International Publishing. https://doi.org/10.1080/00401706.2019.1708675

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

Maydeu-Olivares, A. (2015). Evaluating fit in IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. Routledge.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020. https://doi.org/10.1198/016214504000002069

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

Muthén, L. K., & Muthén, B. O. (n.d.). *Mplus user's guide: Statistical analysis with latent variables* (8th ed.). Los Angeles, CA: Muthén & Muthén.

OECD. (2024). *PISA 2022 Technical Report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en

Paek, I., & Cole, K. (2020). *Using R for Item Response*. Routledge/Taylor & Francis Group.

Putra, M. D. K., & Tresniasari, N. (2015). Pengaruh dukungan sosial dan self-efficacy terhadap orientasi masa depan remaja. *TAZKIYA Journal of Psychology*, *3*(1), 71–82. https://doi.org/10.15408/tazkiya.v20i1.9194

Rahayu, W., Hayat, B., & Putra, M. D. K. (2023). *Analisis rasch: aplikasi dan interpretasi*. UNJ Press.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Samejima, F. (1990). *Redictions of reliability coefficients sand standard errors of measurement using the test information function and its modifications*. University of Tennessee.

Samejima, F. (1994). Some critical observations of the test information function as a measure of local accuracy in ability estimation. *Psychometrika*, *59*(3), 307–329. https://doi.org/10.1007/BF02296127

Samritin. (2018). Kalibrasi tes campuran dikotomus 2PLM dan politomus grm menggunakan prosedur GRM dan GPCM. *JEC (Jurnal Edukasi Cendikia)*, *2*(2), 55–66.

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, *52*(1), 279–294. https://doi.org/10.3758/s13428-019-01224-2

Sims, T. (2017). *Comparison of IRTPRO 3 and Mplus 7 for multidimensional item response item parameter and examinee ability estimation* [Georgia State University]. https://doi.org/10.57709/10130483

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model-data fit of ideal point and dominance models. *Applied Psychological Measurement*, *35*(4), 280–295. https://doi.org/10.1177/0146621610390674

Thissen, D., Nelson, L., Rosan, K., & McLeod, L. D. (2009). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. Lawrence Erlbaum Associates., Inc.

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

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*(6), 389–406. https://doi.org/10.1177/0146621604268734

Wainer, H., & Thissen, D. (2009). True score theory: the traditional method. In H. Wainer & D. Thissen (Eds.), *Test Scoring*. Lawrence Erlbaum Associates., Inc.

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*, *XX*(X), 1–27. https://doi.org/10.3102/10769986231207879

Wang, J. (2018). *Technical report: does it matter which IRT software you use? yes*.

Wang, J., & Wang, X. (2020). *Structural equation modeling: applications using Mplus* (D. J. Balding, N. A. C. Cressie, G. Fitzmaurice, & H. Goldstein (eds.); 2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119422730

Wind, S. A. (2023). Detecting Rating Scale Malfunctioning With the Partial Credit Model and Generalized Partial Credit Model. In *Educational and Psychological Measurement* (Vol. 83, Issue 5). https://doi.org/10.1177/00131644221116292

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers*. Springer Nature Singapore.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the mathematics and science scales. In *TIMSS 1999 Technical Report* (pp. 237–263). International Study Center, Lynch School of Education, Boston College.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(19). https://doi.org/10.1186/s41155-016-0040-x

Appendix

Appendix 1

Syntax Mplus

```
INPUT INSTRUCTIONS

  TITLE:
  Pengujian OMD scale dengan GPCM Mplus;

  DATA:
  FILE IS OMD_GPCM.txt;

  VARIABLE:
  NAMES ARE u1-u8;
  USEVAR ARE ALL;
  CATEGORICAL ARE u1-u8 (gpcm);
  MISSING is .;

  ANALYSIS:
  ESTIMATOR= MLR;

  MODEL:
  O BY u1-u8*;
  O@1;

  OUTPUT:
  TECH1 TECH8;

  PLOT:
  TYPE = PLOT1;
  TYPE = PLOT3;

  SAVEDATA:
  FILE IS FSCOREOMD.txt; SAVE = FSCORES;
```

- Initial step: Prepare a data file in *txt* format consisting of all item response options without headers. Don't forget to set missing data with *dot* ".".
- The second step is to create a syntax according to the Syntax attached in Appendix 1
- In the DATA command, it is necessary to adjust the data file name, in this study the data file name is OMD_GPCM.txt
- After the Syntax is created, click run in Mplus and Mplus will result the output for GPCM.
- To interpret GPCM output or calibration results using classical parametrization as this study, several manual calculations are needed, it is subtracting Item Categories from their respective Item Location in the IRT Parameterization section output, or dividing the Step Parameters of each category in each item by the Factor Loading (Item Discrimination) in the Model Results (Unstandardized Output) section.

- To see the ICC Plots, got to the Plot command in toolbar menu -> view Plots -> Item Characteristic Curve

Appendix 2

Syntax Mirt GPCM

```
# Generating Data
> library (readxl)
> dataset <- read_excel(OMD_R_GPCM.xlsx)
> View(dataset)

# Estimate model GPCM
> GPCM = mirt(OMD_R_GPCM, 1, 'gpcmIRT', SE=T)
> coef(GPCM, simplify=TRUE)

# Obtaining factor score
> peoplegpcmMAP = fscores(GPCM, method = "MAP", full.scores = T, full.scores.SE = T)

# Creating plot ICC for all items
> plot(GPCM, type = "trace")

# Generating TIF
> plot(GPCM, type="info")

# Fitting model using M2*
> M2(GPCM, type = "M2*", calcNull = TRUE, na.rm = FALSE, quadpts = NULL,
      theta_lim = c(-6, 6), CI = 0.9, residmat = FALSE, QMC = FALSE, suppress = 1)
```

- Initial step: Prepare a data file in *xlsx* or *excel file* format consisting of all item response options with headers.
- The second step is to import the dataset into R Studion the select import from excel (Required to install and activate the 'Readxl' Package first)
- Then create a Syntax R according to Syntax attached in Appendix 2 and the specific explanation each syntax accords to the method section explained in this article
- After Syntax created, you can run each command or each Syntax and output for GPCM will be printed in R.