Issue V, 22 November 2022
e-ISSN 2707-9481 Institute of Metallurgy and Ore Beneficiation, Satbayev University, Almaty, Kazakhstan
ISBN 978-601-323-288-1 https://doi.org/10.31643/2022.10

Agus Santoso
Faculty of Science and Technology
Universitas Terbuka, Indonesia
E-mail: aguss@ecampus.ut.ac.id

Challenge of Analysis of Polytomous Item Characteristics with Item Response Theory

Abstract: In the world of education, a lot of scoring is done with a polytomous, for example, on items that are constructed responses. Likewise, at the Open University of Indonesia, the questions for the final exam of a course called the take-home exam (THE) are presented in the form of a constructed response. This problem is done by students who take this course, but the number of students who take this course is not stable. Sometimes, this course is taken by a few participants. On the one hand, it is necessary to carry out an analysis of item characteristics. On the other hand, doing so will face many challenges. In this study, the challenge of analyzing polytomous data is presented on the polytomous score in 3 courses whose final exam is presented in the form of a take-home exam (THE). This study is a mixed research, quantitative analysis packaged in qualitative research with a narrative tradition. Documentary data in the form of students' answers to 3 courses' take-home exams are then analyzed for their characteristics by using various models of IRT polytomous data analysis. Obstacles in conducting analysis are told in narrative form. The analysis was carried out on the take-home exam package of three subjects, namely the Statistical Method II course (89 test participants and two questions), Experimental Design (67 test participants and three questions), and the Sampling method (206 test participants and three questions). Based on the results of the analysis, there is only one package of questions that can be thoroughly analyzed with item response theory, namely the package of questions with the sampling method. Based on the analysis process, it was found that there are challenges in conducting an analysis with item response theory. The challenges are mastery of the R language, the syntax of the selected analysis package, the length or many items in one test package, many test takers, and, last, foresight in rescoring to produce a more proportional pattern. This limitation can be used as a consideration for other researchers in analyzing the polytomous data.

Keywords: challenge, item characteristic analysis, polytomic scoring, item response theory.

Cite this article as: Santoso, Agus. (2022). Challenge of Analysis of Polytomous Item Characteristics with Item Response Theory. *Challenges of Science*. Issue V, 2022, pp. 82-92. https://doi.org/10.31643/2022.10

Introduction

Almost everything that is carried out en masse and structured is carried out through an assessment process. The goal is to measure the achievement of each program goal, including in the world of education. The educational process has goals that lead to the development of student competencies or students according to targets (Arlinwibowo, Retnawati, & Kartowagiran, 2021; Retnawati et al., 2016). To measure the extent to which a person's competence develops, educational institutions use measuring instruments or commonly referred to as certain instruments. The selection of the measuring instrument is highly dependent on the targeted assessment domain. For example, the affective domain uses a questionnaire instrument, and the cognitive domain uses a test instrument (Ebel & Frisbie, 1980; Kubiszyn & Borich, 2003; Miller et al., 2009; Nitko & Brookhart, 2011). These examples are just a few examples of the many instruments that can be used in the measurement process.

The Open University is a university that applies a distance learning model (Mizal et al., 2021; Yaumi, 2007). Just like learning in general, distance learning also has an obligation to measure student learning achievement (Mizal et al., 2021; Sunday A. Itasanmi et al., 2020). This is done to produce a competency achievement profile for each student. The data is used as the institutional basis for determining the status of student learning outcomes (classification of grades and graduation status in each course) and becomes university data to continue to develop the education system (Retnawati et al., 2017).

The concept of distance education makes open universities have to innovate to develop learning achievement measurement instruments that can be done anywhere (Arlinwibowo, Retnawati, et al., 2020; Hamid et al., 2020). One of the assessment mechanisms owned by the Open University is the take-home exam (THE). Take-home exam is a test-based assessment technique carried out by open universities to measure student learning achievement. THE is developed with attention to content that is in accordance with the objectives of each course. Each instrument consists of 2 to 5 items that must be completed by each student.

THE is carried out in the homes of each student. THE will appear for 6 hours. In the span of 6 hours, students are expected to be able to solve all the problems presented in the question package. Then, the system will close access to questions and record the responses submitted by students. Student responses in working on THE questions are then archived by the system. The response archive is ready to be assessed so as to produce a score conclusion for each student.

However, until now, the questions used in preparing the THE question package have not been calibrated and analyzed comprehensively. Supposedly, each item that composes the test package has known its psychometric character before being taken into a test package. Psychometric characters are needed to conclude the quality of each item. The character of the items can be used as the basis for selecting which items have good performance to estimate students' abilities and which items should not be involved in the measurement process. In addition, the characteristics of the items can be used as a basis for inferring the abilities of each student so that a fair assessment process is produced and represents the student's abilities.

Currently, the system at the Open University has archived response patterns that have been converted into assessment scores. The data is polytomous. The response pattern becomes a very valuable asset for analyzing item items so that item quality mapping can be carried out. To analyze polytomous data, item response theory provides various analytical models, namely the graded response model (GRM), partial credit model (PCM), and generalized partial credit model (GPCM).

In the analysis process, researchers must choose one of the most suitable polytomous data analysis models (Arlinwibowo, Retnawati, Hadi, et al., 2021). In establishing the model, the first thing to do is to conduct an analysis based on the instrument's suitability with the philosophy of developing an analytical model. The second analysis is to perform a statistical fit test that tests the suitability of the model with the student response pattern (Arlinwibowo, Retnawati, & Kartowagiran, 2021). Philosophical and statistical studies become the basis for researchers to determine which model will be used as an item analysis tool.

Then, before conducting item analysis using a particular model, researchers need to test the assumptions of item response theory. The assumption test is used to determine the dimensions measured by the instrument (Retnawati, 2014). Test these assumptions using exploratory factor analysis. The results of the assumption test will show the grouping of response patterns. If the response pattern clusters into one dimension, the instrument will be analyzed with unidimensional item response theory, and if the response pattern clusters into more than one dimension, then the instrument will be analyzed with multidimensional item response theory (Arlinwibowo, Achyani, & Galih Kurniadi, 2021; Arlinwibowo, Hadi, et al., 2020). This assumption test is crucial to show the estimation of students' abilities, whether measuring a single ability or splitting it into several abilities. If the instrument measures several abilities, it will be continued with a search for abilities in each dimension.

However, instrument analysis with item response theory has some limitations that must be anticipated. The first limitation is that this theory will be stable when analyzing data with a relatively large number of samples. In addition, this analytical technique also requires a sufficient test length to produce a stable profile estimate. The longer the test (the number of items), the estimation results of students' abilities will be influenced by more minor errors (DeMars, 2010; Uyigue & Orheruata, 2019). Item response theory is an analytical technique that adopts probability theory so that if a test only contains a few items, the standard error generated by the response pattern is not yet stable (Retnawati, 2016).

Thus, the purpose of this study is to determine the quality of THE based on analysis with item response theory. The description of the item profiles produced on the day of the analysis process is expected to be a reference for developing a better THE question bank so that the quality of the tests carried out by the Open University will be better.

Research Methods

This study is mixed, a mix of quantitative analysis and qualitative research with a narrative tradition. Documentary data in the form of student answers on the take-home exam (THE) for the subject of statistical method II, experimental design, and sampling method. The following is a description of the data on the number of samples and the number of items carried out by students.

Table 1. The Result of Regression Coefficients Reading Habits (X) towards Writing Skills (Y)

Code	Course	Number of Students	Number of Items
SATS4222	Statistical Method II	89	2
SATS4321	Experimental Design	67	3
SATS4421	Sampling Method	206	3

The collected data was then analyzed by the polytomous item response theory to determine the character of the evidence. But before that, the data will be analyzed using exploratory factor analysis to determine the dimensions measured by the instrument. Model analysis with item response theory is carried out by considering the results of factor analysis, unidimensional or multidimensional. Then, the researcher analyzed the data using various analytical models in the item response theory, namely the partial credit model, graded response model, and generalized partial credit model. The results of the analysis are then tested for model fit with response patterns so as to produce statistical recommendations for which model can be used to analyze the related data.

Factor analysis and item response theory were carried out by utilizing the R software using the mirt package. The analysis of the constraints in conducting the analysis is told in the form of a narrative that elaborates on the technical and theoretical constraints.

Research Results

In this study there were 3 data sets analyzed in this study. The data set is data obtained from the Take Home Exam (THE). The score in THE is the result of the correction of the student's response to the exam. These three data sets were obtained from the Statistics study program, Faculty of Mathematics and Science at the Open University of Indonesia. The three courses are Statistical Method II, Experimental Design, and Sampling Method.

Students' answers in the home exam are presented in a table (Excel), then the coding is done. Coding is done by considering the scoring rubric that has been designed by the Open University team. With this coding, the score is converted into a simpler power polytomous, namely 0, 1, 2, 3, 4, 5, and 6, by considering the many steps in work. The data were then analyzed using GRM, PCM, and GPCM to see the fit of the model and depicted a categorical response function (CRF) graph for each item. By paying attention to the functioning of the score for each item, the score is updated and then used as material for re-analysis for model fit.

Of the three test sets, there is 1 set, namely SATS4421, which can be analyzed for the characteristics of the items in full, while 2 data sets, SATS4222 and SATS4321, cannot be analyzed completely. Both sets of questions cannot be analyzed completely because of the small size of the data analyzed. Each analysis is presented in detail as follows.

The first description is the result of the analysis in the Statistics Method II course with the code SATS4222. This device consists of 2 items that were responded to by 89 people. After being coded into a simpler polytomy data into six categories for item 1 and 5 categories for both items. After the polytomy data is coded, then the data is analyzed. The frequency distribution for each scale is presented in Table 2.

Table 2. Distribution of Student THE Scores in the Statistical Method II

core	**Item 1**		**Item 2**	
	Fr equency	**Pr oportion**	**Fr equency**	**Pr oportion**
	7	0,0 79	11	0,1 24
	7	0,0 79	8	0,0 90
	12	0,1 35	3	0,0 34
	10	0,1 12	23	0,2 58
	47	0,5 28	44	0,4 94
	7	0,0 79		

The coding results were then analyzed using item response theory. However, the results of the analysis show that item number 1 cannot be analyzed further, and item number 2 does not fit any model. The complete results are presented in Table 4. Observing these results, the analysis of the model fit on the SATS4222 data cannot be continued.

Table 3. Summary of the Appropriateness of the Instrument Model for the Statistical Method II

tems	**GRM**			**PCM**			**GPCM**		
	MSEA	**I -val**	**Inter pretation**	**MSEA**	**I -val**	**Inter pretation**	**MSEA**	**I -val**	**Inter pretation**
	.493	- (.000	- Not Fit	.154	- (.000	- Not Fit	.526	- (.000	- Not Fit

Note: “-“ indicates that the value or interpretation cannot be determined

Next is the analysis related to the results of the response pattern of the SAT4321 Experimental Design device, which consists of 3 items. The device was filled by 67 participants. The analysis process shows that it is necessary to re-score to produce a more proportional response pattern. The re-scoring resulted in a category score of 0, 1, 2, and 3. The consequence of the change in the score was a change in the scoring rubric. The distribution of participants' scores for each score category is presented in Table 4.

Table 4. Distribution of Student THE Scores in Experimental Design Courses

C ategory	**Item 1**		**Item 2**		**Item 3**	
	Fr equency	**Per centage**	**Fr equency**	**Per centage**	**Fr equency**	**Per centage**
0	6	0,0 90	5	0,0 75	10	0,1 49
1	35	0,5 22	17	0,2 54	14	0,2 09
2	13	0,1 94	43	0,6 42	31	0,4 63

3	13	0,1 94	2	0,0 30	12	0,1 79

The response pattern generated from 67 participants has been estimated to be stable. The stability of the score is indicated by the reliability index. In this study, the reliability of the score was estimated using the Cronbach Alpha formula. The estimation results show that the reliability in the medium category is 0.771.

Then before the item analysis is carried out, dimensional analysis is carried out first to determine the many dimensions measured by the instrument. Dimensionality analysis to test the assumptions was carried out by exploratory factor analysis. By utilizing the eigenvalues, a scree plot can be drawn to test the unidimensional assumption. The scree plot results are presented in Figure 1.

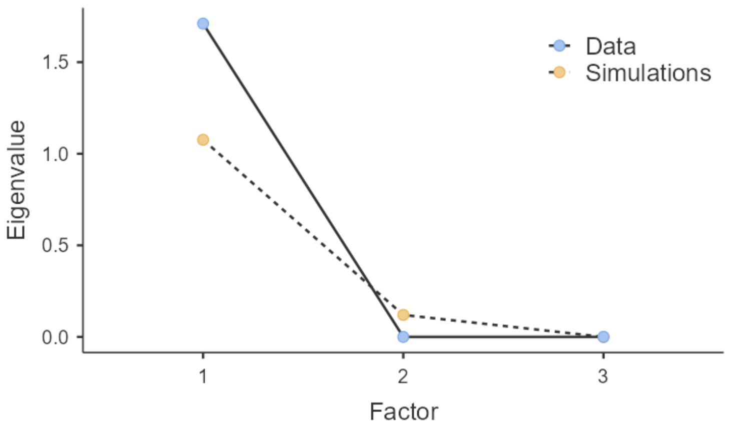

Figure 1. Scree Plot of Exploratory Factor Analysis of Experimental Design Course Data

Based on Figure 1, there is only one factor that has an eigenvalue above one. Thus, it can be concluded that the response pattern instrument of the experimental design course test results only measures one dimension. Therefore, the data were analyzed using a unidimensional polytomous model. The results of the analysis of the suitability of the item model were then carried out using R software. The summary of the results of the analysis presented to see the suitability of the GRM, PCM, and GPCM models is presented in Table 5.

Table 5. Summary of the Suitability of the Experimental Design Course Instrument Model

tems	**GRM**			**PCM**			**GPCM**		
	MSEA	**I** **-val**	**Inter** **pretation**	**MSEA**	**I** **-val**	**Inter** **pretation**	**MSEA**	**I** **-val**	**Inter** **pretation**
	.125	- (.154 -	- Fit -	.105	- (.189 -	- Fit -	.143	- (.125 -	- Fit -

Note: "-" indicates that the value or interpretation cannot be determined

The summary of the fit of the model with the response pattern in Table 5 shows that only item 2 has a model fit that can be identified. The response pattern for item 2 has a good match with all models. However, for items 1 and 3, the fit of the model cannot be concluded. Thus, the analysis process cannot be continued to the next stage.

The Sampling Method test kit consists of 3 items and is carried out by 206 participants. The analysis process shows that it is necessary to re-score to produce a more proportional response pattern. The re-scoring resulted in a scoring category of 0, 1, 2, and 3. The consequence of the change in the score was a change in the scoring rubric. The distribution of the scores of the participants in each score category is presented in Table 6.

Table 6. Distribution of Student Scores in Sampling Method

C **ategory**	**Item 1**		**Item 2**		**Item 3**	
	Fr **equency**	**Per** **centage**	**Fr** **equency**	**Pr** **esentase**	**Fr** **equency**	**Per** **centage**
0	1	0,0 05	20	0,0 98	19	0,0 93
1	76	0,3 71	68	0,3 32	86	0,4 20
2	98	0,4 78	55	0,2 68	78	0,3 80
3	30	0,1 46	61	0,2 98	22	0,1 07
4	0	0	1	0,0 05	0	0

There are categories with too small a frequency (item 1 scores 0, and item 2 scores 4). Thus, it is necessary to improve the scoring process. Improvements are made by combining the scores with a small frequency with a higher frequency. The revised scoring is presented in Table 7.

Table 7. Distribution of Student Scores in Revised Sampling Method Courses

C **ategory**	**Item 1**		**Item 2**		**Item 3**	
	Fr **equency**	**Per** **centage**	**Fr** **equency**	Pr esentase	**Fr** **equency**	**Per** **centage**
0	77	0,3 76	20	0, 098	19	0,0 93
1			68	0, 332	86	0,4 20
2	98	0,4 78	55	0, 268	78	0,3 80
3	30	0,1 46	62	0, 3004	22	0,1 07
4						

The response pattern generated from 206 participants was estimated to be stable. The stability of the score is indicated by the reliability index. In this study, the reliability of the score was estimated using the Cronbach Alpha formula. By using the Cronbach Alfa formula, it can be obtained that the reliability of the Sampling Method test kit of 0.584 is in the medium category.

Then before doing the item analysis, dimensional analysis was first carried out to determine the many dimensions measured by the instrument. Dimensionality analysis to test the assumptions was carried out by exploratory factor analysis. The complete results are presented in the scree plot of Figure 2. These results indicate that the Sampling Method test kit measures the ability dimension only so that it can be said to have unidimensional properties.

Based on Figure 2, there is only one factor that has an eigenvalue above one. Thus, it can be concluded that the experimental design course instrument only measures one dimension. Therefore, the analysis using item response theory was analyzed using a unidimensional polytomous model. The results of the analysis of the suitability of the item model were then carried out using R software. The summary of the results of the analysis presented to see the suitability of the GRM, PCM, and GPCM models is presented in Table 8.

Considering the RMSEA, it can be found that the smallest RMSEA is achieved when the model is in the form of GRM. This indicates that the best model for analyzing this data is the GRM model. This is also

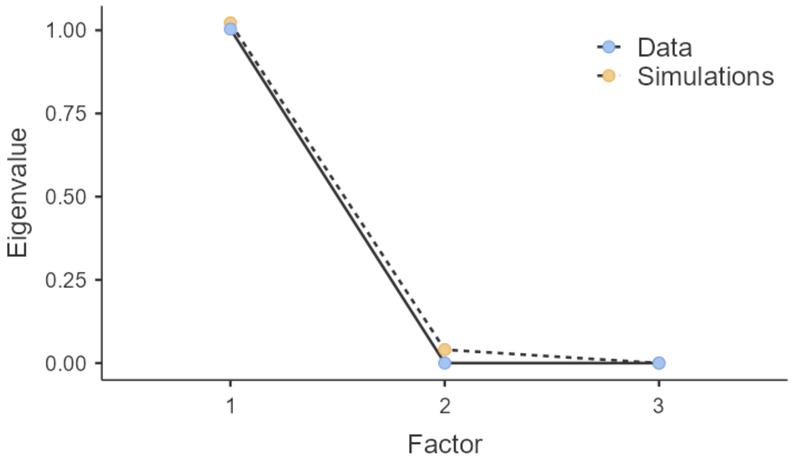

Figure 2. Scree Plot Analysis of Exploratory Factors Data for Subject Test Participants Sampling Method

Table 8. Summary of Instrumental Model Fit for Sampling Method Course Test

	GRM			PCM			GPCM		
tems		**I**	**Inter**		**I**	**Inter**		**I**	**Inter**
	MSEA	**-val**	**pretation**	**MSEA**	**-val**	**pretation**	**MSEA**	**-val**	**pretation**
	.000	(.453	Fit	.044	(.219	Fit	.028	(.325	Fit
	.072	(.127	Fit	.085	(.084	Fit	.081	(.094	Fit
	.083	(.120	Fit	.075	(.142	Fit	.074	(.144	Fit

supported by paying attention to the p-value results. The larger the p-value, the more suitable the model. Thus, based on the p-value criteria, the model fit test showed the same results, namely, GRM became the best model for analyzing empirical data from the take-home exam of the sampling method course.

The results of the analysis show that the data measures one dimension, and the most suitable model is GRM. Thus, the analysis will continue with the unidimensional item response theory with GRM. The results of the R analysis also show the estimation results of the item parameters. The parameter estimation results with the GRM model are presented in Table 9.

Table 9. Characteristics of items of student response patterns on the take-home exam of the Sampling Method Course

	ı	**b**	**b**	**l**	**l**
tems	**1**	**2**	**3**	**ocation**	
.059	: 0.609	- .018	2	- .704	0
.634	: 1.909	- 0.337	- .732	(0.504	-
.545	: 1.993	- .036	0 .908	: 0.016	-

Based on the table of item characteristics, the quality of the items can be traced through the value of a, the sequence of step parameters (bi), and the conclusion of the level of difficulty (location). An item is said to be able to distinguish students' abilities well when it has a value of less than 2 and greater than 0.3. Thus, the three items have a good value. The second consideration is the step parameter (bi). The step or bi parameter is the intersection of the score characteristic curves. An item is said to be good if the intersection is coherent from small to large. Based on the results of the analysis, all the step parameter values are coherent. This means that in item 1, to get a score of 1, at least someone must have theta -0.609. If you want to get a value of 2 then at least students have theta 2.018. The same applies to other items, where the step parameter indicates the theta transition to get a certain value.

Based on the item character profile, it is possible to estimate the function response curve (CRF). Profile visualization based on item characteristics can make it easier for readers to understand the quality of an item. The CRF of the sampling method course test package is presented in Figure 3 as follows.

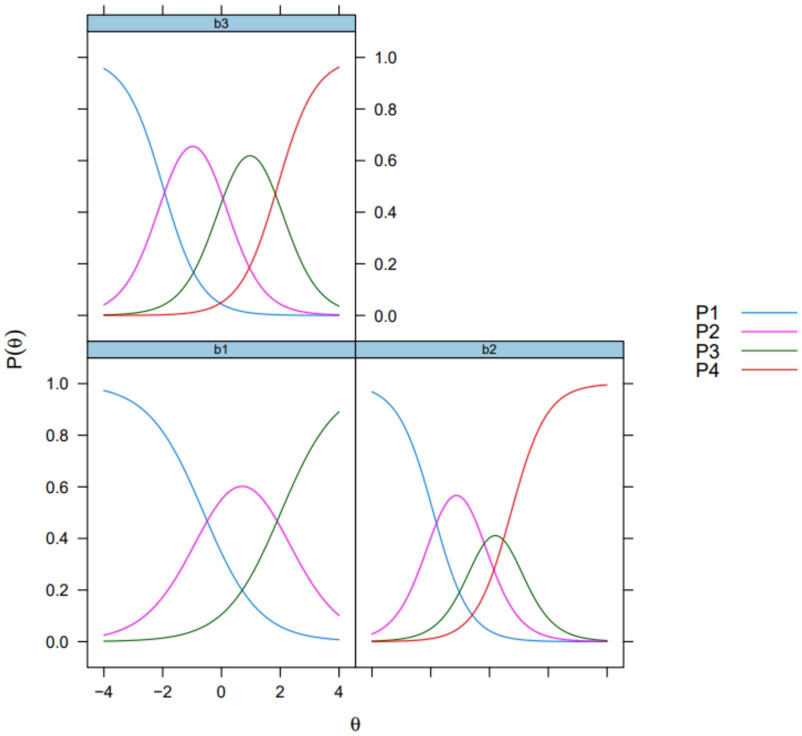

Figure 3. Curve Response Function (CRF) of THE Problem Package in the Sampling Method Course

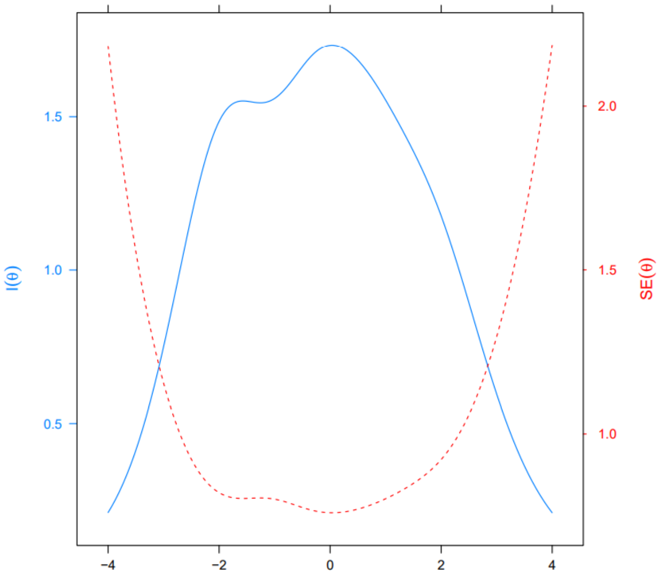

Figure 4. The intersection of the information function and standard error curve

Based on the estimated item parameters, the researcher can trace the value of the information function along with its standardized error. The value of the information function and standard error can be drawn on a single screen so that it shows the intersection of two specific points. The second picture of the graph is shown in Figure 4. Based on Figure 4, it is shown that the device can measure the ability in the range of abilities ranging from -4 to +4, which has covered 96.5% of the overall ability of the test takers.

Challenges Conducting Item Response Theory Analysis on The Take-Home Exam. Item response analysis is the most current analytical technique to determine the quality of an instrument. This analysis continues to grow from time to time, ranging from unidimensional to multidimensional. In addition, along with the development of technology, tools for analysis continue to grow, and there are more choices. Until now, R software has been one of the most complete and powerful options. There are many packages in R that can be used for item analysis with various models. In this study, researchers used a relatively complete package, namely mirt.

In the analysis process, various challenges were found that have the potential to provide obstacles. Knowing this challenge is very important to be able to anticipate potential problems that may occur in instrument analysis with item response theory. The first challenge faced was understanding the language and syntax that worked within a package. Regarding language, R has a special language that must be understood so that we can order things as needed. This aspect is a challenge for people who are not familiar with the R language. Thus, the introduction of the R language is one of the main assets. The second challenge related to technical analysis with R is understanding the syntax of an analysis package. In R, there are many packages with different syntax characters. For example, mirt and ltm are two packages that can be used for item response theory-based analysis, but they have very different syntaxes. Thus, we must master the commands in the package through a specific package guide before using it.

The third challenge is the problem of scoring. There is a possibility of changing the scoring technique during the analysis process. This can be caused because there is a very small score frequency, so the characteristics of that score cannot be analyzed. Thus, there is a difference between the scores that serve as guidelines for the assessment by the institution and the results of the analysis. Changes in scores will have an impact on changes in rubrics and interpretation of scores, especially if the test administering agency is still using the assessment model with classical techniques.

The fourth challenge is related to the character of the item response theory-based analysis technique. Analysis of the quality of the instrument with item response theory is recommended for using data with a large sample size. In fact, the field conditions sometimes do not allow to get many participants. The small number of samples makes the stability of the analysis results low. More extreme things can happen. Namely, the fit of the model and the item parameters cannot be estimated. Thus, the analysis cannot produce any information.

The next challenge is the demand for a relatively large number of items. The number of items that are only 2 or 3 can potentially make the analysis tool unable to function properly. Even if the analysis can produce information, the estimation of the participants' abilities is likely still in the high error range. With the high error content in the participant's theta estimation, the results of the student's ability estimation do not accurately show the original ability.

Research Discussions

Item Response Theory (IRT) was constructed as a modern item analysis to address the shortcomings of classical theory. IRT is a test theory based on a probabilistic model derived from the pattern of examinees' responses to a series of test items (Price, 2017). IRT has the characteristics of (1) the character of the item does not depend on the sample of the examinee, (2) the focus of the analysis is more on the quality of the item than the test, and (3) the model measures students' abilities with precision (Hambleton et al., 1991).

To perform an IRT-based item analysis, we are given a large selection of applications (Marsigit et al., 2020) ranging from paid to free. One free application that has the ability to perform analysis of various models is R (Chalmers, 2012; Ince Araci & Tan, 2022; Rizopoulos, 2006). Of the many packages, researchers chose the mirt package because it was considered the most complete and in accordance with the purpose of the analysis (Chalmers, 2012). However, the thing that is challenging in the analysis process with R is mastering the R language and the package that will be used. The R language is the initial modal, while the advanced modal is the mastery of the syntax in each package.

There are recommendations for many samples and length of questions in the item response theory analysis so that the analysis results are accurate (DeMars, 2010; Sahin & Anıl, 2017; Suwarto et al., 2019). The number of samples and the length of the questions are determined by the model selected in the analysis process. For dichotomous data, the 1PL model requires at least 10 items with a sample size of 150 (Şahin & Anıl, 2017), while some suggest 200 (DeMars, 2010; Uyigue & Orheruata, 2019). For the 2PL and 3PL models, it takes at least 10 items with a sample lot of 750 (Şahin & Anıl, 2017). The fewer parameters considered, the fewer participants and the minimum items needed (Suwarto et al., 2019). In addition, the number of test items will also have an influence on the standard error in theta estimation. If there are too few items involved in the test, the resulting theta estimate contains a larger error (Arlinwibowo, Retnawati, Hadi, et al., 2021). Thus, it is natural that a small data set will encounter difficulties in the analysis process.

Before determining the analysis model, it is necessary to test the fit of the model first. The test is used to show that the model has conformity with the empirical data (response pattern data). RMSEA, designated, is an absolute fit index scaled as a badness-of-fit statistic where a value of zero indicates the best result (Kline, 2016: 273). The RMSEA value of 0.08 is the limit set for the fit of the data model in the analysis (Price, 2017: 340). Kline (2016: 274) states that the approved model has a good fit when RMSEA < 0.05. (Finch & French, 2019: 153) and (Coulacoglou & Saklofske, 2017: 301) stated that the 0.05 < RMSEA < 0.08 model had a sufficient fit.

With the limitation of many items and the size of the sample, the results of the analysis show that there is only one package that can produce a complete analysis output. The results of the model fit test show that GRM is the best model for analysis. GRM is very suitable for analyzing polytomous data with the character of the instrument having graded answer choices and aims to measure a person's attitude (Reckase, 1997). The results of the model fit show that the GRM is the best model for the analysis of the problem package. The results of the model fit test support the previous statement that philosophically, instruments with graded options are suitable for analysis with the GRM model (Chalmers & Ng, 2017). The value of bi is a step parameter resulting from the intersection of the mn and mn+1 categories of graphs (Embretson & Reise, 2000). bi refers to the minimum ability to enter the higher category points (Retnawati, 2014).

Table 9 data shows that the values of b1, b2, and b3 have a good (ideal) order, namely b1 < b2 < b3 (Reckase, 1997). Therefore, the difficulty level of each item meets the criteria of good quality and can represent the ability of the test takers. GRM is an analysis of the response of polytomous data items that take into account the parameter a (discriminant index). According to (Hambleton & Swaminathan, 1985), the item is said to be good if the discriminant index value is between 0 to 2 (Arlinwibowo, Retnawati, & Kartowagiran, 2021). Thus, all items of the collaborative ability assessment instrument have the ability to distinguish good student abilities, namely $1.059 \leq ai \leq 1.634$.

Conclusions

The analysis was carried out on the take-home exam package of three subjects, namely the Statistical Method II course (89 test participants and two questions), Experimental Design (67 test participants and three questions), and the Sampling method (206 test participants and three questions). Based on the results of the analysis, there is only one package of questions that can be thoroughly analyzed with item response theory, namely the package of questions with the sampling method. Based on the analysis process, it was found that there are challenges in conducting an analysis with item response theory. The challenges are mastery of the R language, the syntax of the selected analysis package, the length or many items in one test package, many test takers, and, last, foresight in rescoring to produce a more proportional pattern. This limitation can be used as a consideration for other researchers in analyzing the polytomous data.

Cite this article as: Santoso, Agus. (2022). Challenge of Analysis of Polytomous Item Characteristics with Item Response Theory. *Challenges of Science*. Issue V, 2022, pp. 82-92. https://doi.org/10.31643/2022.10

References

Arlinwibowo, J., Achyani, I., & Galih Kurniadi. (2021). Multidimentional item respose utilization for validating mathematics national examination in Indonesia. PVJ_ISComSET 2020, 1–7. https://doi.org/10.1088/1742-6596/1764/1/012113

Arlinwibowo, J., Hadi, S., & Firdaus, E. M. (2020). Validasi Persepsi Siswa Terhadap Pembelajaran Sains Dengan Teori Respon Butir

Multidimensi. Jurnal Ilmu Komputer …, 1(2), 7–14. https://www.ejr.stikesmuhkudus.ac.id/index.php/jikoma/article/view/971%0Ahttps://www.ejr.stikesmuhkudus.ac.id/index.php/jikoma/article/viewFile/971/686

Arlinwibowo, J., Retnawati, H., Hadi, S., Kartowagiran, B., & Kassymova, G. K. (2021). Optimizing of item selection in computerized adaptive testing based on efficiency balanced information. Journal of Theoretical and Applied Information Technology, 99(4), 921–931.

Arlinwibowo, J., Retnawati, H., & Kartowagiran, B. (2021). Item Response Theory Utilization for Developing the Student Collaboration Ability Assessment Scale in STEM Classes. Ingenierie Des Systemes d'Information, 26(4), 409–415. https://doi.org/10.18280/ISI.260409

Arlinwibowo, J., Retnawati, H., Kartowagiran, B., & Kassymova, G. K. (2020). Distance learning policy in Indonesia for facing pandemic COVID-19: School reaction and lesson plans. Journal of Theoretical and Applied Information Technology, 98(14), 2828–2838.

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

Chalmers, R. P., & Ng, V. (2017). Plausible-Value Imputation Statistics for Detecting Item Misfit. Applied Psychological Measurement, 41(5), 372–387. https://doi.org/10.1177/0146621617692079

Coulacoglou, C., & Saklofske, D. H. (2017). Psychometrics and psychological assessment: principles and applications. Academic Press Inc.

DeMars, C. (2010). Item response theory. Oxford University Press, Inc.

Ebel, R. L., & Frisbie, D. A. (1980). Essentials of Educational Measurement (5th ed.). Prentice-Hall.

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

Finch, W. H., & French, B. F. (2019). Educational and psychological measurement. Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Springer Science and Business Media, LLC.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory Library. Sage Publications.

Hamid, R., Sentryo, I., & Hasan, S. (2020). Online learning and its problems in the Covid-19 emergency period. Jurnal Prima Edukasia, 8(1), 86–95. https://doi.org/10.21831/jpe.v8i1.32165

Ince Araci, F. G., & Tan, Ş. (2022). Multidimensional Computerized Adaptive Testing Simulations in R. International Journal of Assessment Tools in Education, 9(1), 118–137. https://doi.org/10.21449/ijate.909616

Kline, R. B. (2016). Principles and practice of structureal equation modeling (4th ed.). The Guilford Press.

Kubiszyn, T., & Borich, G. (2003). Educational Testing and Measurement. John Wiley & Sons, Inc.

Marsigit, M., Retnawati, H., Apino, E., Santoso, R. H., Arlinwibowo, J., Santoso, A., & Rasmuin, R. (2020). Constructing Mathematical Concepts through External Representations Utilizing Technology : An Implementation in IRT Course. TEM Journal, 9(1), 317–326. https://doi.org/10.18421/TEM91

Miller, M. D., Linn, R. L., & Grondlund, N. E. (2009). Measurement and assessment in teaching (10th ed.). Pearson Education, Inc.

Mizal, B., Basith, R. I., & Tathahira, T. (2021). Critical Thinking Through Distance Learning: An Analysis of Indonesian Open University. International Journal of Education, Language, and Religion, 3(1), 17. https://doi.org/10.35308/ijelr.v3i1.3667

Nitko, A. J., & Brookhart, S. M. (2011). Educational assessment of student. Pearson Education, Inc.

Price, L. R. (2017). Psychometric methods: Theory into practice. The Guilford Press.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of Item Response Theory (pp. 271–286). Springer Science and Business Media, LLC. https://doi.org/10.1201/b19166

Retnawati, H. (2014). Teori respons butir dan penerapannya [Item response theory and its application]. Nuha Medika.

Retnawati, H. (2016). Analisis kuantitatif instrumen penelitian [Quantitative analysis of research instruments]. Parama Publishing.

Retnawati, H., Hadi, S., & Nugraha, A. C. (2016). Vocational high school teachers' difficulties in implementing the assessment in Curriculum 2013 in Yogyakarta Province of Indonesia. International Journal of Instruction, 9(1), 33–48.

Retnawati, H., Hadi, S., Nugraha, A. C., Arlinwibowo, J., Sulistyaningsih, E., Djidu, H., Apino, E., & Iryanti, H. D. (2017). Implementing the computer-based national examination in Indonesian School: The challenges and strategies. Problems of Education in The 21st Century, 75(6), 612–633.

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

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. Kuram ve Uygulamada Egitim Bilimleri, 17(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Sunday A. Itasanmi, Mathew T. Oni, & Omobola O. Adelore. (2020). Students' Assessment of Open Distance Learning Programmes and Services in Nigeria: A Comparative Description of Three Selected Distance Learning Institutions. IJORER : International Journal of Recent Educational Research, 1(3), 191–208. https://doi.org/10.46245/ijorer.v1i3.64

Suwarto, Widoyoko2, E. P., & Setiawan, B. (2019). The Effects of Sample Size and Logistic Models on Item Parameter Estimation. Proceedings of the 2nd International Conference on Education, 323–330. https://doi.org/10.4108/eai.28-9-2019.2291082

Uyigue, A. V., & Orheruata, M. U. (2019). Test Length and Sample Size for Item-Difficulty Parameter Estimation in Item Response Theory. Journal of Education and Practice, 10(30), 72–75. https://doi.org/10.7176/jep/10-30-08

Yaumi, M. (2007). the Implementation of Distance Learning in indonesian Higher Education. Learning, X(2), 196–215..